5G Ultra-Reliable and Low-Latency Systems Design

Chih-Ping Li, Jing Jiang, Wanshi Chen, Tingfang Ji, and John Smee
Qualcomm Research, Qualcomm Technologies, Inc., San Diego, CA, 92121
Email: {cpli,jingj,wanshic,tji,jsmee}@qti.qualcomm.com

***Abstract*—5G New Radio (NR) is envisioned to support three broad categories of services: evolved mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC). The URLLC services refer to future applications that require secure data communications from one end to another with ultra-high reliability and deadline-based low latency requirements. This type of quality-of-service is vastly different from that of traditional mobile broadband applications. In this paper, we discuss the systems design principles to enable the URLLC services in 5G. Theoretical queueing analysis and system-level simulations are provided to support these systems design choices, many of which have been considered as work items in the 3GPP Release 15 standards, which will be the first release for 5G NR.**

I. Introduction

Ultra-reliable and low-latency communications (URLLC) is a new service category that will be supported in 5G New Radio (NR). It is targeted at emerging applications in which data messages are time-sensitive and must be securely delivered end-to-end subject to high reliability and hard latency requirements. The hard latency requirement means that a data transmission that cannot be decoded at the receiver before a deadline is of no use and can be dropped from the system, resulting in the loss of reliability. As an example, the current 3GPP requirement for URLLC includes the hard latency of one millisecond over the air interface and the system reliability of 99.999%. That is, the quality of service (QoS) of URLLC is violated if more than one out of 10^5 packets fails to be delivered in one millisecond over the wireless medium. This is in contrast to the QoS of mobile broadband applications that optimize data throughput and average delay. Use cases of URLLC include autonomous vehicles that perform cooperation and safety functions, monitoring and control in smart grids, tactile feedback in remote medical procedures, control and coordination of unmanned aviation vehicles, robotics, and industrial automation.

The QoS of URLLC introduces several new challenges in the wireless systems design. Mobile broadband applications can rely on a sufficient number of hybrid automatic repeat request (HARQ) retransmissions to achieve ultra-high reliability, but this approach is limited for URLLC due to the hard latency requirement. Low latency can potentially be achieved by allocating enough resources to minimize the block error rate (BLER) of HARQ transmissions, but it may result in low spectral efficiency and system capacity. All URLLC users connected to the radio access network are expected to satisfy the QoS and receive equal grade of service (EGOS), for which the outage capacity is of interest but not the ergodic capacity. The outage capacity is affected by the tail behavior of the wireless system, e.g., the tail of random traffic demand and intra/inter-cell interference, and users that are at the cell edge, power-limited, or in deep fade, that needs to be optimized. When the services of URLLC and evolved mobile broadband (eMBB) in 5G are deployed in the same radio spectrum, their coexistence needs to have a future-proof design to best accommodate the mixed demand that is unknown at the moment.

In this paper, we provide the high-level systems design for URLLC that is supported by both theoretical queueing analysis and system-level simulation results. The focus is on the downlink of URLLC transmissions in a frequency-division-duplexing (FDD) system; the same design principles are applicable to the uplink case. Our contributions include:

- An efficient HARQ design with shortened transmission time interval (TTI) and round trip time (RTT) is shown to improve the outage capacity of URLLC and the minimally achievable low latency.
- System-level simulations and theoretical queueing analysis are provided to demonstrate the fundamental tradeoff between the outage capacity, system bandwidth, and the latency requirement for URLLC. These tradeoffs lead to the optimized systems design.
- We show that multiplexing the URLLC and eMBB traffic by frequency division multiplexing (FDM) is highly inefficient, and wideband resource allocation is needed for URLLC to maximize the outage capacity. Dynamic multiplexing between URLLC and eMBB in both time and frequency domains is also desired to optimize the system spectral efficiency and the capacity of both services.

The remainder of the paper is organized as follows. Section II shows the need of examining the impact of queueing effect on the system performance, that wideband resource allocation maximizes the URLLC capacity, and that dynamic multiplexing of eMBB and URLLC is desired to maximize the overall system spectral efficiency. Section III discusses how the downlink systems design are applicable to the uplink scenario, followed by the conclusion in Section IV.

II. The Systems Design for URLLC

A. Queueing effect

In the downlink of a radio access network (RAN), the end-to-end delay of a successful data transmission comprises of scheduling delay (the time between packet arrival and the next scheduling instant), queueing delay, transmission delay, receiver-side processing and decoding delay, and multiple

978–1–5386–3873–6/17/$31.00

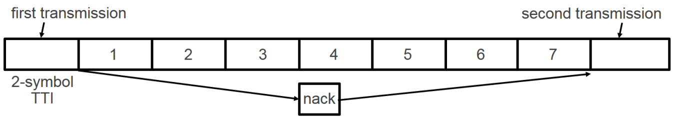

Fig. 1: A HARQ RTT timeline that has 2-symbol TTI and 16-symbol RTT.

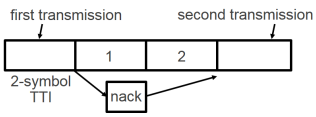

Fig. 2: A HARQ RTT timeline that has 2-symbol TTI and 6-symbol RTT.

HARQ RTTs. The queueing delay results from the statistical multiplexing of data flows destined for multiple URLLC users. The data flows may also be sporadic and bursty because the traffic models of various URLLC use cases may not be standardized to have a periodic arrival pattern as in the voice-over-IP (VoIP) services in 4G Long Term Evolution (LTE). The queueing effect needs to be considered in the systems design of URLLC because of the hard latency requirement. In general, sufficient HARQ retransmissions are needed to achieved high reliability while leaving enough delay margin from the hard latency bound to mitigate the queueing effect, which is worsened as the admitted traffic load increases so as to maximize the spectral efficiency for the URLLC services.

B. Shortened TTI and HARQ RTT timeline

One simple method to meet the URLLC QoS is to reduce the durations of TTI and HARQ RTT so that more HARQ retransmissions are allowed to achieve high reliability. Reducing the TTI duration includes using fewer orthogonal-frequency-divison-multiplexing (OFDM) symbols in one TTI and shortening the OFDM symbols by increasing the subcarrier spacing. Short OFDM symbols also enable more efficient pipeline processing that lead to a more tightened HARQ RTT timeline. Figures 1 and 2 show the HARQ RTT timelines with 8 and 3 HARQ processes, respectively. Figure 3 shows the system capacity of URLLC under these two HARQ timelines and different hard latency requirements, where all URLLC users are at the worst-case -3dB cell edge and have the Poisson traffic. We observe that the system capacity and the minimally achievable latency can be substantially improved with the short HARQ RTT because (i) less time is needed to have enough HARQ retransmissions to meet the reliability target; and (ii) URLLC packets have wider delay margin to tolerate more queueing delay before their deadlines.

To obtain qualitative insights on the relationship between the URLLC capacity and the hard latency requirement in Figure 3, we consider an $M/M/m/k$ queueing model: The first M means Poisson packet arrivals. The second M means exponential service times that reflect the time to decode a data packet after multiple HARQ retransmissions follows approximately a

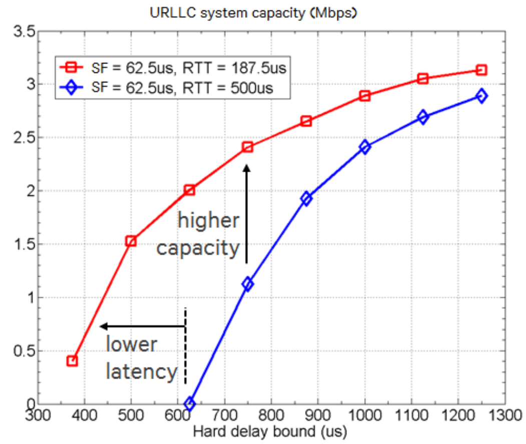

Fig. 3: The URLLC system capacity under different hard latency requirements and HARQ RTT timelines. Simulation assumptions include system bandwidth of 20MHz, that all URLLC users are at the worst-case of -3dB geometry and have the Poisson traffic, and the reliability requirement of 99.99%. The term "SF" stands for subframes and is equivalent to TTI here.

geometric distribution.[1] The notation m indicates the number of allowed concurrent transmissions, which is proportional to the system bandwidth available to the URLLC services. The notation k means that if an arriving packet observes k outstanding packets in the system for some $k > m$, both being queued and undergoing HARQ retransmissions, it will be dropped from the network. The value of k increases with the hard latency requirement, say d milliseconds, in the sense that if an arriving packet sees d-millisecond worth of packets awaiting in the system, it will surely miss its deadline and shall be discarded from the network. In this queueing model, the probability p_{block} of dropping packets, i.e., the loss of system reliability, is [2]

$$p_{\text{block}} = \left(G p_0 \frac{m^m}{m!} \right) \rho^k = \Theta(\rho^k), \quad \rho = \lambda/(m\mu),$$

where G is a constant, p_0 is the probability that the system is empty, $(1/\mu)$ is the mean service time, and λ is the Poisson arrival rate. We have $\lambda = \Theta\left(\sqrt[k]{p_{\text{block}}}\right)$. The URLLC capacity λ_{URLLC} is the admitted arrival rate in the $M/M/m/k$ queue:

$$\lambda_{\text{URLLC}} = (1 - p_{\text{block}})\lambda \approx \lambda = \Theta\left(\sqrt[k]{p_{\text{block}}}\right) = \Theta\left(p_{\text{block}}^{(1/\text{latency})}\right),$$

which is an increasing function of the latency requirement with diminishing return because $p_{\text{block}} < 1$, and reflects the delay-capacity scaling curve in Figure 3.

The drawback of short TTI duration includes more control overhead which reduces the data capacity for URLLC but could be alleviated by grant-free transmissions (e.g., semi-persistent scheduling). More resource blocks (RBs) need to be allocated in the frequency domain to a packet transmission

[1] This is only an approximation in a sense that a packet is continuously serviced in the queueing model, but the resources are made available in between HARQ retransmissions of the packet in the wireless network.

to achieve a target BLER, in which the loss of trunking efficiency adversely affects the outage capacity of URLLC. Similarly, reducing the duration of OFDM symbols increases the subcarrier spacing, and fewer RBs are available in the frequency domain, causing more queueing effect. Overall, the TTI and RTT durations should be chosen carefully to optimize these performance tradeoffs. The TTI durations of one and two OFDM symbols for URLLC have been agreed in 3GPP.

C. Wideband allocation for URLLC

The traffic demand of the URLLC services may be sporadic in some use cases and cannot fully utilize the system resources. Efficiency multiplexing between URLLC and eMBB becomes important to maximize the overall system spectral efficiency. One approach is to partition the system bandwidth statically or semi-statically and serve the URLLC and eMBB traffic by frequency-division multiplexing (FDM). Next, we present a queueing analysis and system-level simulations to show that reserving bandwidth for URLLC yields low spectral efficiency and resource utilization. Instead, wideband allocation for URLLC is desired.

1) Queueing analysis: Assume some bandwidth is reserved for URLLC. The URLLC traffic model is Poisson. The QoS requirements include the delivery of fixed-size data packets over the air interface within the hard latency bound of L milliseconds and the system reliability of $(1-p_{\text{loss}})$, e.g., $L=1$ and $p_{\text{loss}} = 10^{-5}$ [1]. The amount of reserved bandwidth decides how many URLLC transmissions can be packed in the frequency domain in one TTI. From these assumptions, we consider an $M/D/m/m$ queueing model: "M" means the Poisson arrival process, "D" means the over-the-air delay is a fixed value (e.g., assuming all URLLC transmissions can be decoded successfully in one transmission and there is no need for HARQ retransmissions nor ACK/NACK feedback), the first "m" denotes the maximum number of concurrent transmissions in one TTI, and the second "m" indicates that packets that have positive queueing delay are dropped from the system due to missing a stringent hard deadline. In this queueing model, the loss of system reliability is the Erlang-B formula [2, pp.179]:

$$p_{\text{loss}} = \frac{(\lambda/\mu)^m/m!}{\sum_{n=0}^{m}(\lambda/\mu)^n/n!}, \tag{1}$$

where λ is the Poisson arrival rate in the unit of packets per TTI and $(1/\mu)$ is the fixed service time in the unit of TTIs; let $\mu = 1$ here. The corresponding system utilization, defined as the time-average proportion of allocated resources, is

$$\sum_{k=1}^{m}\left(\frac{k}{m}\right)\frac{(\lambda/\mu)^k/k!}{\sum_{n=0}^{m}(\lambda/\mu)^n/n!}. \tag{2}$$

Figure 4 shows the relationship among the loss of system reliability, packet arrival rate, and resource utilization according to (1) and (2). We observe that increasing the packet arrival rate yields more severe queueing effect and loss of system reliability. As the reliability requirement is tightened, we must reduce the traffic load with lower resource utilization in order to continue meeting the QoS requirement. Figure 5 shows the

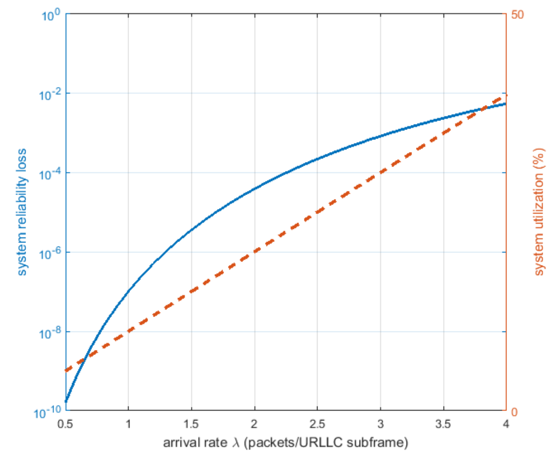

Fig. 4: The relationship among the loss of system reliability, resource allocation, and the packet arrival rate in an $M/D/m/m$ queueing model with $m = 10$. The hard latency requirement is implied in the queueing model, in which queueing delay is disallowed and leads to packet drop.

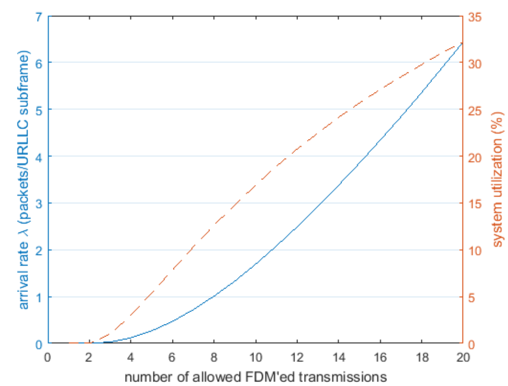

Fig. 5: Maximally supportable Poisson arrival rate and the resource utilization under the reliability of $p_{\text{loss}} = 10^{-5}$ in an $M/D/m/m$ queueing model. The first "m" is the number of allowed simultaneous data transmissions and scales with the reserved bandwidth for URLLC.

maximally supported arrival rate (i.e., system capacity) and the resource utilization under the requirement of $p_{\text{loss}} = 10^{-5}$. As the number of allowed concurrent transmissions (i.e., the available bandwidth for URLLC) decreases, the system capacity decreases exponentially with low resource utilization. As an example, when the number of concurrent transmissions is reduced from 16 to 8 to 4, the maximally supported arrival rate changes from 4.2 to 1.0 to 0.1 packets per TTI, respectively. If we consider the spectral efficiency over the available bandwidth to be the ratio of the achievable packet arrival rate to the number of concurrent transmissions, then the spectral efficiency is improved from $0.1/4 = 0.025$ to $1.0/8 = 0.125$ to $4.2/16 = 0.2625$ as the number of concurrent transmissions is doubled. It indicates that the spectral efficiency of the URLLC traffic increases with the available bandwidth.

	hard latency requirement		
system bandwidth	500μs	750μs	1ms
5MHz	0Mbps	0Mbps	0Mbps
10MHz	0.28Mbps (0.03bps/Hz)	3.94Mbps (0.39bps/Hz)	6.2Mbps (0.62bps/Hz)
20MHz	10.7Mbps (0.54bps/Hz)	15.8Mbps (0.79bps/Hz)	16.9Mbps (0.85bps/Hz)

TABLE I: The URLLC system capacity and spectral efficiency under different available system bandwidth and the hard latency requirements. The reliability requirement is 99.999%.

2) System-level simulations: We perform system-level simulations on the FDD downlink in the rural scenario. There is one URLLC serving cell with 20 eMBB neighboring cells in a wrapped-around model. Overhead of the control channels PDCCH and PUCCH is not modeled. The gNB (i.e., base-station) in the serving cell periodically sends reference signals to the randomly distributed URLLC users for channel estimation, and the users report the channel station information to the gNB for the multiple-input-multiple-output (MIMO) operation. The scheduling policy is delay-based and focuses on providing EGOS to the users to optimize the outage capacity. A packet is dropped from the system when its deadline is expired. The serving cell is subject to full-buffer inter-cell interference from all neighboring cells. All users are assumed to have the same Poisson traffic load. The system capacity for URLLC is defined as the largest arrival rate at which the hard latency and reliability requirements are satisfied for all users, including the cell-edge ones. As an example, if the maximum arrival rate at which all URLLC users meet the QoS is λ packets/second/user, where there are 22 users in the serving cell and 256-bit packet payload is used, then the system capacity is $\lambda \times 256\,\text{bits/packet} \times 22\,\text{users/cell} = 5.63\lambda$ kbps/cell. A range of arrival rates is swept to find the system capacity. See Appendix A for the details of simulation assumptions.

Table I shows the URLLC system capacity under different available bandwidth and the hard latency requirements with the fixed reliability requirement of 99.999%. The system capacity scales super-linearly and the spectral efficiency improves with the available bandwidth as predicted in the queueing analysis in Section II-C1. The URLLC capacity under the reserved bandwidth of 5MHz is negligible because the lack of RBs in the frequency domain requires more HARQ retransmissions and causes severe queueing effect so that the hard latency and reliability requirements cannot be met simultaneously. Making the wide system bandwidth available is beneficial to the practical deployment of the URLLC services. Table II shows the resource utilization of the reserved bandwidth when the URLLC users are fully loaded at the capacity-achieving points shown in Table I. The resource utilization is generally low, especially over the narrow bandwidth, and cannot be increased by admitting more traffic because the QoS requirement will be violated. Thus, it is inefficient to reserve narrow bandwidth for the URLLC traffic.

	hard latency requirement		
system bandwidth	500μs	750μs	1ms
5MHz	0%	0%	0%
10MHz	2.3%	32.4%	50.9%
20MHz	43.3%	64%	68.6%

TABLE II: Resource utilization of the available bandwidth at the capacity-achieving points in Table I.

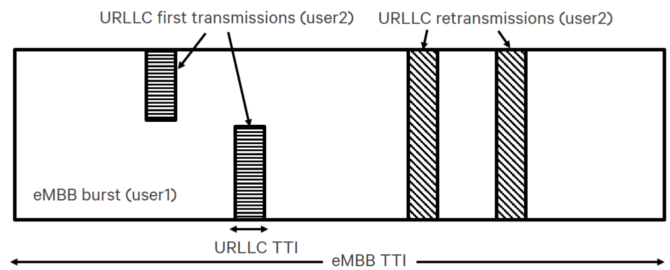

Fig. 6: Dynamic multiplexing of the eMBB and URLLC traffic in both time and frequency domains.

D. Dynamic multiplexing of eMBB and URLLC

Following the above discussions, the eMBB and URLLC services, when sharing the same radio spectrum, need to be dynamically multiplexed over the entire system bandwidth to maximize the spectral efficiency of both services. As an example in Table II, the URLLC services have the maximum resource utilization of 50.9% under the system bandwidth of 10MHz and the latency requirement of one millisecond. The 49.1% of the system resources are wasted under the static FDM between eMBB and URLLC. If dynamic multiplexing is applied, the eMBB capacity could jump from zero to 20.52Mbps following the calculation of 4.18bps/Hz $\times$ 10MHz $\times 0.491 = 20.52$Mbps, where 4.18bps/Hz is the eMBB full-buffer spectral efficiency on the downlink [3].

In the time domain, URLLC requires short TTI duration, e.g., two OFDM symbols or 71.43μs under the subcarrier spacing of 30KHz, to be the smallest scheduling time unit to reduce both the HARQ RTT and scheduling delay as discussed in Section II-B. For the eMBB traffic that typically involves bulk data transfer, long TTI duration, e.g., 500μs, is preferred to minimize control overhead and exploit the coding gain of long codewords to improve the data capacity and spectral efficiency. Due to the different scheduling granularity, eMBB and URLLC services need to be dynamically multiplexed in the time domain as well. See Figure 6 for an example. When URLLC packets arrive at the gNB during an ongoing eMBB transmission, they cannot wait until the end of the eMBB burst to be scheduled because of the hard latency requirement. In this case, the gNB may suspend the eMBB transmission to free up resources to transmit the URLLC data in the next URLLC TTI, which is referred to as the puncturing mechanism. In such cases, an indication mechanism is useful for the gNB to inform the punctured eMBB user about the location of the preempted resources to improve its decoding performance.

In summary, dynamic multiplexing of eMBB and URLLC

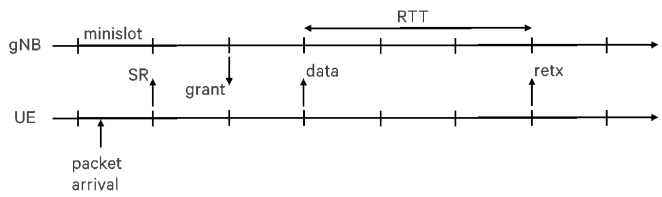

Fig. 7: The hand-shaking procedure for the user to request resources from the gNB for the upcoming uplink transmissions.

system bandwidth (MHz)	system capacity (Mbps)	spectral efficiency (bps/Hz)	Resource utilization
5	4.51	0.9	51.5%
10	12.39	1.24	51.5%
20	28.16	1.41	74.4%

TABLE III: System capacity, spectral efficiency, and resource utilization in the case of uplink URLLC transmissions.

services in both time and frequency domains is desired. This is a flexible scheme that shall be able to handle the mixed yet unknown demand of both services.

III. The Systems Design of URLLC on the Uplink

The systems design for the URLLC services on the uplink is more difficult than the downlink case for several reasons. The user equipment (UE) may be power-limited on the uplink so that it is harder to satisfy the high reliability requirement especially for the users that are temporarily at the cell edge or in deep fade. When URLLC packets arrive at the user, a hand-shaking procedure is initiated by the user to request resources from the gNB for the uplink transmissions; see Figure 7 for an example. This procedure incurs additional delay overhead and reduces the delay budget for the actual transmission and combating the queueing effect. In the dynamic multiplexing of eMBB and URLLC services, preempting ongoing eMBB uplink bursts to make room for new URLLC arriving packets requires the gNB to explicitly signal the eMBB user who is responsible for making the resources available, which needs to be completed in a very short period of time.

Despite the challenges outlined above, the systems design principles discussed in this paper are applicable to the uplink case. Table III presents the system-level simulation results for the uplink URLLC transmissions under idealized assumptions (see [4] for more details). Similarly, the system capacity grows super-linearly and the spectral efficiency improves with the available bandwidth. Dynamic multiplexing of eMBB and URLLC services is desired on the uplink.

IV. Conclusion

We present the systems design principles to enable the URLLC services in both downlink and uplink scenarios in 5G NR. The queueing effect is shown to have a significant impact on the performance of URLLC because of the hard latency requirement. From the first principles, we show that wideband resources allocation is desired to maximize the spectral efficiency for URLLC, and dynamically multiplexing eMBB and URLLC traffic in both time and frequency helps to maximize the overall system spectral efficiency. Indication mechanisms are useful to facilitate the coexistence of the eMBB and URLLC traffic that has different scheduling granularity.

Acknowledgment

The authors would like to thank the systems and software teams working on the 5G project at Qualcomm for the fruitful discussions and the support of simulation environments.

References

[1] 3GPP Technical Report 38.913, "*Study on Scenarios and Requirements for Next Generation Access Technologies*."

[2] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 1992.

[3] R1-166390, "*Updated Sub6 DL Full-Buffer KPI Evaluation for eMBB*," Qualcomm Incorporated, 3GPP RAN1 meeting #86, Gothenburg, Sweden, 2016.

[4] R1-1705624, "*UL URLLC capacity study and URLLC eMBB dynamic multiplexing design principle*," Qualcomm Incorporated, 3GPP RAN1 meeting #88bis, Spokane, WA, 2017.

Appendix A
System-level Simulation Assumptions

Scenario	Rural
Layout	Single macro layer. Hexagonal grid with 21 cells wrapped around
Inter-BS distance	1732m
Carrier frequency	2GHz
System bandwidth	$\{5, 10, 20\}$MHz
Channel model	3D UMa
Transmission power	BS: 49dBm PA scaled with bandwidth. UE: 23dBm
Antenna config	2Tx/2Rx (X-pol)
BS antenna height	35m
BS antenna element gain and connector loss	8dBi
BS/UE receiver noise figure	5/9dB
Traffic model	eMBB: full-buffer. URLLC: Poisson with 32-byte packets.
UE distribution	URLLC: 22 UEs in the serving cell with 50% indoor and 50% outdoor. eMBB: one UE in each neighboring cell.
Tone spacing	60KHz
Cyclic prefix duration	NCP
TTI/RTT duration	2/6 OFDM symbols
System reliability	99.999%
Hard latency	$\{500, 750, 1000\}\mu s$
MIMO	2×2 SU-MIMO