Journal of Engineering and Science Research 7 (1): 48-57, 2023
e-ISSN: 2289-7127
© RMP Publications, 2023
DOI: 10.26666/rmp.jesr.2023.1.8

Survey of Enablers to Overcome Latency in 5G and Beyond Radio Access Networks

Mohammad Emad Alolaby, Rudzidatul Akmam Dziyauddin, Norulhusna Ahmad

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia (UTM), Malaysia
*Corresponding Author: rudzidatul.kl@utm.my

Received: 05 December 2022; Revised: 25 December 2022; Accepted: 20 January 2023; Published: 20 February 2023

Abstract: One of the most important 5G requirements is to minimize latency as low as 1ms within the network for new delay-sensitive applications, such as traffic efficiency and safety, and remote control of machinery besides satisfying higher date rate and high number of connected devices. This paper reviews major contributors include Hybrid Automatic Repeat Request (HARQ) retransmissions, current Transmission Time Interval (TTI) length, coding and legacy scheduling mechanisms for low latency purpose. To address latency challenge, each contributor has to be independently considered and can be either minimized or eliminated by defining some important latency sources for air interface. In addition, it suggests some technical enablers or existing recent algorithms towards overcoming latency issue. We also discuss some key challenges and future works in supporting the latency requirements for next generation networks.

Keywords: *Massive MIMO (MMIMO), Latency, Scheduling, Ultra-reliable machine-type communications (uMTC)*

1. Introduction

5G services can be categorized in three generic services: extreme mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) also called ultra-reliable machine type communications (uMTC) [1]. uMTC applications require much lower latency besides very higher reliability compared to today's communication systems. Important example cases are real-time remote computing for moving terminals, teleprotection in smart grid network, and traffic efficiency and safety [2].

In Long Time Evolution (LTE) networks, end-to-end round-trip time (RTT) latencies between user equipment (UE) and server for internet access case, are around 60 ms (10 - 20 ms in laboratory environment). The pure Radio Access Network (RAN) user plane one trip time (OTT) latency is about 4 ms for Frequency Division Duplexing (FDD) and 4.9 ms for Time Division Duplexing (TDD) [3]. To fundamentally support these new kind of delay sensitive applications, 5G networks need to deliver 5-10 times reduced latency compared to LTE. This means end-to-end latency of less than 5 ms and air latency (RAN air interface part only) of less than 1 ms [4].

In the literature, couple of surveys on latency reduction are available, each focuses on some contributors and certain technology enables. For example, in [5], different ways are reviewed in the domain of RAN, cashing and core network for achieving lower latency. This includes digital beamforming for control, flexible TTI, software-defined networking (SDN) and network functions virtualization (NFV). Nevertheless, the paper does not make enough

Corresponding Author: Rudzidatul Akmam Dziyauddin, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia (UTM), Malaysia. Email: rudzidatul.kl@utm.my

attempt to overcome some major contributors to latency. Such as, the reduction of Automatic Repeat Request (ARQ) retransmissions. It just refers to the importance of shorter frames to reduce the bad impact of ARQ retransmissions without offering a solution to reduce retransmissions. Also, while the paper highlights the importance of beamforming, nevertheless, it limits the usage to a switched architecture system where control signals are sent using low-resolution digital beamforming with analog beamforming in the data plane. The usage of analog beamforming has the challenge of limiting the transceiver to communicate in only direction at a time. It would have been useful if they adopted more newly novel approaches such as hybrid beamforming.

In [6], they surveyed some challenges and enablers for delivering end-to-end, reliable, low-latency services in mmWave mobile systems. Special focus has been paid on congestion control, Medium Access Control (MAC) layer, and core network architecture. For each of the three areas, they have discussed current and possible innovative solutions, and some example results that show the potential of the different techniques for the reduction of the overall delay. However, it would have been more comprehensive if the author had considered other contributors such as HARQ retransmissions. Also, the paper highlights the importance of the usage of beamforming techniques without giving further details on how they can be utilized and what is the expected impact.

In [7], a holistic analysis and classification of the main design principles and enabling technologies that will make it possible to deploy low-latency wireless communication networks were done. They investigated how the delay adds from PHY to transport layer, and they showed how to divide the end-to-end delay into multiple components. Then, they discussed how different techniques may leverage one or multiple delay components. Although the study is relatively comprehensive, however, it is generic and not focused on 5G. Also, it misses or does not pay enough focus on some contributors such as, TTI length, HARQ and on some potential technology enablers such as massive multiple-input multiple-output (Massive MIMO or shortly MMIMO). Actually, the paper states that the usage of MMIMO may not be applicable for low latency services and consider this as a largely open problem.

In this paper the reduction of air latency for delay sensitive applications in 5G will be addressed. Our attention focuses on major contributors to radio latency problem. We discuss four major contributors in current 4G systems, namely HARQ retransmissions, TTI length, channel coding, and legacy scheduling mechanisms. Our contribution lies in the analysis of some technologies and schemes that could be enablers to overcome latency challenge for the 5G RAN systems or beyond. To the best of our knowledge, these contributors and enablers are either not covered or not detailed enough in other similar surveys.

The rest of this paper is organized as follows: Section 2 details major contributors to air latency along with possible enablers to overcome them. Section 3 highlights some technologies and schemes proposed in literature. Finally, conclusion is drawn in Section 4.

2. Latency Sources and Possible Solutions

There are multiple contributors to the end to end latency of current LTE systems. They can be divided into two major categories: latency resulted from air interface nature and design, and the latency resulted from packets flow along network structure [8]. Considering the last, a major part of the latency comes from the core and transmission networks of the system. So having an optimized architecture is necessary to achieve 5G latency target. Localized contents, caching, localized traffic flows and device-to-device (D2D) communications could be key enablers to attain that target [9], [10]. Giving further details about latency because of network structure is beyond the scope of this paper as it is focusing more on air interface contributors.

With regard to latency inherent in air interface, the wireless channel can be characterized as fast and varied over time. The most important reason of these variations is due to fading that significantly affect the average received signal power. Accordingly, fading leads to increased risk of temporary outages and packet losses, that makes it hard to build low latency radio links. Fading can be handled either by simply compensating with a fading margin added to the SNR, diversity, or by further complicated techniques to manage the disparities in the instantaneous radio-link quality, including link adaptation, scheduling, and error detection and correction mechanisms, such as HARQ. Despite the importance of these enablers for reliability, some of them do it as a trade off on latency. Briefly, there are multiple contributors to air interface delay, such as HARQ retransmissions, current transmission time interval (TTI) lengths, channel coding, and legacy scheduling. To satisfy 5G latency requirement these contributors have to be considered one by one and either minimized or eliminated. In the following sections these contributors are discussed along with promising technologies to overcome their challenges.

2.1 HARQ retransmissions

As mentioned earlier, wireless channels suffer from fast and significant variations due to multiple reasons. The most important reason is fading which significantly affects the average received signal strength. Accordingly, fading leads to increased risk of temporary outages and packet losses, that makes it hard to build low latency radio links. Fading can be handled either by simply compensating with a fading margin added to the signal-to-noise ratio (SNR), diversity, or by further complicated techniques, including link adaptation, scheduling, and error detection and correction mechanisms, such as HARQ. The variations can be partially overcome prior to data transmission through link adaptation and scheduling. However, a perfect link adaption is unlikely feasible due to the random variations in the link quality mainly from Rayleigh fading, interference, and receiver noise, perfect adaptation is not feasible. Therefore, all wireless systems utilize a sort of forward error correction (FEC). Briefly, FEC coding relies on additional bits, named as parity bits, appended to the information bits before transmission. Another method to deal with errors is to employ automatic repeat request (ARQ). The receiver uses an error-detection code often a cyclic redundancy check (CRC) to determine the integrity of the received packet. If there is no error, the received data can be said error-free and a positive acknowledgement (ACK) is notified to the transmitter. Otherwise, if there is an error detected, the receiver discards the received data and notifies the transmitter by sending a negative acknowledgement (NAK) via a feedback channel, so that the data retransmission can be triggered. Most feasible hybrid ARQ schemes selected are a CRC code for error detection while convolutional or turbo codes for error correction despite that many types of error-detection and error correction can also be an option. In principle, all modern communication systems use hybrid ARQ (HARQ) which is a combination of ARQ and forward error-correction coding. So, scheduling, link adaptation and HARQ are mechanisms to overcome instantaneous variations in radio-link quality and complement each other. While scheduling and link adaptation work before the transmission of data, HARQ works after transmission [11].

Despite its importance for reliability and spectral efficiency, HARQ imposes a major challenge considering 5G latency requirements. Actually, HARQ is a major contributor to the end to end delay. The LTE user-plane (U-plane) one way latency for a scheduled UE consists of the fixed node processing delays (which includes radio frame alignment ~0.5 ms [12]) and 1ms TTI duration. Figure 1 shows U-plane delay components for LTE FDD, which can be expressed in ms as [13]:

$$T(n)\,[ms] = 1.5 + 1 + 1.5 + 8 \times n = 4 + 8 \times n \quad (1)$$

where n is the total number of HARQ retransmissions.

This highlights the major contribution of HARQ retransmissions to end-to-end latency, the major question is how to minimize these re-transmissions? Following is a brief discussion about some enablers.

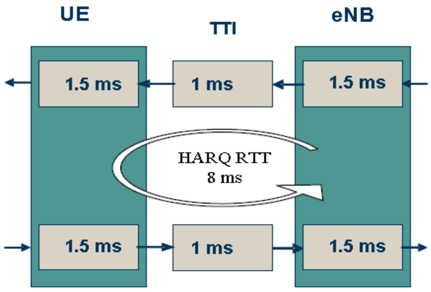


Figure 1. User-plane delay components of LTE FDD in 3GPP standards

2.1.1 Diversity

Significant signal fluctuations resulted from fading increases the chance of temporary outage, which consequently increases packet errors and losses. One way to compensate this is by adding a fading margin to the average SNR. However, in order to achieve high levels of reliability of $1–10^{-5}$ to $1–10^{-9}$, a significant margin of 50 dB – 90 dB needs to be added. Alternatively, adding diversity to the transmission provides robustness against fading losses and consequently can reduce needed fading margin. With a diversity order of 16, the fading margin for a reliability of $1–10^{-9}$ can be reduced by 90% (from 90 dB to 9 dB). Diversity can be realized in the dimensions of time, frequency and space. Considering we are targeting low-latency, time dimension cannot be utilized. Space could be one of the most attractive option considering the emerging technology of MMIMO. In the last few years, MMIMO with its beamforming capability has proved to be a promising technology to enhance wireless system performance, especially in capacity increase and interference reduction [14]. MMIMO is similar to traditional MIMO systems but with much higher number of antenna elements. The number of base station (BS) antennas should be significantly larger than the number of users (usually minimum 8-10 fold). This allows the system to utilize beamforming to communicate with multiple users simultaneously in the same time frequency resources, which has positive reflection on system performance and capacity. Multiple researches have been done to highlight MMIMO advantages on capacity, spectral efficiency, and HARQ retransmissions minimization like [15], [16], [17], [18]. MMIMO relies on the law of large numbers and beam-forming in order to avoid fading dips, so fading no longer affects latency.

5G wireless systems use millimeter wave (mmWave) bands to gain from their wider available bandwidth. To overcome the severe propagation loss in the mmWave band, 5G systems use MMIMO with large antenna arrays. Compared to the current wireless systems, the wavelength in the mmWave band is smaller enabling the employment of huge sized arrays at the BS and consequently a major improvement in MMIMO advantages [14], [15].

To get targeted improvement in performance from MMIMO, there is a need for perfect channel state information (CSI) at the BS. This is a challenging factor, as it needs major overhead especially when talking about mm-Wave frequencies, which have very short coherence time compared to legacy radio channels. Practically, in TDD systems, CSI is calculated from uplink training according to channel reciprocity [14]. Theoretically, a training sequence needs to be sent on each channel, that is, on each transmit – receive antenna pair. It is even worse in FDD systems, due to the fact that these systems cannot rely on channel reciprocity. In these systems, there is need for extensive overhead for downlink training and uplink feedback. In addition, hardware is another challenging factor to be considered. The mm-Wave transceivers are quite costly with high power consumption and accordingly it is not practical to build a complete radio frequency (RF) chain for each antenna element [19] [18]. These two challenging factors triggered the research for agile methodologies. One of the most promising methodologies is to use hybrid digital/analog (HDA) beamforming structure first proposed in [20]. In this approach, hybrid transceivers uses a combination of analog beamformers in the RF and digital beamformers in the baseband domains, with fewer RF chains than the number of transmit elements. Hybrid beamforming is new and hot research topic and multiple researchers have started introducing variant novel schemes to actualize it, such as the one in [21].

2.1.2 Conservative Channel Estimation

A general design choice to have reliable low-latency design is to make adaptive transmissions, which is dependent on channel estimation though CSI [10]. Generally, bad estimation will lead to errors. More knowledge about CSI by applying conservative channel estimation will improve system performance, consequently minimizing HARQ retransmissions.

Considering that MIMO will definitely be part of 5G system, we will talk about MIMO channel estimation methodologies. MIMO channel models can be classified in different ways. One useful model classification is based on the modeling approach taken. It divides the models in physical models and analytical models. Analytical models characterize the channel in a mathematical/ analytical way without explicit consideration of wave propagation. Contrary to analytical models, physical channel models characterize the channel based on electromagnetic wave propagation by describing the double-directional multipath propagation between the transmit antenna array and the receive antenna array. They explicitly model wave propagation factors, such as the complex amplitude, direction of departure (DoD), direction of arrival (DoA), delay for each multipath component, polarization and time variation. Based on the complexity of the physical model, it enables accurate reproduction of radio propagation. Physical channel models can be further divides into deterministic models, geometry-based stochastic models (GSCM), and non-geometric stochastic models. Deterministic models, such as ray tracing, characterize the channel in a completely deterministic manner. While in GSCM, the impulse response is characterized by the laws of wave propagation applied to specific transmitter, receiver, and scatterer geometries, which are chosen in a random manner. On the other hand, non-geometric stochastic models determine physical parameters (DoD, DoA, delay, etc.) in a completely stochastic way by proposing underlying probability distribution functions without assuming an underlying geometry. Interested reader can refer to [22], [23], [24], [25], [26] for details about channel models. [24] highlights the tradeoff in terms of accuracy, generality, and simplicity for different modern models. They proof that the deterministic models, such as raytracing and METIS, have the advantage of accuracy compared to other ones but at the cost of simplicity. However, the price of increasing complexity is delay, which may be a limiting factor. On the other hand, stochastic channel models (SCM) including geometry-based stochastic channel models (GSCM), have the advantage of simplicity and generality while sacrificing some accuracy. The tradeoff between accuracy and simplicity for delay sensitive applications could be an interesting and wide-open topic for future research.

2.2 Current TTI length

To realize low latency system, it is mandatory to consider shorter TTI. The TTI is the smallest unit of time the transmission over radio channel is allowed in. Firstly, it affects access delay, which is the time an arriving packet has to wait until the next access opportunity occurs. In other words, if an arriving packet reaches within a TTI it has to wait until it completes before data transmission can start as shown in Figure 2 [10]. Short TTI will reduce this access delay. Furthermore, TTI defines resource blocks (RB), the atomic time/frequency resource unit assigned to a user to package its data in during transmission. Thus, the corresponding TTI has to be fully received prior to data delivery at the receiver end, which leads to further latency (See Figure 3). To gain a better understanding, LTE could be a good case for illustration. The smallest resource unit to allocate in LTE is scheduling block (SB), which consists of two RBs (an RB occupies the time of 0.5 TTI). Under favorable channel conditions, RB is capable of transmitting several kilobytes of data. However, in the case of 5G communication, both narrowband and broadband applications have to be considered. If one RB is allocated to a single machine to machine (M2M) device with data transmission of just a few bytes, then it might cause severe wastage of radio resources as well as increased latency as detailed before. Accordingly short TTI, in the order of 100 μs needs to be considered if 1 ms latency is targeted [10].

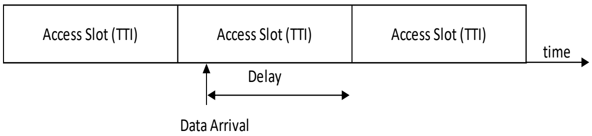


Figure 2. Access delay

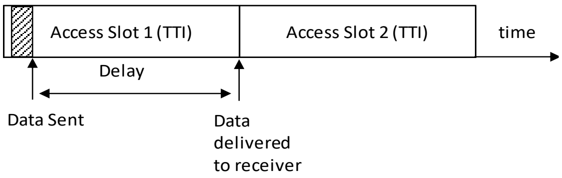


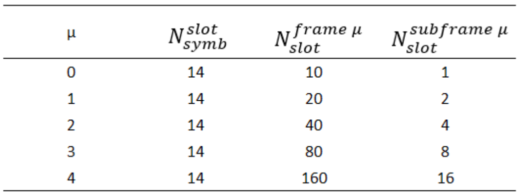
Figure 3. TTI size impact on delay

ETSI in its 5G new radio (5G NR) technical specifications: 5G NR; Physical channels and modulation (TS 138 211), triggered this by specifying different transmission numerologies [27].

Figure 2. Access delay **Figure 3. TTI size impact on delay**

. Transmission numerology μ = 4 with 240Khz subcarrier spacing and normal prefix corresponds to smallest access slot (TTI), which is desirable for low latency applications [10].

Table 1. Number of OFDM symbols per slot, slots per frame, and slots per subframe for normal cyclic prefix



μ	N_{symb}^{slot}	$N_{slot}^{frame\,\mu}$	$N_{slot}^{subframe\,\mu}$
0	14	10	1
1	14	20	2
2	14	40	4
3	14	80	8
4	14	160	16

2.3 Channel Coding

An essential factor of ensuring reliable data transmission lays within channel coding. However, channel code design impacts the receiver processing delay. Ensuring highly reliable and low latency transmission would imply the benefit of using convolutional codec rather than iterative ones such as turbo codes or low-density parity check codes [28]. The reason behinds that is because the performance of convolutional codes is round about level with iterative ones for short messages (usually the case of uMTC). Moreover, iterative codes may have an error floor which prevents them from achieving very low error rates (e.g. 10^{-9}). On the contrary, convolutional codes do not have such an error floor and the receiver enjoys more simplicity. Furthermore, in convolutional coding the receiver can start data decoding once it is being received; to the contrary, iterative codes would decode data blocks upon completely receiving it. Thus, convolutional coding enables shorter receiver processing periods and early on-the-fly channel decoding. Therefore, reference symbols (RS) should be placed at the very beginning of a TTI. This enables having channel estimation immediately after receiving the first symbols in a

TTI and then start decoding the code words, assuming that the channel will not variate significantly under very short TTI [10].

2.4 Legacy Scheduling Mechanisms

Scheduling and link adaptation are techniques for handling variations in the instantaneous radio-link quality prior to transmission of the data to satisfy quality-of-service (QoS) requirements and differentiation. Scheduling deals with the assignment of the shared resources among different users, while link adaptation controls how to set the transmission parameters of a radio link, such as modulation and coding schemes, to handle variations of the radio-link quality. Link adaptation and scheduling is closely related and they are often seen as one joint function. In the remainder of this paper, the word scheduling is used to refer to both functions.

In LTE, scheduler is part of MAC layer. It is located in the eNodeB whether for the uplink or downlink. It controls the dynamic assignment (dynamic scheduling) of uplink and downlink radio resources between users in terms of RB pairs. Typically, this implies minimizing the amount of resources needed per user and thus allowing for as many users as possible in the system, while still satisfying whatever quality-of-service requirements that may exist. Scheduling part of radio resource management (RRM) and has important interactions with other RRM functions, such as power control, link adaptation and Inter-cell Interference control (ICIC) [29]. For example, it plays a role in interference coordination by controlling the inter-cell interference on a slow basis. Based on users momentary traffic demand, quality of service (QoS) requirements and estimated channel quality, the eNodeB in each 1 ms interval takes a scheduling decision and sends scheduling information to the selected set of terminals. There is also a possibility for “semi-persistent” scheduling where a semi-static scheduling pattern is signaled in advance to reduce the control-signaling overhead. Scheduling strategy is not mandated by 3GPP specifications.

Scheduling is a major contributor to delay due to the time spent in buffering and processing. Current legacy schedulers, such as semi-persistent, fair queuing (FQ) or enhanced fair queuing (EFQ), mainly optimize on system utilization, fairness or rate while sacrificing other performance dimensions such as latency. This make them not suitable for delay sensitive services, such as uMTC. It is essential to come with new scheduling mechanisms that give highest priority for delay sensitive packets while less priority packets would also be managed and conveyed fairly. In other words, it is vital to develop scheduling mechanisms that guarantee QoS differentiation for delay applications while still giving reasonable balance among all services. Accordingly, further advanced sachems have started evolving to satisfy these challenges. This includes considering the usage of more resource dimensions beside time and frequency such as space. Also, contribution from layers other than MAC and physical layers such as a network layer has started to be considered in what so called cross-layer scheduling. However, the cost of further advanced scheduling mechanisms is a higher computational delay, which may be a limiting factor and is subject of further research. Alternatively, other emerging technologies, such as MMIMO and Network slicing, may reduce or even eliminate the need for complex schedulers, which is essential to overcome their delays. As discussed earlier, MMIMO has the vital benefit of simplified signal processing because it creates “channel hardening” such that small-scale fading is essentially eliminated due to the large beamforming gain [30]. In other words, the need for complex scheduling mechanism to assign best resource for a corresponding user could be totally eliminated, which will be positively reflected in the reduction of processing time and consequently on the reduction of over all latency. Equally important, network slicing, which is a promising technology to guarantee QoS service differentiation, could overcome the need for complex schedulers. Network slicing is an architectural approach that facilitates presenting multiple faces of the network for different use cases. Slicing addresses precise quality-of-service (QoS) requirements to each use case. Network slicing, is still in early stages of adoption on a widespread basis. It is defined in the 3GPP Release 15 specifications [31]. Further enhancements to the 5G core network slicing in Release 16, and RAN slicing is targeted for the Release 17 specifications. The described architecture gives the operator the option to offer multiple services with various nature of performance. Every network slice works as a stand-alone, virtualized network, so that a certain application deals only with one network slice that is optimized to its QoS requirements. Other slices that the subscriber is not subscribed to will be unseen and inaccessible. This architecture gives the operator the advantage of creating isolated slices that are well-tuned for specific user-behavior cases [32]. 3GPP TS.23.501 [31] has identified standardized Slice/Service Types (SSTs) for each of the 5G generic services. Accordingly, Network slicing is vital to accommodate competing QoS requirements for different 5G services. It has special importance when talking about uMTC sort of services as among other advantages, it may be an option to substitute complex schedulers for QoS differentiation, to overcome their computational latency.

3. Survey on Suggested Methods / Techniques for Latency Minimization in Literature

Several enablers and techniques has been proposed in literature to overcome latency, like shorten the packet and minimize TTI, advanced scheduling schemes, minimizing HARQ retransmissions, network slicing, and other techniques.

As mentioned earlier HARQ retransmissions cause a major part of air interface latency and it is extremely vital to reduce the number of retransmissions to limit delay. A comprehensive study on the enhancement that can be achieved by MMIMO on the number of HARQ retransmissions to reduce the delay has been done in [16]. However, this study is relatively basic and would have been more interesting if it considered more realistic environment such as using OFDM rather than QPSK narrow band wireless communication system. In addition, the study could be more comprehensive if the author included how to utilize space (coming from MMIMO) as new resource dimension for scheduling.

Some researchers, and in order to reduce HARQ latency for delay sensitive applications, suggested having a pre-scheduled resource for retransmission which is shared by a group of UEs as considered in [33]. They proved that this can minimize the time and resources needed for retransmissions, however it would have been more useful if they integrated this approach with mechanisms to reduce retransmission as a first step.

[34] suggests a data priority aware traffic aggregation model for M2M type communication. They suggest that this algorithm will improve the radio resource utilization without affecting E2E delay. In their proposal, they relied on relay nodes (RN) for aggregation (Multiplexing) of small sized data from different users over the same RB. Also, they compared advantages and drawbacks of different data scheduling techniques, namely FIFO, Priority Queuing (PQ), and Weighted Fair Queuing (WFQ), and found that PQ technique as the appropriate scheduling technique in case of delay sensitive applications. Based on that, they proposed traffic slicing model to satisfy the diversified QoS requirements for 5G services. While the idea of network slicing is interesting, one major drawback of their approach is that they did not consider reducing TTI time and instead they considered aggregation which will introduce further latency resulted from additional intermediate nodes, buffering and additional signaling. This makes it more convenient for applications with moderate latency restrictions, such as mMTC, rather than those with strict latency constrains, such as uMTC. Also, another weakness is that pure dependence on PQ may lead to starvation on lower priority classes and they did not highlight how to overcome this.

Other researches, such as [35] focused on physical layer mechanisms. They started from the air interface of Long Term Evolution (LTE) networks, and suggested several modifications to allow lower delays and higher reliability for uMTC applications. They suggested reduced TTI length, shorter OFDM symbol durations, usage of convolutional codes instead of turbo codes, high diversity levels and physical channels design that enables early channel estimation and reliable transmission. They proved that it is possible to have an OFDM based 5G radio interface that satisfies the requirements of delay sensitive and highly reliable applications, such as uMTC. The study would have been more comprehensive if the author had considered other schemes, such as cross layer scheduling and other physical layer mechanisms such as MMIMO.

Paying special attention on scheduling, in [36], opportunistic distributed multiuser scheduling in the presence of a fixed packet deadline delay constraint is addressed. They proposed a scheduling mechanism that utilizes channel gain and buffering time to make scheduling decisions with an objective to optimize power consumption while satisfying delay constraint. The paper does not directly address 5G latency challenge of 1 ms, instead it assumes packet deadline of 100ms and simply gives priority to packets with shorter time-to-live.

In [8], it is proposed to have a proactive scheduling method to minimize the delay (PDMS). The authors designed a centralized context-aware scheduler to get, as a primary object, the number of dropped packets due to missed deadline to the minimum and therefore reduce the latency, utilizing future channel information. It is assumed that the scheduler is aware of the deadline and of future channel knowledge. It was demonstrated that this scheduler is able to reduce average packet delay by more than 50%. However, this paper does not explain how to predict channel properties in the future. Also, the work would have been more comprehensive, if lower layers had been considered.

Considering channel estimation, a novel idea of progressive channel estimation through iteratively refining multiple training beams is suggested in [37]. They claim 80% capacity gain while saving 95% of training time. A more comprehensive study would include the impact of less accuracy of estimation of the proposed algorithm.

4. Key Challenges and Future Works

The challenge is with a short TTI may result to a delay spread problem and also Intersymbol Interference (ISI) at a receiver for a fast fading channel. Thus, the optimal TTI is important to ensure the data rate is well kept despite achieving the latency requirement. Another challenge is that the M-MIMO must be able to support reliable connectivity so that the ACK packet loss is not encountered during the transmission for reducing the HARQ. In addition, the hardware is another challenging factor since the mm-Wave transceivers are quite costly and consumed high power that make it is not practical to build a complete radio frequency (RF) chain for each antenna element [19] [18]. Another challenge is to come up with a time-saving encoder and decoder besides correcting the errors efficiently. It is known that the complex encoder such as Turbo codes may achieve high error correction with high number of iterations. However, the encoder can accumulate to the whole system component delay. Next challenge is to develop hybrid or hierarchical scheduling that integrated with the network slicing. The scheduling itself has been well investigated to satisfy the QoS requirements, however, the challenge is that when the QoS becomes stringent for 5G and also next generation networks. In 6G, internet of senses and teleportation has been highlighted and this required high data rate and also low latency as well as low jitter. Therefore, most scheduling works coupled with network virtualisation under spatial diversity assumption, such as MMIMO and beamforming. The network virtualisation including RAN slicing is a great challenge as the resources of the networks will be sliced according to certain QoS groups. However, within the same QoS group, a number of applications can be created and may required various specification. Thus, the challenge is that to have a number of intra-slices in the network slices. At certain extent, the limitation of number of slices will be laid due to the maximum system capacity. The major challenge will be to flexibly and adaptively share RAN resources among slice owners so that the RAN infrastructure can be efficiently utilized, while maintaining a certain degree of slice isolation.

More knowledge about CSI by applying conservative channel estimation will improve system performance, consequently minimizing HARQ retransmissions. Therefore, it is important to investigate the CSI at the perspective of reducing the latency, particularly for delay sensitive application. The tradeoff between accuracy and simplicity of different channel models for delay sensitive applications could be an interesting and wide-open topic for future research. Another technique on reducing the HARQ will be a great topic to explore as the next generation networks required minimal latency for delay sensitive application, like medical, and can be integrated with scheduling and resource allocation. Another interesting topic is to investigate the RAN slicing in reducing the latency problem despite accommodating QoS differentiation for spatial diversity channel. The ultimate aim of the overall system is to assess the improvement that can be achieved by such system on reduced latency. Another potential work is to explore on the channel coding with machine learning that can reduce the iteration for high complexity encoders and also evaluated the performance in terms of latency and jitter as well besides the bit error rate.

5. Conclusion

Reducing air latency to 1 ms is one of the major challenges in 5G networks and the requirement is likely stringent in next generation networks. In this paper major contributors to air latency have been analyzed including HARQ retransmissions, current TTI length, coding and legacy scheduling mechanisms. To address latency challenge these contributors have to be considered one by one and either minimized or eliminated. Multiple enablers have been pointed out including reduced TTI, reduced retransmissions utilizing MMIMO and better channel estimation, and better Quality of Service (QoS) differentiation that satisfies latency requirements. As a result we believe that contrary to earlier generations there is no perfect approach suitable for all 5G services or beyond, and that the network needs to be sliced to have better QoS differentiation that matches each of the different groups. Also, we believe that combination of technology components needs to be integrated in order to address latency problem. Specifically speaking, there is need first to reduce TTI size in order of 100 µs and to combine that with suitable coding, new context-aware cross-layer scheduling, network slicing, and new physical layer enablers, such as MMIMO, that help in reducing or eliminating retransmissions.

Acknowledgements

This work is partially funded by Universiti Teknologi Malaysia (Grant No. Q.K130000.2456.08G12)

References

[1] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, *et al.*, "The METIS 5G System Concept: Meeting the 5G Requirements," *IEEE Communications Magazine,* vol. 54, pp. 132-139, 2016.

[2] "Scenarios, requirements and KPIs for 5G mobile and wireless system," *Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS),* vol. Deliverable D1.1, 29-04-2013 2013.

[3] "The outcome of the evaluation, consensus building and decision of the IMT-Advanced process (Steps 4 to 7), including characteristics of IMT-Advanced radio interfaces," ITU-R Report M.2198November 2011 2011.

[4] Samsung, "5G Vision," *DMC R&D Center, Samsung Electronics Co., Ltd.,* Feb-2015.

[5] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys & Tutorials,* vol. 20, pp. 3098-3130, 2018.

[6] R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving Ultra-Low Latency in 5G Millimeter Wave Cellular Networks," *IEEE Communications Magazine,* vol. 55, pp. 196-203, 2017.

[7] X. Jiang, H. Shokri-Ghadikolaei, G. Fodor, E. Modiano, Z. Pang, M. Zorzi*, et al.*, "Low-Latency Networking: Where Latency Lurks and How to Tame It," *Proceedings of the IEEE,* vol. 107, pp. 280-306, 2019.

[8] R. Holakouei and P. Marsch, "Proactive Delay-Minimizing Scheduling for 5G Ultra Dense Deployments," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall),* 2015, pp. 1-5.

[9] C. B. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Communications Magazine,* vol. 50, pp. 46-53, 2012.

[10] A. O. J. F. M. P. Marsch, *5G Mobile and Wireless Communications Technology* vol. 1st. New York, NY, USA: Cambridge University Press 2016.

[11] S. P. Erik Dahlman, Johan Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. UK: ELSEVIER, 2011.

[12] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)," Cedex - FRANCE2009.

[13] 3GPP, "LTE; Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)," Cedex - FRANCE2009.

[14] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine,* vol. 52, pp. 186-195, 2014.

[15] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, *et al.*, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Processing Magazine,* vol. 30, pp. 40-60, 2013.

[16] N. Zarifeh, A. Kabbani, M. El-Absi, T. Kreul, and T. Kaiser, "Massive MIMO exploitation to reduce HARQ delay in wireless communication system," in *2016 IEEE Middle East Conference on Antennas and Propagation (MECAP)*, 2016, pp. 1-5.

[17] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, pp. 742-758, 2014.

[18] L. A. Adeeb Salh, Nor Shahida M Shah, and Shipun A Hamzah, "Adaptive Antenna Selection and Power Allocation in Downlink Massive MIMO Systems," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, p. 3521~3528, 2017 2017.

[19] Z. Li, S. Han, and A. F. Molisch, "Hybrid beamforming design for millimeter-wave multi-user massive MIMO downlink," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1-6.

[20] Z. Xinying, A. F. Molisch, and K. Sun-Yuan, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Transactions on Signal Processing*, vol. 53, pp. 4091-4103, 2005.

[21] B. M. S. Yasmine M. Tabra, "Hybrid MVDR-LMS beamforming for massive MIMO," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, pp. 715-723, November 2019 2019.

[22] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, *et al.*, "Survey of channel and radio propagation models for wireless MIMO systems," *EURASIP J. Wirel. Commun. Netw.*, vol. 2007, pp. 56-56, 2007.

[23] B. C. C. Oestges, *MIMO Wireless Networks Channels, Techniques and Standards for Multi-Antenna, Multi-User and Multi-Cell Systems*, 2 ed.: Elsevier, 2013.

[24] P. Ferrand, M. Amara, S. Valentin, and M. Guillaud, "Trends and challenges in wireless channel modeling for evolving radio access," *IEEE Communications Magazine*, vol. 54, pp. 93-99, 2016.

[25] J. M. Pekka Kyösti, Lassi Hentilä, Xiongwen Zhao, Tommi, Jämsä, Christian Schneider, Milan Narandzić, Marko Milojević, Aihua Hong, Juha Ylitalo, Veli-Matti Holappa, Mikko Alatossava, Robert Bultitude, Yvo de Jong, Terhi Rautiainen, "WINNER II Channel Models (IST-4-027756 WINNER II D1.1.2 V1.2)," vol. IST-4-027756 WINNER II D1.1.2 V1.2, V1.2 ed, 2008.

[26] J. K. Yong Soo Cho, Won Young Yang, Chung G. Kang, *MIMO-OFDM Wireless Communications with MATLAB*: Wiley-IEEE Press, 2010.

[27] ETSI, "5G NR; Physical channels and modulation," vol. 3GPP TS 38.211 version 15.2.0 Release 15, ed: ETSI, July 2018.

[28] N. A. Johansson, Y. P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1184-1189.

[29] Ericsson, *LTE L14 Air Interface*, 2014.

[30] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, *et al.*, "Hybrid Beamforming for Massive MIMO: A Survey," *IEEE Communications Magazine*, vol. 55, pp. 134-141, 2017.

[31] 3GPP, "Technical Specification Group Services and System Aspects; System Architecture for the 5G System (5GS); (Release 15)," in *3GPP TS 23.501 V15.7.0* vol. 3GPP TS 23.501 V15.7.0, ed. Valbonne - FRANCE: 3GPP, 2019.

[32] Rysavy Research, "Global 5G: Implications of a Transformational Technology," September 2019.

[33] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-Scheduled Resources for Retransmissions in Ultra-Reliable and Low Latency Communications," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1-5.

[34] M. Dighriri, A. S. D. Alfoudi, G. M. Lee, T. Baker, and R. Pereira, "Comparison Data Traffic Scheduling Techniques for Classifying QoS over 5G Mobile Networks," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2017, pp. 492-497.

[35] O. N. C. Yilmaz, Y. P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1190-1195.

[36] M. M. Butt, K. Kansanen, and R. R. Muller, "Individual Packet Deadline Constrained Opportunistic Scheduling for a Multiuser System," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, 2011, pp. 1-5.

[37] H. Cheng, C. Liao, and A. A. Wu, "Progressive channel estimation for ultra-low latency millimeter-wave communications," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 610-614.

Journal of Engineering and Science Research 7 (1): 48-57, 2023
e-ISSN: 2289-7127
© RMP Publications, 2023
DOI: 10.26666/rmp.jesr.2023.1.8

Corresponding Author: Rudzidatul Akmam Dziyauddin, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia (UTM), Malaysia. Email: rudzidatul.kl@utm.my