A New Approach for Measuring Delay in 5G Cellular Networks

David Candal-Ventureira, Felipe Gil-Castiñeira, Francisco Javier González-Castaño(✉), and Pablo Fondo-Ferreiro

Information Technologies Group, atlanTTic, University of Vigo, Vigo, Spain
{dcandal,xil,javier,pfondo}@gti.uvigo.es

Abstract. 5G networks have introduced new technologies and paradigms to support new use cases with very demanding quality of service (QoS) requirements, in terms of metrics such as delay or reliability. Accordingly, operators need new tools to measure these metrics in different realistic, often extreme, conditions, so that they can evaluate the degree of fulfilment of their service levels. In this work we propose a simple practical framework to evaluate the delay between two nodes in such a cellular network. This framework allows evaluating the delay in the uplink and downlink channels independently. We have validated the proposed framework on top of a real 5G network, by measuring the delay as a function of the requested data rates of the network in different network slices.

Keywords: 5G · Network performance analysis · Network slicing

1 Introduction

Many services requiring wireless connectivity benefit from the advantages of mobile network technologies over other radio access technologies (RATs), such as ubiquity, better performance on the move, and efficient modulation and codification management. However, some quality of service (QoS) requirements of the former are difficult to meet when a large number of users demanding heterogeneous services share the network.

5G standardization is defining architectures, technologies and mechanisms to support use cases that were unfeasible in previous generations of mobile networks. On the one hand, the performance of mobile networks has been significantly improved, with very high transmission rates, very low delays, better

This research has been partially supported by the Spanish grant PRE2021-098290, funded by MCIN/AEI/10.13039/501100011033 and FSE+, PDC2021-121335-C21, PID2020-116329GB-C21 and ED431C 2022/04, and was conducted under the framework of the project "Factory competitiveness and electromobility through innovation", with reference IN854A 2020/01, funded by the agency GAIN from the Xunta de Galicia regional government of Spain.

W. Wang and J. Wu (Eds.): BROADNETS 2023, LNICST 511, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-40467-2_1

mobility management and lower power consumption. In addition, the 5G standard introduces the possibility of creating private mobile networks. This allows end users to create their own wireless local area networks (WLANs) using mobile network technology, similarly to Wi-Fi. This kind of architecture requires, however, the operation in shared frequencies, either in industrial, scientific and medical (ISM) bands or bands regulated by governments for specific use cases such as industrial ones. As a result, private mobile networks have significantly less frequency resources than operator networks to operate. In order to support simple devices with limited power and computation resources, the concept of bandwidth parts (BWPs) allows 5G network channels to be divided into sub-channels, in pre-set time periods, so that these devices can be connected to the network by monitoring and transmitting on a short range of frequencies while the rest of the channel is exploited by devices with higher capacities.

A fundamental feature for supporting use cases with special requirements is network slicing. The network slicing paradigm allows the creation of logical networks on a common hardware architecture, which are isolated from each other and have different priority levels. In other words, network slicing allows verticals sharing hardware resources while avoiding mutual interference. This makes possible to guarantee a ratio of the network resources to a certain use case, but at the same time these reserved resources can be assigned to other slices in case they are not being used. Therefore, the network slicing paradigm supports the compliance of a certain level of performance in terms of instant transmission rate, delay, jitter and confidentiality, among other metrics, without network oversizing and allocation of dedicated isolated resources to each vertical, so that the instantaneous demands of one do not affect the others, saving costs for operators. Nevertheless, as different network slices may share a common physical infrastructure, adequate performance measurement systems are essential to operators to meet the requirements they agree with their users.

In this paper we propose a novel framework to evaluate the delay between two nodes within a cellular network. The framework allows to measure the communications delay of the uplink and downlink channels individually. We experimentally validate the proposed framework in a real 5G network setup using commercial end devices.

The rest of the paper is structured as follows. Section 2 discusses related work. Section 3 describes the framework proposed for measuring the delay between two nodes in a 5G network. The setup in which the experiments were conducted is introduced in Sect. 4. Section 5 presents the results of the evaluations. Finally, Sect. 6 concludes the paper.

2 Related Work

Previous analytical works have focused on estimating the theoretical maximum performance supported by the radio interface of a mobile network based on the physical layer characteristics defined in their standards. For example, in [4], the authors analyze the details of the 5G radio interface to compute the maximum

theoretical performance it can achieve in terms of peak rate, delay, efficiency and mobility. Similarly, in [7], the authors compare the maximum performance of the radio link for the different 5G radio configurations. In [1], a theoretical analysis of the scalability of a 5G core is presented in terms of CPU load, response time and number of requests per second from different components of the core. This evaluation focuses on the computing resources required to deploy the core rather than on network performance.

There is scarce practical work in the literature evaluating the performance of mobile networks. For example, in [9], the authors evaluate a real commercial LTE network, in particular signal quality parameters perceived by a UE in motion, and compare them to those observed when the UE is static, as well as the throughput in both cases.

Nevertheless, there exist experimental tools and research projects on experiments with mobile networks. Simulation tools such as 5G-LENA [8] and 5G K-SimNet [3] allow estimating the performance of a 5G network by taking into account parameters such as received signal power and interference between base stations. Testbeds such as that of the 5GENESIS project [13] have tested mobile networks using development tools such as software-defined radio (SDR) and open source software.

Network slicing is one of the latest innovations introduced in 5G. As a result, there is plenty of related work in the literature. Most of these articles, however, consider the problem of allocating hardware resources to slices based on demands and QoS requirements [10–12]. In [6], the authors propose a list of specific Key Performance Indicators (KPIs) for network slicing. However, these are intended only for Operations Support Systems (OSSs), for measuring the times to deploy, configure and retire a network slice instance or the efficiency of resource allocation of a slice.

One of the main difficulties when evaluating end-to-end network slicing in a 5G network is the lack of experimentation tools supporting this feature, especially at the Radio Access Network (RAN) side [2]. As a result, there is very little experimental work, mostly led by major cellular network equipment vendors. In [5], a carrier-grade Ericsson 5G testbed supporting network slicing is demonstrated. The authors configure two network slices for Unmanned Aerial Vehicle (UAV) control and data traffic communications, and they demonstrate that control communications in one slice are not affected by the transmissions in the other.

Despite being one of the main concerns of mobile operators, the evaluation of the performance of mobile networks is a field in which there is still considerable work to be done. Certainly, many approaches for evaluating the performance of wired networks or WLANs could be adapted to measure a mobile network, but these networks have particularities that should be accounted for. As previously said, next-generation mobile networks, especially 5G networks, are designed to support services with very diverse requirements and include new paradigms such as network slicing that are relevant to the performance evaluation methodology. Besides, there are substantial differences with previous technologies. For

example, cellular transmissions are scheduled, so that end users can perceive different QoS levels. An outstanding difference is the performance asymmetry of the downlink and uplink channels. They usually have widely differing amounts of resources, and, in the uplink transmission, resource grants must be notified in advance to the UE. Therefore, traditional tools to measure network delays, which compute the round-trip time (RTT) delay, are unsuitable for these networks. In this work, we present a framework for separately measuring the delay of any channel/network slice, which is disaggregated by uplink and downlink components. We are not aware of any previous work addressing this practical problem methodically.

3 Evaluation Framework and Methodology

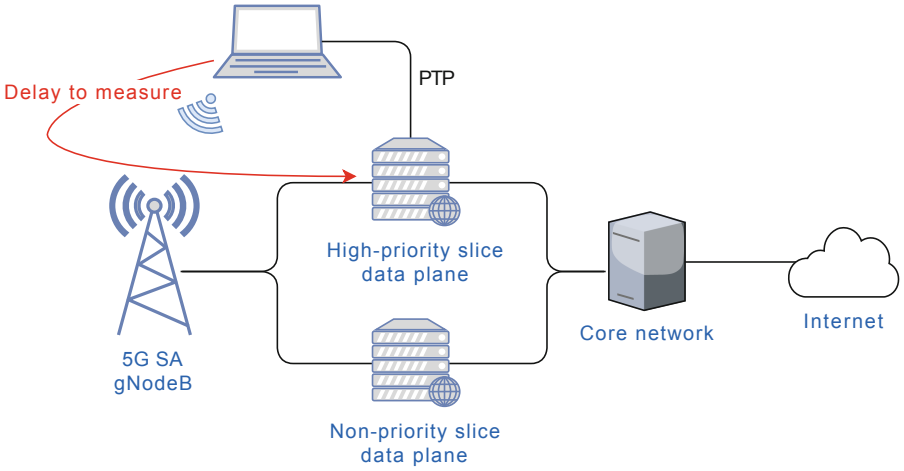

Fig. 1. Architecture of the evaluation framework to measure delays in a cellular network.

Figure 1 shows the framework that is proposed for measuring delays in cellular networks. In this framework, it is necessary to synchronize the clocks of the two nodes between which the delay is being measured. Once they are synchronized with enough precision, the delay can be measured by simply comparing the timestamps at which a series of packets where transmitted and received.

In our evaluations, we have measured the delay in the uplink channel from the UE to the User Plane Function (UPF). The architecture of the measurement testbed is shown in Fig. 1. Both the UE and the server running the UPF were directly connected through an Ethernet cable and ran the Precision Time Protocol (PTP), which has sub-microsecond clock accuracy, to get their clocks synchronized.

Even though perfect synchronization between the corresponding nodes is unfeasible, the deviation between the clocks of the UE and the UPF was three orders of magnitude lower than the values of the delay measurements in our evaluations, which validates the feasibility of our proposal.

In order to facilitate the measurement of the delay, we generated a synthetic packet trace for each scenario under evaluation, by setting the appropriate timestamps to meet the desired requested data rate. We used the well-known *tcpreplay* tool to send these packets through the 5G network from the UE.

4 Experimental Setup

The 5G Standalone (SA) network testbed that was used for the evaluation conducted in this work is composed of an Open5GS core network and a carrier-grade Release 15 indoor gNB donated by a mobile operator. The network is deployed in our premises, using commercial frequencies temporarily ceded by the operator, which currently does not use them for any service in the area. That is, we were the only users operating on these frequencies in the coverage area of the gNB. Four network slices were implemented and configured in the core network and RAN, covering the most relevant data services.

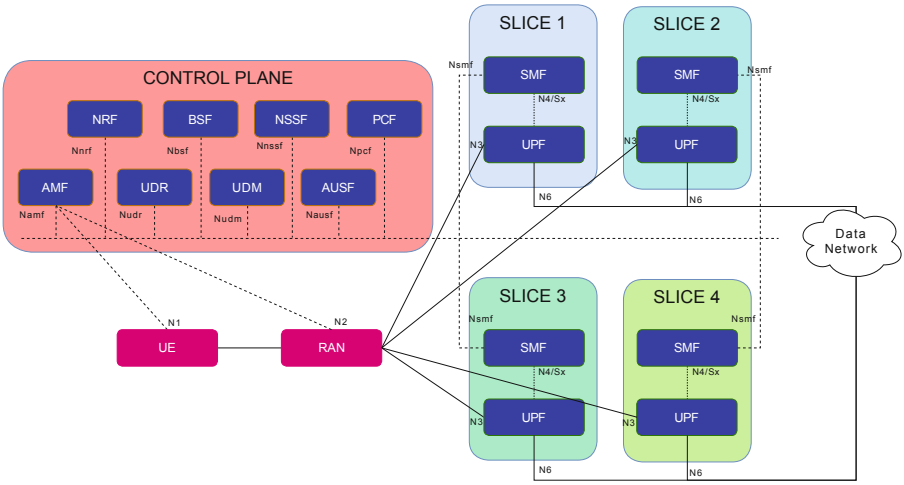

Fig. 2. Architecture of the 5G SA testbed core network.

The 5G core is deployed on an Kubernetes cloud. We isolated the control plane virtual network functions (VNFs) from the data plane VNFs of each slice by deploying them in different machines. A partition like that in Fig. 2 was chosen for its deployment. In this architecture, we have one independent Session Management Function (SMF) and one UPF for each slice, both running on a single host. The control plane host runs the shared Access and Mobility management Function (AMF), the Authentication Server Function (AUSF), the Binding

Support Function (BSF), the Network Repository Function (NRF), the Network Slice Selection Function (NSSF), the Policy Control Function (PCF), and the Unified Data Management (UDM) and Unified Data Repository (UDR) VNFs, which are shared by the four slices. Each host is a server equipped with an Intel Xeon Gold 6230 processor, with 20 cores with a maximum turbo frequency of 3.90 GHz.

The gNB is composed by an indoor carrier-grade Remote Radio Head (RRH) managed by a carrier grade Base Band Unit (BBU). These elements are directly connected through an Ethernet cable. The gNB operates in the n78 band using a bandwidth of 50 MHz with a subcarrier spacing of 30 kHz, and supports 4×4 Multiple-Input Multiple-Output (MIMO). A network switch interconnects the RRU with the data center were the core network is deployed.

The evaluation was conducted in the uplink channel, in which the network was able to provide up to 55 Mbps to Single-Input Single-Output (SISO) UEs. Two networks slices were used, for priority and non-priority services: 20% uplink resources were reserved for the high-priority slice in the RAN, whereas the non-priority slice had no reserved resources. This means that the network guaranteed at least 11 Mbps at the uplink channel. The remaining uplink resources were equally distributed within active network slices until they did not have any pending data to be transmitted. Therefore, in the worst case, when there was contention between the two slices, the high-priority slice would be allocated 33 Mbps (11 Mbps from its reserved resources and 22 Mbps owing to the equal distribution of the remaining resources between the two slices).

The UE used in the evaluation was a Quectel RM500Q-AE[1] 5G modem connected to a commodity computer. This modem complies with 3GPP Release 15 and operates in a wide range of Non-SA (NSA) and SA 5G Sub-6 GHz bands. According to its specifications, it supports 4×4 MIMO in the downlink channel and SISO in the uplink channel in band n78, but in practice it can only handle three data layers for the Physical Downlink Shared Channel (PDSCH) channel.

5 Evaluation

Table 1. Deviation between the clocks of the UE and the UPF (µs).

Mean	4.017e−05
Median	−0.004
Minimum	−9.643
Maximum	10.368
10th percentile	−0.417
90th percentile	0.401
Std. deviation	1.102

[1] https://www.quectel.com/product/5g-rm500q-ae.

Before measuring the delay of the transmissions in the 5G network, the deviation between the clocks of the UE and the UPF was computed. If the UPF clock is ahead of the UE clock, the value is positive (and negative otherwise). Table 1 shows the value of this metric during the evaluations. As shown in the table, the deviation between the clocks of the two nodes, once synchronized using PTP through an Ethernet interface, was less than 10 microseconds with an average of 0.04 nanoseconds. These values are far below the lowest delays of a 5G network, which are in the order of few milliseconds. Thus, the error introduced by the synchronization of the devices is negligible.

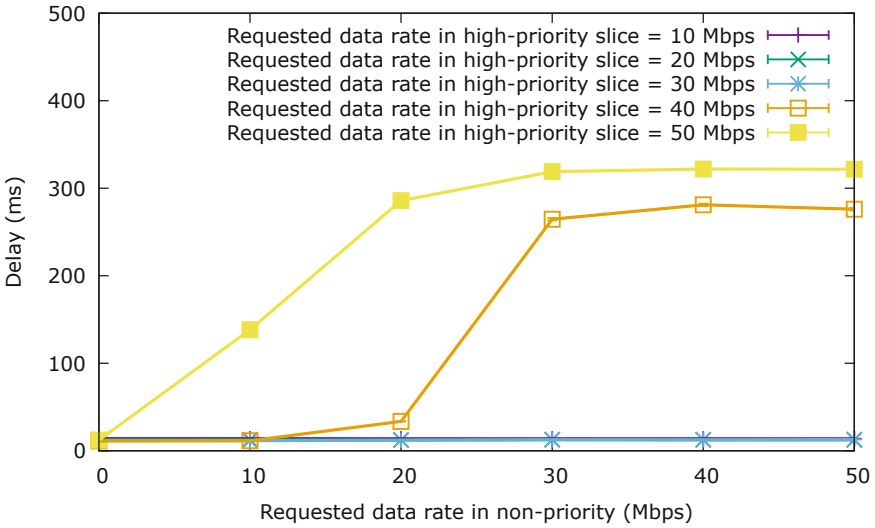

Fig. 3. Delay of the transmission of independent packets of 1,500 B on the high-priority slice versus the requested data rate on slices 1 and 2.

Figure 3 shows the observed delays in milliseconds of the transmissions of independent UDP packets with a payload of 1,500 Bytes in the uplink channel through the high-priority slice between the UE and the corresponding UPF, as a function of the data rate requested by the UE in slices 1 and 2. The figure shows that delay grows very fast when the aggregate requested data rate from both slices is greater than the uplink channel capacity, 55 Mbps, or when the requested data rate in the high-priority slice is higher than 30 Mbps. On the other hand, the average delay stays around 12 ms when the requested data rate in the high-priority slice is 30 Mbps at most, for any requested data rate in the non-priority slice.

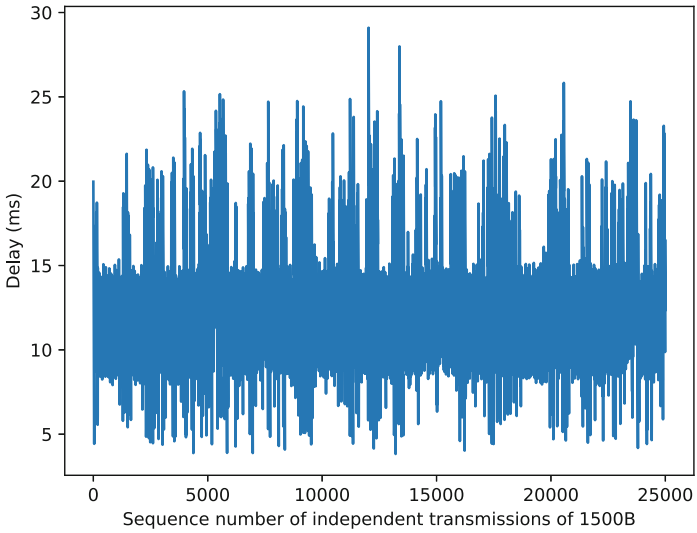

Fig. 4. Delay of the transmission of independent packets of 1,500 B on the high-priority slice with a requested data rate of 20 Mbps in the high-priority slice and 50 Mbps in the non-priority slice.

Figures 4 and 5 provide a better illustration of the cause of this behaviour. These figures show the delay, in milliseconds, of 25,000 UDP transmissions with a payload of 1,500 B when the requested data rates in slices 1 and 2 are 20 Mbps and 50 Mbps (Fig. 4) and 40 Mbps and 30 Mbps (Fig. 5), respectively. In both cases, the aggregate requested data rate is 70 Mbps, which exceeds the capacity of the uplink channel. Nevertheless, whereas in the first case the delay seems constant and has a standard deviation of 3.427 ms around an average of 12.541 ms, it increases gradually in the second, growing from tens of milliseconds to up to 426.541 ms. This is because, as the high-priority slice has 20% of the resources reserved at the RAN while the remaining resources are distributed equally between both slices and the capacity of the uplink channel is 55 Mbps, the high-priority slice receives at least 33 Mbps regardless of the requested data rate in the non-priority slice. When the requested data rate exceeds the resources that are provided to the slice, more and more packets get queued, increasing the delay.

By replicating our setup, any operator or researcher can easily measure the delay KPI of a 5G network to better characterize its operation and ensure that its slices provide the expected quality.

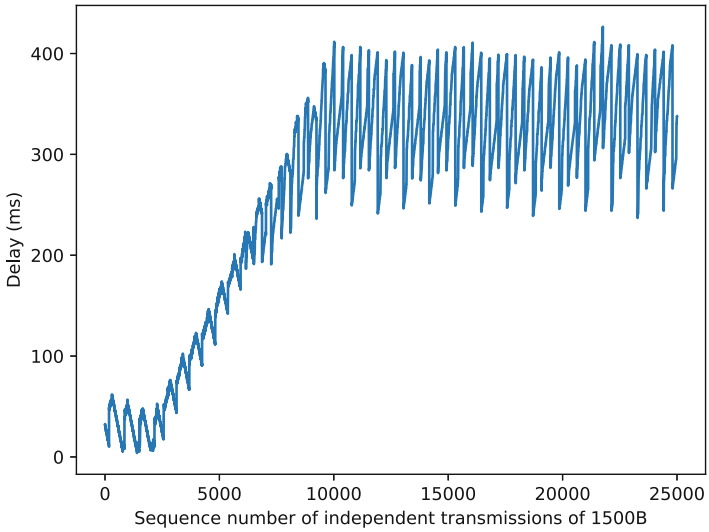

Fig. 5. Delay of the transmission of independent packets of 1,500 B on the high-priority slice with a requested data rate of 40 Mbps in the high-priority slice and 30 Mbps in the non-priority slice.

6 Conclusion

The new stringent use cases that new generation cellular networks seek to support require operators to provide differentiated traffic treatment, highly tailored to customers' needs. Operators should evaluate the compliance of their SLAs under different conditions. In this paper we propose a novel framework to evaluate delay in 5G networks. This framework is specifically suited for cellular networks, as it allows to measure the target metric in the uplink and downlink channels independently. It has been validated in a real 5G network setup using commercial end devices. Evaluation results are consistent with the expected performance of the network.

References

1. Arteaga, C.H.T., Ordoñez, A., Rendon, O.M.C.: Scalability and performance analysis in 5G core network slicing. IEEE Access **8**, 142086–142100 (2020). https://doi.org/10.1109/ACCESS.2020.3013597
2. Chavhan, S., Ramesh, P., Chhabra, R.R.S., Gupta, D., Khanna, A., Rodrigues, J.J.P.C.: Visualization and performance analysis on 5G network slicing for drones. In: DroneCom 2020, pp. 13–19. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3414045.3416208
3. Choi, S., et al.: 5G K-SimNet: end-to-end performance evaluation of 5G cellular systems. In: 2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC), pp. 1–6 (2019). https://doi.org/10.1109/CCNC.2019.8651686

4. Fuentes, M., et al.: 5G new radio evaluation against IMT-2020 key performance indicators. IEEE Access **8**, 110880–110896 (2020). https://doi.org/10.1109/ACCESS.2020.3001641
5. Garcia, A.E., et al.: Performance evaluation of network slicing for aerial vehicle communications. In: 2019 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6 (2019). https://doi.org/10.1109/ICCW.2019.8756738
6. Kukliński, S., Tomaszewski, L.: Key performance indicators for 5G network slicing. In: 2019 IEEE Conference on Network Softwarization (NetSoft), pp. 464–471 (2019). https://doi.org/10.1109/NETSOFT.2019.8806692
7. Mhedhbi, M., Morcos, M., Galindo-Serrano, A., Elayoubi, S.E.: Performance evaluation of 5G radio configurations for industry 4.0. In: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–6 (2019). https://doi.org/10.1109/WiMOB.2019.8923609
8. Patriciello, N., Lagen, S., Bojovic, B., Giupponi, L.: An E2E simulator for 5G NR networks. Simul. Model. Pract. Theory **96**, 101933 (2019). https://doi.org/10.1016/j.simpat.2019.101933. https://www.sciencedirect.com/science/article/pii/S1569190X19300589
9. Sevindik, V., Wang, J., Bayat, O., Weitzen, J.: Performance evaluation of a real long term evolution (LTE) network. In: 37th Annual IEEE Conference on Local Computer Networks - Workshops, pp. 679–685 (2012). https://doi.org/10.1109/LCNW.2012.6424050
10. Wang, G., Feng, G., Quek, T.Q.S., Qin, S., Wen, R., Tan, W.: Reconfiguration in network slicing-optimizing the profit and performance. IEEE Trans. Netw. Serv. Manag. **16**(2), 591–605 (2019). https://doi.org/10.1109/TNSM.2019.2899609
11. Wang, H., Wu, Y., Min, G., Xu, J., Tang, P.: Data-driven dynamic resource scheduling for network slicing: a deep reinforcement learning approach. Inf. Sci. **498**, 106–116 (2019). https://doi.org/10.1016/j.ins.2019.05.012. https://www.sciencedirect.com/science/article/pii/S0020025519303986
12. Xu, Q., Wang, J., Wu, K.: Learning-based dynamic resource provisioning for network slicing with ensured end-to-end performance bound. IEEE Trans. Netw. Sci. Eng. **7**(1), 28–41 (2020). https://doi.org/10.1109/TNSE.2018.2876918
13. Xylouris, G., et al.: Experimentation and 5G KPI measurements in the 5GENESIS platforms. In: Proceedings of the 1st Workshop on 5G Measurements, Modeling, and Use Cases, 5G-MeMU 2021, pp. 1–7. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3472771.3472776