OPTIMISING 5G FOR LOW LATENCY BROADCAST PRODUCTION

M. B. Waddell[1,†], J. Kröger-Mayes[1], S. R. Yoffe[2,†], D. G. Allan[2],
M. Brandstrup[3], J. Christoffersen[3], M. R. Brew[2], and R. W. Stewart[2]

[1] BBC, United Kingdom
[2] StrathSDR, University of Strathclyde and Neutral Wireless Ltd., United Kingdom
[3] TV2, Denmark
[†] Corresponding authors: mark.waddell@bbc.co.uk and sam.yoffe@strath.ac.uk

ABSTRACT

Public and private 5G networks have proved useful for electronic newsgathering [1, 2, 3, 4] where latencies of up to 2 seconds can be tolerated. 5G Networks introduce jitter and delay into the data path as consequences of the uplink scheduling process and retransmission mechanisms that manage transmission errors and packet loss. However, low latency broadcast production and audio applications need much lower delays and more stable performance. The performance depends upon the vendor implementation and its configuration. Previous projects [5] reported significant jitter spikes and concluded that 5G networks cannot readily support the 4 ms target required for audio performances requiring in-ear monitors. Tuning the 5G configuration to reduce latency helps, but there is often a trade-off between latency and spectrum efficiency.

This paper examines the collaborative efforts of BBC R&D, TV2 and Neutral Wireless to investigate and optimise 5G for low latency production in TV studio settings and multi-camera outdoor broadcasts, which in many aspects differ from newsgathering. We discuss the performance and optimisation of the scheduler, along with the latency/capacity trade-off. Achieving an 80 ms latency for video production using commercial hardware is now feasible, and we explore the potential for further reduction. 5G remains an attractive option for future bi-directional PMSE (Programme Making and Special Events) technology, and we identify areas for future research.

INTRODUCTION

News contribution workflows can tolerate video delays of up to 2 seconds and mobile connectivity (3G/4G/5G) is in widespread use. A single camera typically bonds over several public mobile networks to provide a contribution. The long latency prevents a two-way conversation between the studio and the news event, but this can be managed. However, in many broadcast workflows, there are several data-paths with strict latency requirements. Presenters or other stage talent often use wireless microphones and in ear monitors (IEMs) that must achieve extremely low levels of latency. This is particularly critical for musicians where the microphone to IEM latency must be minimised for a successful performance. Previous work suggests a latency target of 4 ms is required [1, 2]. Audio PMSE systems in current use, such as radio mics and IEMs, have one-way latencies under 3 ms.

For production talkback, used to communicate between a studio gallery and the production staff, a high-quality digital intercom system is required. Existing systems offer a latency below 40 ms to a wireless belt-pack and <70 ms from belt-pack to belt-pack. This is typically now achieved using Digital Enhanced Cordless Telecommunications (DECT) technology.

Real-time camera control operations that rely on feedback to the operator, such as camera shading, focus, or pan-tilt-zoom (PTZ) control also demand low latency. When the delay from an operator providing an input to the perceived response exceeds about 100 ms their accuracy and speed in making adjustments is reduced. Beyond 300 ms the operator will experience a significantly reduced sense of control [5] and as delay increases further it is eventually necessary to 'adopt a "move and wait" strategy' [6], which is not suitable in a fast-paced live broadcast environment. This end-to-end path includes the 5G network delay, but usability is usually limited by waiting for visual feedback (video feeds). The measure of latency that includes the video capture and presentation delays in addition to any signal transmission is often referred to as 'glass-to-glass latency', while the same term is often confusingly used to describe just the delay between the encoder input and decoder output. Since different cameras and particularly displays introduce varying levels of delay, this paper will concentrate on the latter.

A growing area of interest is remote production, where the production gallery is not located at the broadcast event. This is increasingly popular to reduce costs and environmental impact. Remote vision mixer operators can tolerate a higher delay of over 600 ms from camera to vision mixer, provided that return feeds to the remote event are not required. Camera operators must adopt an 'always on air' philosophy and respond to talkback directions to cope with the absence of accurate tally signalling.

MEASUREMENT OF UDP TRANSIT TIME FOR NETWORK CHARACTERISATION

Time sensitive applications, such as audio and video streams, typically use the User Datagram Protocol (UDP) at the network transport layer. This enables low latency communication because there is no requirement to acknowledge data between the sending and receiving application. The Real-time Transport Protocol (RTP) [7] is one example where UDP is used for media streams. Transport protocols may also provide additional buffering, forward error correction and/or packet retransmission at the application layer.

The transit time of UDP traffic across a 5G connection and its statistical variation are key in determining the latency performance of a 5G network. Buffering is required to accommodate any jitter in the transit time and ensure that a media decoder is fed with a steady stream of data. The experimental configuration used by the authors is shown in **Figure 1**. The arrangement shows a User Equipment (UE) device that connects to a traffic server over a 5G Non-Public Network (NPN). The sender and receiver would correspond to the encoder and decoder, respectively, in a real broadcast application.

The iperf3 tool is used to establish a steady stream of UDP uplink traffic via the 5G system for transit time analysis. The UDP traffic is captured to file using a packet capture tool at both the server and the UE sender. To ensure captured packets are timestamped with synchronous clocks, Network Time Protocol (NTP) is used over a separate wired Ethernet connection, ensuring a low jitter path for the NTP traffic. The latency introduced by the 5G network for each UDP packet can be calculated by comparing the times at which the packet was sent and received. A typical traffic session runs for 300 seconds and the captured UDP packets are stored to disk. By comparing the received packet data with that sent, the transit time across the 5G network can be computed. The variation in transit time can be analysed statistically to give a performance metric for the 5G system and inform achievable video system latencies.

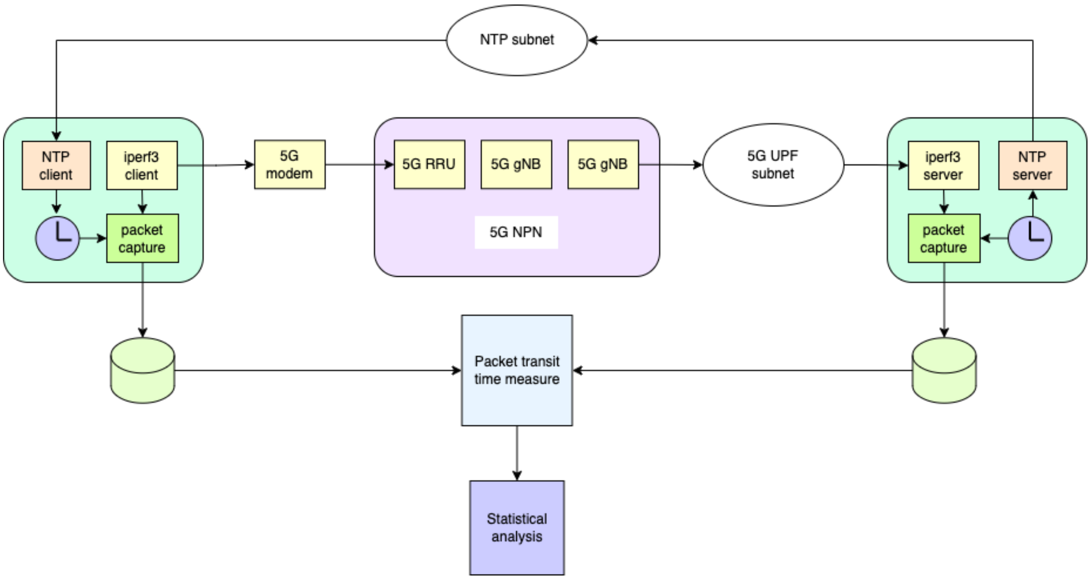

Figure 1: Experimental arrangement for testing UDP packet transit time.

LATENCY AND JITTER RESULTS

In this investigation, four different commercially available 5G NPNs were tested using the same methodology: a system using an 80 MHz channel (n78) configured for downlink; a system configured for low latency using a 100 MHz channel (n77u); and two software-defined radio systems using various configurations also with 100 MHz (n77u) channels. The n77 band spans 3.3-4.2 GHz, with the lower portion 3.3-3.8 GHz also designated n78. The upper portion (n77u – 3.8-4.2 GHz) is available for shared access licensing in the UK.

A typical uplink-biased 5G network with a Single Input Single Output (SISO)-connected UE using a Data Radio Bearer (DRB) with Radio Link Control (RLC) operating in Acknowledged Mode (AM) exhibits the following transit time characteristics:

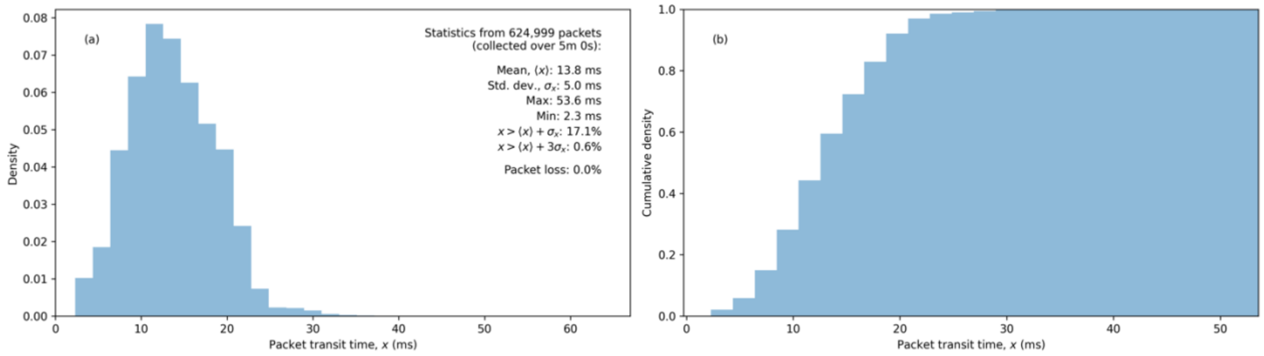

Figure 2: Packet transit times for a typical 5G connection configured for UL-biased traffic.

Three vendor implementations were tested and the 5G configurations were adjusted where possible. The uplink latency for traffic at 20 Mb/s (unless otherwise stated) was measured and the results are summarised in **Table 1**.

The observed results depend on vendor implementation, Time Division Duplexing (TDD) frame structure, low latency optimisations, target frame error rates and available Modulation and Coding Scheme (MCS). The buffer depth of any 5G-connected video decoder must be adjusted to accommodate the longest packet delays present in the network. In this case of **Figure 2**, the video latency can never be shorter than 55 ms for a perfect video codec with no delay. In practice, a further 50 ms of latency is typical for low latency H.265 codec implementations, resulting in a total video delay (encoder input to decoder output) of around 105 ms. **Table 1** shows how low latency optimisations can be made to reduce this maximum transit (and end-to-end) latency, but usually at the expense of capacity.

Table 1: Summary of results.
(DL = downlink-biased, UL = uplink-biased, LL = scheduler optimised for low latency, TFER = Target Frame Error Rate)

	Configuration notes	Mean (ms)	Std. Dev. (ms)	Min. (ms)	Max. (ms)	Uplink Capacity (Mb/s)	Packet Loss (%)
1	DL, 5 Mb/s	12.4	3.7	3.0	32.5		0
2	DL, 10Mb/s	12.2	3.8	3.3	43.7		1.7
3	DL	12.9	3.2	3.9	46.3		16.9
4	Balanced, LL	6.5	1.4	3.3	28.0	200	0
5	Balanced, LL, 50 Mb/s	7.6	2.3	3.3	25.4	200	0
6	UL	13.8	5.0	2.3	53.6	400	0
7	UL, short TDD	12.9	3.8	2.3	38.2	375	0
8	UK, short TDD, LL	8.1	1.5	2.5	26.1	340	0
9	UL, short TDD, TFER=0	8.0	1.4	2.8	19.5	275	0
10	UL, short TDD, LL, fixed MCS 6	6.4	1.7	2.3	10.2	100	0

CAUSES OF LATENCY AND JITTER IN 5G NETWORKS

In 5G New Radio, spectral resources are split among UEs into Physical Resource Blocks (PRBs) and allocated on a per time slot basis. A UE with data to send needs to request resources from the scheduler. The UE sends a scheduling request (SR) and is allocated a small initial uplink grant to provide its buffer status report (BSR), which tells the scheduler how much data is waiting to be sent. These resources are then granted, and the data eventually transmitted. The SR and BSR reporting mechanism is often quite time consuming and can lead to jitter, especially if it must be done repeatedly during a single transmission. The impact can be minimised by appropriate choice of TDD pattern and optimising processing delays. We note that the BSR mechanism is not ideal for live streaming [8].

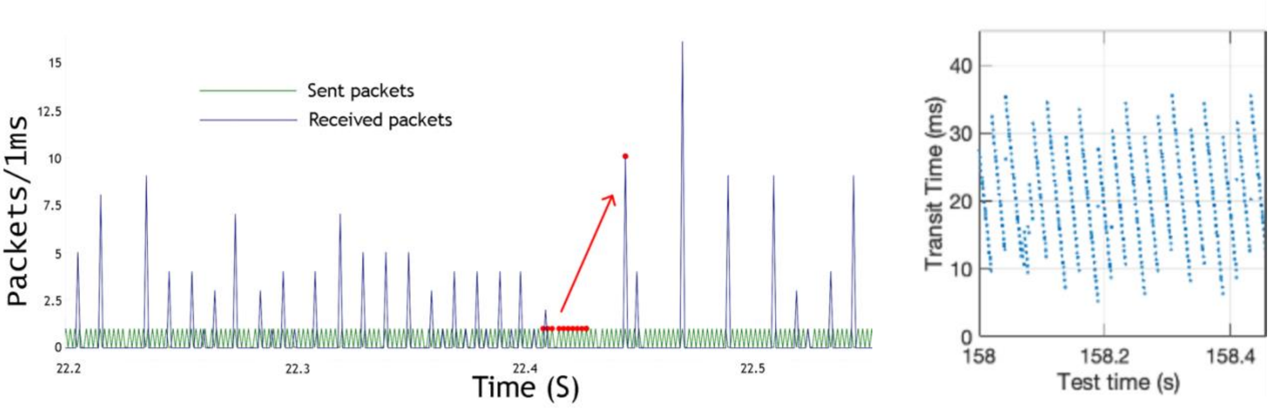

Figure 3: (Left) Example packet transmission and receive times.
(Right) Variation in transit time dues to TDD scheduling.

The interplay of these processes in the scheduling of packets imparts jitter on to the measured packet transit time, leading to a range of values as some packets have to wait longer than others for available resources to be transmitted. The left-hand panel of **Figure 3** shows that regularly spaced UDP packets generated by iperf3 and sent by the UE (green) are received in batches by the server (blue). Tagging a set of 10 packets (red dots) we can see they are created and pass out the device network interface to the modem at a regular interval but have to wait for uplink resources before transmission and are then received together. This leads to a striation or "saw tooth" pattern of values observed in the packet transit times (right-hand panel), since the earliest (first) packet into the modem has to wait the longest.

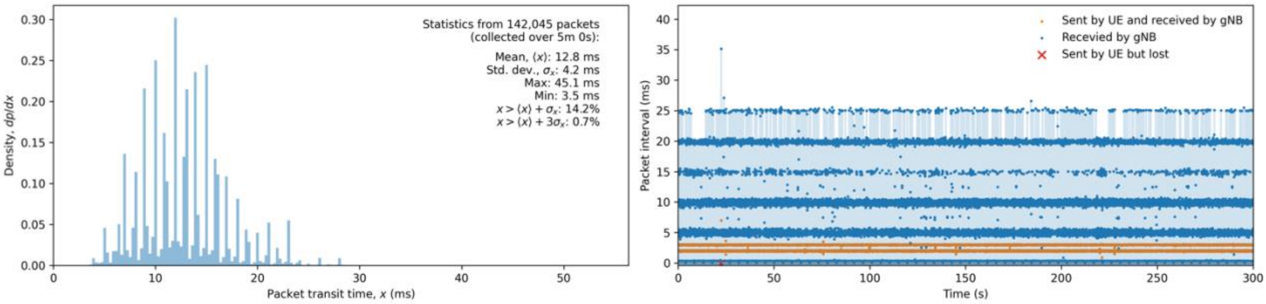

Figure 4: (Left) Discretised transit time distribution. (Right) Received packet interval showing the discrete "banding" of values due to retransmissions.

The left panel of **Figure 4** shows the discontinuous nature of packet transmission. Since the packets must occupy discrete TDD slots, transit times cannot form a continuous distribution. This is exacerbated by the retransmission mechanism to correct for radio transmission errors.

At the Medium Access Control (MAC) layer, the so-called Hybrid Automatic Repeat reQuest (HARQ) mechanism (which combines high-rate Forward Error Correction (FEC) and ARQ) is used to ensure robust delivery of data to the application layer. The retransmission process introduces jitter as additional time above the baseline is needed to successfully transmit the packet. Moreover, there is a probability of multiple transmission errors, whereby the first retransmission is also errored triggering the need for further retransmissions and a further increase in the total packet latency. Placing a sensible upper limit on the number of HARQ retransmissions can ensure jitter is kept low at the risk of errors propagating to higher layers.

For a bearer operating in Unacknowledged Mode (UM), unsuccessful HARQ leads to packet loss. These can often be recovered at the application layer, but not for low latency applications with simple UDP transmission. In this case, an AM bearer (which adds "slow ARQ" at the RLC layer) is preferred.

The modulation is continually adjusted according to the observed Frame Error Rate (FER) of the radio channel. A typical target FER value is 1%, which results in a steady "rumble of HARQ retransmissions". If the radio frequency (RF) channel degrades, the FER increases and the MCS control loop responds by reducing the complexity of the modulation to lower the FER to the target. Should the channel improve, the FER will fall and the MCS loop will select a less robust modulation to improve the capacity of the channel. In this way the system optimises the capacity of the channel in the presence of noise, interference and fading.

The right-hand panel of **Figure 4** presents the measured delay between received packets using configuration 1 in **Table 1**, and nicely shows the retransmission process at work. The packets with almost zero delay arrived together as a bunch, as seen in **Figure 3** (left). There is then a gap of ~5 ms (which is related to the system's RX to TX latency, TDD frame structure and period, and k-values) before the next round of packets are delivered. However, if there is a transmission error, then we have to wait this period again, and the band of packets at ~10 ms were instead successful on the first retransmission. (Similarly, the bands at ~15 ms, ~20 ms and ~25 ms correspond to two, three and four retransmissions, respectively.) The number of allowed retransmissions can be implementation dependent.

SCHEDULER AND LOW LATENCY OPTIMISATION

The 5G scheduler grants resources to attached terminals according to the BSRs that are periodically received from the UEs. Capacity on the channel is then shared between the connected terminals that have buffered traffic ready to send. Idle terminals do not require capacity and can remain attached without overhead. In this way the radio capacity can be dynamically allocated on an as-required basis and shared fairly, efficiently releasing spectral resources when not required.

In streaming applications, the UE buffer is constantly being “topped up" with new data, so in order to continue to be granted uplink resources it needs to periodically update its BSR. Reducing this period can help reduce latency by ensuring resources continue to be scheduled and preventing gaps in transmission.

There are several optimisations that can be made to reduce latency and jitter introduced by the scheduling procedure and retransmission mechanism:

- **Shorter TDD frame structures (configuration 7 in Table 1)**
  An example uplink-biased frame structure uses 2 DL slots and 7 UL slots [4]: **DDSUUUUUUU** (where S is the special slot that covers the transition from downlink to uplink). This pattern uses 10 slots so has a period of 5 ms. An uplink grant request on the first **U** slot has to wait until at least the next **D** slot to be allocated (3.5 ms) and then for a suitable uplink slot (at least 1 ms). Whereas, the shorter **DSUU** pattern only has to wait at most 1 ms to receive the grant then 1 ms for the uplink slot. This affects both latency and jitter caused by retransmissions. Note: k-values can be used to fine-tune scheduling.
- **Report BSR more regularly (configuration 8 in Table 1)**
  Ensures resources continue to be scheduled for incoming packets topping up the UE’s buffer.
- **Reduce the target FER (configuration 9 in Table 1)**
  The allowed steady rumble of transmission errors allows the MCS control loop to attempt to maximise cell capacity. Choosing a lower target FER causes the MCS to remain more conservative to minimise retransmissions. This is at the cost of cell capacity. Note that even with a target FER of zero, retransmissions will occur as the control loop attempts to improve spectral efficiency.
- **Fixed MCS (configuration 10 in Table 1)**
  One final way to reduce latency and jitter is to fix the MCS at a low value. This forces the UE to use an inefficient but robust modulation, significantly reducing capacity but also greatly reducing the probability of retransmissions.

NETWORK SLICING

5G supports network slicing to control the bandwidth and quality of service to attached terminals. It is possible to provide terminals connected to broadcast cameras with guaranteed capacity to ensure that low priority connections (*e.g.* Internet browsing) do not disrupt video traffic. Most discussion of 5G SA slicing focuses on public mobile networks, potentially offering broadcasters priority resources. This is becoming available in several countries as 5G SA hardware rolls out. Indeed, the ability to have a local user plane deployed does mean that these connections can start to meet lower latency demands. However, their cost and viability are not yet clear. While we can certainly envisage a public slice as being ideal for low density electronic newsgathering cameras, their use for low latency multi-camera outdoor broadcasts needs to be explored further. We should note that 5G SA NPNs can also offer network slicing, providing higher-priority resources to particular UEs (such as camera encoders) as well as increasing priority for particular traffic types or endpoints. We acknowledge the huge potential for using a public slice to backhaul an NPN production.

CONCLUSIONS

Our single-cell results indicate that typical 5G NPN implementations necessitate a UE buffer of up to 50 ms to ensure error-free video transmission to a decoder. The total video latency includes this buffer depth and the latency of the video codec, amounting to approximately 100 ms in typical systems. Low latency optimisations allow for 80 ms to be achieved with existing commercial hardware, and this could be reduced further by faster codecs. 5G standalone networks can be used to support IP-based talkback, camera control and tally systems alongside low latency video [9].

Cell handover introduces another source of jitter due to the scheduling gaps required to perform neighbouring cell Reference Signal Receive Power (RSRP) measurements, and latency with the inherent interruption to data transmission while the device moves from the serving cell to the target cell. In our experience, the latter lasts around 20 ms and additional buffer overhead is necessary to cover this period.

Although this latency is higher than that of commercial PMSE systems based on DVB-T COFDM technology (which is typically around 50 ms), it is still suitable for live TV production. The TV2 breakfast studio in Copenhagen has successfully utilized 5G-attached cameras and private 5G SA. Feedback from production staff (including sound engineers, directors, and the Technical Operation Manager) has been positive, with no noticeable degradation in performance. 5G remains an attractive technology for future wireless PMSE and integrates naturally into modern IP workflows using cloud technology.

Note that interlaced video is increasingly unsupported in modern codecs; frame periods and latency are always higher than for progressive. In this paper, it is assumed that future video production will use progressive. The increased frame rate reduces frame period and latency.

The maximum transit times of 20-50 ms observed are not really adequate for critical audio applications, such as musician foldback using IEMs. For such applications, conventional PMSE solutions with single-carrier radio mics and analogue Frequency Modulation (FM) IEMs will continue to be used. The opportunity for optimising 5G to achieve a transit time below 5 ms is unfortunately limited given the fundamental TDD scheduling process and RX to TX system latencies. However, for less critical audio applications there is an opportunity to replace current equipment with a 5G NR solution.

RECOMMENDATION FOR FURTHER WORK

Tests on different vendor implementations highlighted remaining issues when connecting 5G-capable devices to 5G NPNs. Devices that connect readily to one implementation sometimes fail to connect on a different vendor’s base station. Given a valid SIM and the correct Access Point Name (APN), devices should readily connect but quite often do not. Rebooting the UEs and even restarting the 5G system can resolve the issue, but this is not always the case. Vendors need to address this issue.

Several potential latency-optimisation strategies have been identified that require complex configuration of the 5G base station and expert knowledge. No single configuration gives perfect performance for all applications. A set of low latency profiles to assist production staff would be helpful. Such profiles might include:

1. *Best capacity – This would include MIMO operation on the uplink, access to 256-QAM tables for transmission, optimum target FER to suit the performance of the system and would typically be 1% FER.*
2. *Best latency performance – This would include a shorter TDD pattern, reduced target FER (e.g. 0.01%), SISO operation and 64-QAM modulation tables.*
3. *Fixed MCS mode – This would disable the MCS loop and might be appropriate for low-bitrate latency-critical audio applications such as radio microphones.*
4. *An improved BSR mechanism for live streaming or pre-allocation of resources.*

The development of such profiles could be further developed by a 5G user working group. With later releases of the 5G standard, we look forward to the possibility for PTP, which will also make it possible to synchronise cameras and encoders, reducing latency still further.

ACKNOWLEDGMENTS

The authors would like to thank the operational staff at TV2 for access to the Tivoli studio in Copenhagen. We would like to thank Dr. Shuja Ansari and the staff at the University of Glasgow for access to their 5G NPN. We are also grateful to the StrathSDR team at the University of Strathclyde.

REFERENCES

[1] D. Gomez-Barquero *et al.*, “5G-RECORDS: 5G key technology enablers for emerging media content production services” (https://5g-records.eu).

[2] D. Gomez-Barquero *et al.*, “5G-RECORDS: Deliverable 2.1 Use Cases and KPIs” (https://www.5g-records.eu/Deliverables/5G-RECORDS_D2.1_v2.0_web.pdf).

[3] S. R. Yoffe *et al.*, “Pop-Up 5G Standalone Non-Public Networks (SNPNS) For Live Broadcast Production,” *NAB BEITC,* 2023 (https://nabpilot.org/product/pop-up-5g-standalone-non-public-networks-snpns-for-live-broadcast-production/).

[4] M. Waddell *et al.*, *IBC Technical Papers,* 2023 (https://www.ibc.org/technical-papers/ibc2023-tech-papers-5g-standalone-non-public-networks-modernising-wireless-production/10246.article).
S. R. Yoffe *et al.*, *NAB BEITC,* 2024 (https://nabpilot.org/product/using-a-private-5g-network-to-support-the-international-broadcast-of-the-coronation-of-hm-king-charles-iii/).

[5] A. Nankaku *et al.*, “Maximum acceptable communication delay for the realization of telesurgery,” *PLoS ONE* 17(10),e0274328, 2022 (doi: 10.1371/journal.pone.0274328).

[6] B. Berberian *et al.*, “Data Transmission Latency and Sense of Control,” *Engineering Psychology and Cognitive Ergonomics,* 2013 (doi: 10.1007/978-3-642-39360-0_1).

[7] H. Schulzrinne *et al.*, “RTP: A Transport Protocol for Real-Time Applications,” *IETF RFC 3550,* 2003 (https://datatracker.ietf.org/doc/html/rfc3550).

[8] F. Ronteix-Jacquet, “Reducing latency and jitter in 5G radio access networks,” 2023 (https://theses.hal.science/tel-04012662).

[9] Neutral Wireless and QTV, 2024 (https://www.neutralwireless.com/2024/05/tay-5g-showcases-private-5g-for-live-football-broadcasting/).