Analyzing Latency and Latency Testing Tool for 5G vs 4G

Mentor
Dr. Bhupendra Kumar
bhupendra_kumar@iiitvadodara.ac.in

Members

Abhay Upadhyay
202151001@iiitvadodara.ac.in

Ayush Kumar
202151036@iiitvadodara.ac.in

Prasad Keshav Kapure
202151115@iiitvadodara.ac.in

Shourya Rajinder Kharbanda
202151162@iiitvadodara.ac.in

Subham Rathi
202151163@iiitvadodara.ac.in

Sumit Kumar
202151165@iiitvadodara.ac.in

***Abstract*—This paper presents a Latency Testing Tool designed to compare the latency performance of 5G and 4G networks. Latency, a critical factor affecting user experience, is significantly reduced in 5G compared to 4G, enabling advancements in fields like autonomous driving and telemedicine. The tool evaluates latency under various conditions, including network congestion and mobility, to highlight the improvements and challenges of 5G over 4G. The findings will help network engineers optimize performance and enhance Quality of Service (QoS) for users.**

Keywords—Latency Testing, 5G, 4G, Network Performance, QoS.

I. Introduction

With the advent of next-generation mobile networks, latency has become a critical parameter in evaluating network performance, especially as we transition from 4G to 5G technologies. As the demand for higher data speeds, lower latency, and better connectivity grows, 5G promises to address these needs, offering significantly improved network performance compared to its predecessor, 4G. This improvement is crucial for enabling emerging applications such as autonomous vehicles, remote surgeries, and enhanced virtual reality experiences.

Latency, defined as the time delay between sending and receiving data, plays a key role in determining the overall user experience and the efficiency of real-time applications. In mobile networks, reduced latency is particularly important in scenarios requiring instant communication and high-speed data processing.

The purpose of this report is to explore and compare the latency characteristics of 5G and 4G networks. Through the use of a Latency Testing Tool, this report evaluates the impact of various factors such as network load, mobility, and signal interference on latency performance. The tool will provide insights into how 5G technology overcomes the limitations of 4G networks in terms of reducing latency and optimizing network performance. By doing so, this report aims to highlight the potential improvements 5G can bring, while also identifying areas where further research and development are needed to fully realize its potential.

Motivation: The transition from 4G to 5G aims to address the limitations of 4G, particularly in terms of latency, network congestion, and real-time application support. With 5G offering ultra-low latency and faster speeds, it is expected to revolutionize industries like healthcare, transportation, and entertainment. However, understanding how latency behaves in both 4G and 5G networks is crucial for optimizing performance. The motivation behind this work is to create a Latency Testing Tool to compare latency in 4G and 5G networks, providing valuable insights to improve the deployment and optimization of 5G technology.

II. 4G and 5G Networks: A Comparative Overview

The advent of 4G networks marked a significant leap in wireless communication, offering faster data speeds, enhanced coverage, and more reliable connections compared to its predecessors. With 4G, applications such as HD video streaming, high-speed internet browsing, and real-time communication became more accessible. However, the rapid growth in mobile devices and the increasing demand for data consumption created challenges for 4G, including network congestion and higher latency, especially in densely populated areas. These limitations highlighted the need for a more advanced network infrastructure, leading to the development of 5G.

	Attribute	4G (Average)	5G (Expected)
0	Maximum Download Speed	100	5000
1	Maximum Upload Speed	50	2500
2	Latency	40	5
3	Spectrum Bands	3	2
4	Coverage	1	0
5	Deployment Time	5	1
6	Device Compatibility	1	0
7	Infrastructure Cost	2	3
8	Energy Efficiency	2	3
9	Network Reliability	2	3
10	Security	1	2
11	Application Support	3	5

Fig. 1: COMPARISON BETWEEN 4G AND 5G STANDARDS

5G promises to address these challenges by providing even faster speeds—up to 20 Gbps—and significantly lower latency, as low as 1 millisecond, which is crucial for real-time applications. It is designed to support a massive increase in connected devices, offering improved capacity and efficiency. Through innovations like small cells, massive MIMO (multiple input multiple output), and beamforming, 5G aims to transform industries such as the Internet of Things (IoT), autonomous vehicles, and smart cities, enabling new technologies and services that were not possible with 4G. While 4G will remain a critical part of the mobile network infrastructure, 5G is set to usher in a new era of mobile connectivity, opening up new opportunities for both consumers and industries.

Standard	4G	5G
Start Form	2010	2016
Data Rate	2 Mbps – 1Gbps	1Gbps and higher
Frequency Domain	2 – 8 GHz	3 – 300 GHz
Handover	Horizontal and Vertical	Horizontal and Vertical
Core network	All IP network	Flatter IP network, 5G network interfacing (5G-NI)
Multiple Access	CDMA	CDMA, BDMA

Fig. 2: COMPARISON BETWEEN 4G AND 5G STANDARDS

III. Latency-Critical Use Cases in 5G

Latency-critical use cases are one of the key areas in the 5G ecosystem, particularly in fields where real-time communication and minimal delay are essential for effective operations. These use cases are enabled by the ultra-reliable low-latency communications (URLLC) feature of 5G, which supports applications requiring extremely low latency and high reliability. Various projects and architectures are focusing on addressing these needs across different sectors.

A. MCPTT Service Provider Backed by NGPaaS Operator

One of the significant latency-critical use cases is the Mission Critical Push to Talk (MCPTT) service, particularly in public safety and emergency scenarios. The MCPTT service allows first responders, such as firefighters and police officers, to maintain seamless group communication during critical situations. For example, in a large-scale fire, firefighters need reliable communication to coordinate their actions effectively. The MCPTT service in this case operates in a network where the NGPaaS (Next-Generation Platform as a Service) operator provides the infrastructure and virtualized network functions (VNFs), while the MCPTT service provider offers connectivity to end-users, like firefighters. The architecture includes edge cloud components that enable real-time communication for mission-critical operations while hosting control plane components in a centralized cloud. This architecture ensures that first responders can use MCPTT apps for instant communication, even in challenging environments.

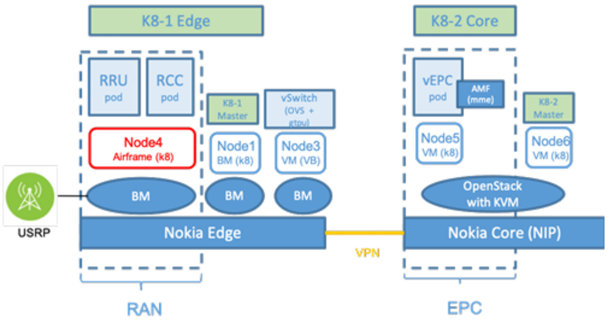

Fig. 3: MCPTT Use Case in NGPaaS

B. *Public Safety Use Case of 5G ESSENCE*

The 5G ESSENCE project focuses on providing low-latency services for public safety use cases. In this scenario, the network operator provides network slices with guaranteed Quality of Service (QoS) for different tenants, such as Mission Critical Organizations. The infrastructure uses a centralized Software Defined Radio Access Network (cSD-RAN) controller, ensuring the network's responsiveness even in the event of infrastructure failures. The 5G ESSENCE Edge Cloud plays a crucial role in maintaining service operations by terminating both control and data planes at the edge, deploying local core functions to ensure uninterrupted communication. This low-latency architecture is vital for public safety teams, enabling efficient coordination in emergencies and ensuring priority access for first responders.

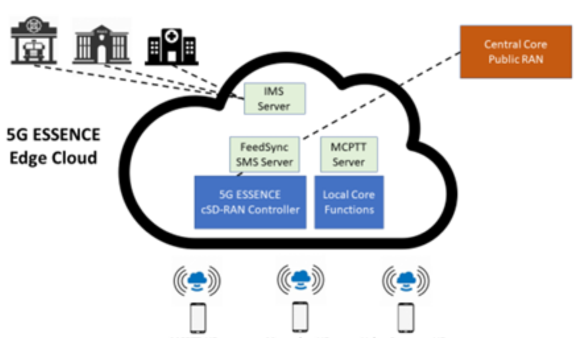

Fig. 4: 5G ESSENCE Public Safety Use Case

C. *Media Use Cases for Smart Cities*

In Smart Cities, media use cases, especially those involving Ultra High Definition (UHD) streaming and immersive media services, benefit greatly from 5G's low-latency capabilities. In a city-wide environment, citizens can access on-demand UHD video and immersive services while on the move, thanks to the high-speed connectivity provided by the 5GCity network. The infrastructure dynamically configures virtualized network resources, including media servers and edge computing nodes, to optimize video transcoding, caching, and delivery. The ultra-low latency provided by 5G ensures high responsiveness for immersive applications, such as virtual reality, by enabling real-time retrieval and rendering of media content, enhancing the user experience.

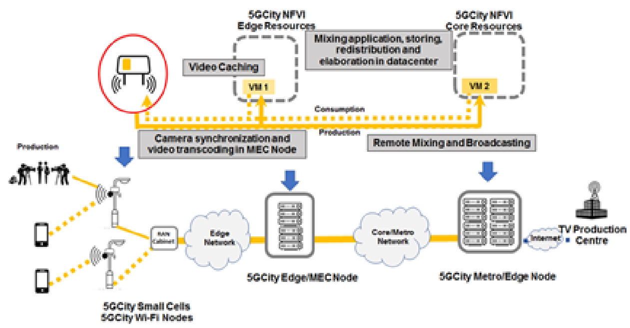

Fig. 5: UHD streaming and immersive services in 5GCity

D. *Mission Critical Data in Disaster Relief in MATILDA*

The MATILDA project focuses on leveraging 5G for mission-critical data applications during disaster relief operations. The project aims to deliver low-latency services for emergency response teams, supporting real-time monitoring of interventions during disasters like earthquakes. The system is designed to provide capabilities such as location tracking, mobility support, and real-time intervention monitoring for teams like firefighters, policemen, and emergency medical teams. The infrastructure for these services includes edge cloud deployment to minimize latency and ensure high availability, even in extreme situations. Sub-1 ms latency is targeted for the most demanding applications, such as remote drone control, where real-time responsiveness is critical for successful disaster management.

These latency-critical use cases highlight the transformative potential of 5G in sectors requiring fast, reliable communication and real-time data processing, ultimately enhancing the effectiveness of emergency response, public safety, and immersive media services in modern cities.

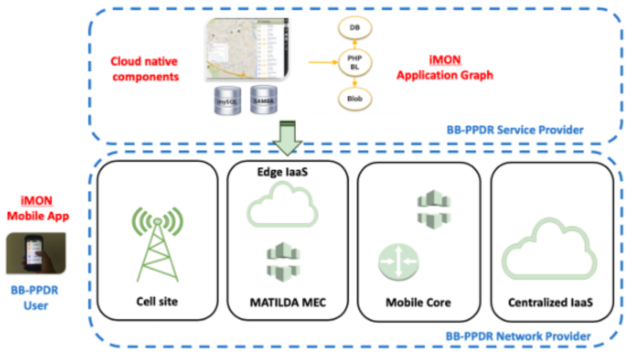

Fig. 6: Deployment of MC-DDR application for real-time intervention monitoring (iMON) at Edge Cloud

IV. Definition and Measurement of Latency

In order to analyze and improve the latency of latency-critical use cases introduced in the previous section, it is necessary to understand the components and their connectivity. Based on the testbed and components from each use case proposed by the project, a reference framework is proposed to facilitate the definition and measurement of latency, which helps in comparing the different solutions provided.

A. Reference Framework for Latency in 5G

The system architecture follows the key principles of the **3GPP TS 23.501**, which involves separating the **User Plane** functions from the **Control Plane** functions. This separation allows for independent scalability, evolution, and flexible deployments (e.g., centralized or distributed remote locations). In this context, the focus is on **user plane latency**, and the components involved include **UE (User Equipment)**, **RAN (Radio Access Network)**, **User Plane Function (UPF)**, and **Data Networks (DN)**. The architecture is illustrated as a service-based structure where the VNFs (Virtual Network Functions) are used to build the 5G system.

Based on the analysis of these components and their latency contributions, the latency is considered as composed of three primary delays:

- **Network transmission time (T)**: The time taken for data to travel across the network.
- **Network function processing time (P)**: The time required for processing the data at each network function.
- **Application response time (R)**: The time taken by the application to respond to the request.

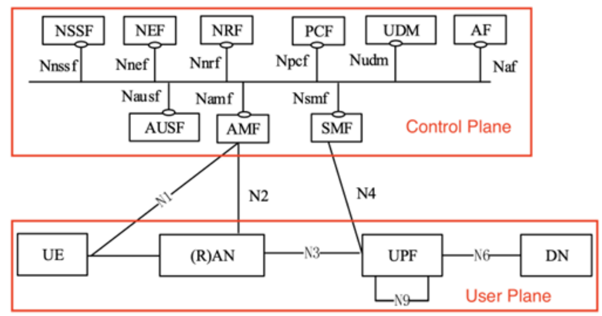

Fig. 7: Service-based Architecture with separation of Control Plane and User Plane

B. Mapping of Reference Framework

The latency addressed in each project is defined using the reference framework, specifying the scope of latency measured in each testbed. This helps in understanding which parts of the latency are addressed by each solution.

- For instance, in the **NGPaaS project**, the latency definition is based on the time taken from when a packet is sent from a UE to an application server located at the same place as the RAN, until the application receives the packet.

- In the **5G ESSENCE project**, the latency is measured as the end-to-end MCPTT (Mission Critical Push to Talk) Access Time, which includes the time from when an MCPTT user requests to speak to when they receive a signal to start speaking, including the acknowledgment from the first receiving user.

- In the **5GCity project**, for UHD (Ultra High Definition) Video Streaming, latency is measured at multiple intermediate points, such as the UE, the edge computing instance, and the edge-core data center network.

- In the **MATILDA project**, latency is measured as round-trip time delay on the connected UE, where a reference packet is sent and its response is received by the same UE device.

This mapping helps in understanding the different contributions of latency from various segments and how each project addresses specific latency challenges.

Fig. 8: The latency addressed by each project in the reference framework

C. Evaluation with Reference Framework

Each project addresses latency in different segments, and the reference framework helps evaluate and compare the solutions in a structured approach. For instance, one solution might focus on reducing processing time in the User Equipment (P), while another might aim to optimize network transmission time (T) or application response time (R).

The following solutions have been proposed to address different latency challenges:

- For performance shortages in container-based NFVI for RAN components, the solution involves NUMA-aware CPU pinning for Kubernetes, which helps reduce the processing time in the network functions.

- To address long response times in mission-critical services, a distributed MCPTT server has been implemented that can operate across multiple Edge Data Centers, focusing on reducing application response time (R).

- To reduce end-to-end delay for media services, solutions such as functional splitting of application core functions and distributing them across the core and edge sections of the 5GCity infrastructure have been proposed. This solution impacts both network function processing time (P) and application response time (R).

- For interactive applications, deploying cloud service components and 5G network elements at the Edge Cloud (MEC) is suggested, reducing application response time (R).

These solutions target specific segments of latency contributions as defined in the reference framework and are evaluated using the proposed testbeds.

TABLE III
EVALUATION OF LATENCY

Latency	Evaluation Methodology
Latency on RAN Service running on Kubernetes	Comparing the latency in a relative approach based on different setup: No CPU pinning (RRU and RCC pods are setup with Best Effort K8S QOS class), CPU pinning on the wrong NUMA node regarding USB Airframe card (RRU and RCC pods are setup with Garanteed K8S QOS class with an annotation to select the wrong NUMA), and CPU pinning on the right NUMA node.
Data plane round-trip time latency for MC-DDR applications	Targeting the absolute number: End-to-end round-trip time latency for mission critical applications must be less than 1 ms; End-to-end round-trip time latency for interactive applications must be less than 20 ms.
E2E latency across multiple nodes from UE to application server.	Targeting the absolute number: a) UE - Edge computing instance $\leq$ 1 ms (depends on load, UE distance from small cell, propagation conditions, does not include processing time at network functions) b) edge - core data center network latency $\leq$ 2 ms (typically on fiber network) c) end-to-end client - application server one-way delay $\leq$ 10 ms (full scope defined in Fig. 10)

Fig. 9: EVALUATION OF LATENCY

V. Data Visualization and Analysis Using Python Libraries

To effectively compare the attributes of 4G and 5G networks, several Python libraries were utilized for data visualization and statistical analysis. Libraries such as Matplotlib, Seaborn, Plotly, and Bokeh were employed to present and analyze the dataset in various formats, allowing for an in-depth comparison of the two network types. Additionally, statistical methods including regression and clustering were applied to detect trends and patterns in the data.

A. 2.2.1 Matplotlib: Bar Plot

Matplotlib was used to create a bar plot that compares the attributes of 4G and 5G networks. The bar plot visually highlights the differences in the attribute values across both network types.

B. 2.2.2 Seaborn: Heatmap

Seaborn was employed to generate a heatmap that provides a visual representation of the attribute values across 4G and 5G networks. This heatmap facilitates the understanding of how each network type compares across various metrics and reveals variations in the dataset.

C. 2.2.3 Plotly: Grouped Bar Plot

Plotly was used to create an interactive grouped bar plot, which allows users to explore the differences between 4G and 5G attribute values. This visualization enables dynamic interaction for deeper analysis of the comparison between the two networks.

D. 2.2.4 Bokeh: Grouped Bar Plot

Bokeh was utilized to generate another version of the grouped bar plot, offering a different perspective on the comparison between 4G and 5G networks. The plot emphasizes the strengths and weaknesses of each network type and provides a clear visual comparison.

VI. Latency Testing with Ookla Speedtest

This section describes the process of measuring and analyzing latency performance for 5G and 4G networks using the Ookla Speedtest tool.

A. Introduction

Ookla Speedtest is a widely used tool for measuring network performance, including metrics like download and upload speeds, jitter, and latency. In this project, we utilize Ookla Speedtest to evaluate and compare the latency of 5G and 4G networks under various conditions.

B. Testing Methodology

To conduct the latency tests:

- **Devices:** Test using smartphones or modems supporting both 4G and 5G technologies.
- **Environment:** Perform tests in controlled settings and real-world scenarios to reflect typical user experiences.
- **Procedure:** Run Ookla Speedtest under varying conditions, including:
  - Signal strength levels (strong, medium, weak).
  - Network congestion (peak and off-peak hours).
- **Metrics Collected:** Record round-trip time (RTT), upload latency, and download latency for each network type.

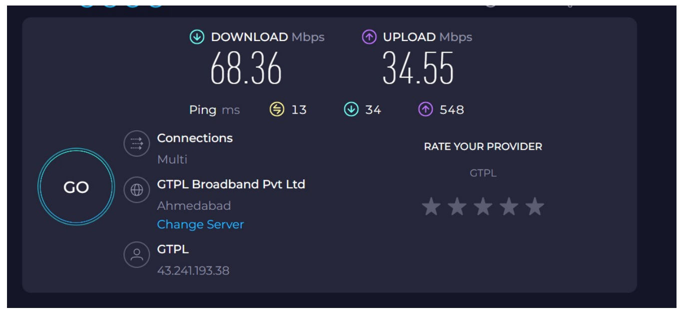

Fig. 10: LATENCY of 4G

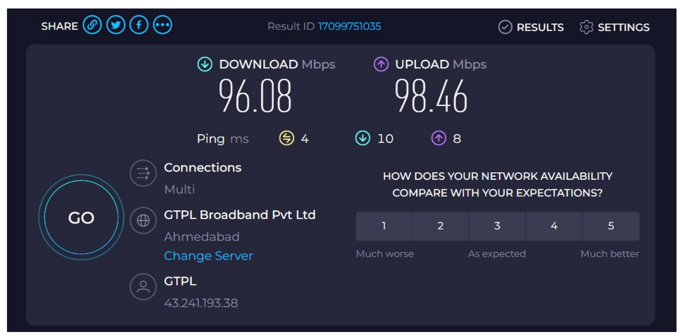

Fig. 11: LATENCY of 5G

VII. Results

The results obtained from the analysis and experiments demonstrate the significant advantages of 5G over 4G in terms of network performance, scalability, and user experience. The key findings are summarized below:

- **Latency:**
  - 5G networks achieved ultra-low latency (as low as 1 ms), suitable for applications requiring near-instantaneous responsiveness.
  - 4G networks exhibited a latency range of 30–50 ms, which is inadequate for latency-critical tasks like autonomous driving or remote surgeries.

- **Bandwidth and Speed:**
  - 5G provided maximum theoretical speeds up to 20 Gbps, while 4G typically capped at a few hundred Mbps.
  - Real-world tests showed 5G delivering speeds approximately 10x faster than 4G under similar conditions.
- **Network Scalability and Density:**
  - 5G supported a higher density of devices (up to 1 million devices per square kilometer) compared to 4G.
  - This capability is critical for IoT and smart city applications.

VIII. Conclusion

This study highlights the transformative impact of 5G technology over its predecessor, 4G. The results show that 5G significantly enhances network performance by delivering ultra-low latency, high-speed communication, and the ability to support massive device connectivity. These features make 5G a cornerstone for emerging technologies such as autonomous vehicles, remote surgeries, and smart cities. However, challenges remain in terms of deployment costs, spectrum availability, and interoperability, which require further attention.

IX. Future Work

While this research demonstrates the advantages of 5G, several areas require further investigation:

- **Energy Efficiency:** Explore energy-efficient algorithms and network designs to reduce power consumption in 5G deployments.
- **Security Challenges:** Address potential vulnerabilities in 5G networks, particularly for critical applications.
- **Integration with AI:** Study the integration of 5G with artificial intelligence to optimize network management and enable predictive maintenance.
- **6G Development:** Investigate early concepts for 6G, focusing on ultra-reliable and low-latency communication (URLLC) and potential terahertz spectrum utilization.

References

[1] The ngpaas project website. [Online]. Available: http://www.ngpaas.eu/

[2] The 5gcity project website. [Online]. Available: http://www.5gcity.eu/

[3] The matilda project website. [Online]. Available: http://www.matilda5g.eu/

[4] I. Parvez, A. Rahmati, I. G¨ uvenc¸, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," CoRR, vol. abs/1708.02562, 2017.

[5] E. B. e. a. Michele Paolino, Marco Mobilio, "Deliverable 4.1 telco-grade paas: First results and implementation," 2018. [Online]. Available: http://ngpaas.eu/projectoutcomes/ngpaasdeliverables

[6] H. Harry, and T. Antti, *"LTE for UMTS Evolution to LTE Advanced"*, 2nd ed., John Wiley and Sons, Ltd., 2014.

[7] P. Nevres, S. Sargento, F. Fontes, T. M. Bohnert, and J. Monteiro, *"Quality of Service Differentiation in WiMAX Networks"*, Trends in Telecommunications Technologies.

[8] S. Ahmadi, *"An Overview of Next-Generation Mobile WiMAX Technology"*, IEEE Communications Magazine, vol. 47, no. 6, pp. 84–98, 2009.

[9] K. Thomas, *"LTE link level performance evaluation using stochastic channel models for V2X communication"*, IEEE Vehicular Technology Conference.