2025 36th International Teletraffic Congress (ITC 36)

A Discrete-Time Model of the 5G New Radio Uplink Channel

Simon Raffeck*, Sebastian G. Grøsvik§, Laura Alexandra Becker¶, Stanislav Lange§, Stefan Geissler*, Thomas Zinner§, Wolfgang Kellerer¶, Tobias Hossfeld*
*Chair of Communication Networks, University of Würzburg, Germany
§Norwegian University of Science and Technology
¶Technical University Munich, Germany
Email: {firstname.lastname}@{uni-wuerzburg.de, ntnu.no, tum.de}

***Abstract*—The introduction of 5G comes with significant advancements over LTE in both core and radio domains, with a shift to microservices in the core and enhanced flexibility in the radio access network. While these changes boost performance, it comes at the cost of increased complexity in the radio access and core configuration, particularly in the radio domain. This complexity is heightened by use case-specific 5G deployments in campus and industrial networks, making the selection of optimal configuration parameters a challenging task. To address this, we present a discrete-time queueing model for the 5G New Radio uplink channel, capable of predicting key performance indicators (KPIs) such as the one-way delay as well as analyzing the channel's impact on traffic streams. The model is validated through simulations and comparisons with 5G campus network measurements.**

***Index Terms*—5G New Radio (NR), discrete-time analysis, uplink channel performance, model validation, 5G campus measurements.**

I. Introduction

As commercial cellular networks transition from LTE to 5G, operators continue to upgrade their infrastructure and enhance the quality of service (QoS) to end users and their devices. To this end, the 5G network technology introduces novel features to increase network flexibility in order to support a wide range of existing and new use cases. By providing high-performance, ubiquitous connectivity across a broad range of verticals from public mobile network operators, over industrial and health deployments, to private campus networks for on-premise connectivity, 5G and its future developments are set to form the base for heterogeneous deployments with vastly different performance expectations and requirements. In addition, as an increasingly heterogeneous user equipment landscape is emerging, sectors in the industry look to 5G LAN and private 5G deployments to fulfill their networking needs. These technologies are expected to provide a holistic private network that can be configured to provide optimal service to sector-specific use cases.

However, with the expectation of custom-tailored mobile service for a heterogeneous set of use cases comes the inevitable challenge of finding relevant configuration parameters as well as optimal configuration values. Due to the newly introduced flexibility in both the core domain and the radio channel, identifying use case-specific and optimal system configurations is a complex challenge. Especially as the scope in which 5G performance can be custom-configured to facilitate service requirements remains unclear and largely unexplored in literature. To this end, we address the two critical open challenges of (1) identifying configurable and relevant parameters of the 5G New Radio (NR) uplink channel and (2) evaluating the impact of the identified parameters on key performance metrics of the 5G NR uplink channel in terms of packets' sojourn times from the user equipment (UE) to the next-generation NodeB (gNB) egress and the traffic pattern in terms of batch size of reassembled packets at the gNB egress.

In previous work [1], we have investigated the 5G performance across different deployments of 5G campus networks; despite differences in the options, some key parameters and fundamental behaviors emerge across the board and are, hence, worth investigating in a more generic manner. These parameters lend themselves to generalization and modeling.

In this paper, we explore the latency behavior of the wireless 5G NR uplink channel. To this end, we propose a discrete-time queuing model of the data transmission from UE to gNB and the subsequent packet reassembly process at the gNB. Our model provides a means of understanding the wireless link, deriving the impact of various configuration parameters on expected packet latency through the predicted sojourn time, simplifying the intricate 5G NR mechanisms and optimizing the system to support service requirements. As a result, we establish a relationship between over-the-air packet latency, gNB configuration, and the latency impact of specific changes to the gNB performance parameters. We validate our model by comparing results to both abstract and detailed simulations as well as real-world measurements in our 5G campus network, quantifying gaps and limitations of the assumptions and abstractions made in our model.

The remainder of this work is structured as follows. Section II reviews prior work on queuing models and 5G performance modeling. Section III covers the necessary technical background, while Section IV introduces our discrete-time model. Section V describes our validation methodology, followed by an accuracy analysis in Section VI. Finally, Section VII concludes and suggests future research directions.

978-3-948377-03-8/19/$31.00 ©2025 ITC

II. Related Work

In this section, we discuss related work in the areas of modeling various components within the 5G architecture, specifically radio resources, as well as the development of queuing models in general. The relevant research for this publication falls into two broad categories. Firstly, we provide a brief overview of the methodological landscape of queuing models and discrete-time analysis in general, before discussing similar works dealing with the performance evaluation of the 5G radio channel.

Using queueing networks to model complex networked systems has been a staple methodology for several years in the research community. Bharath-Kumar in [2] investigated the performance of different networks using queuing theory. These results have later been expanded in [3]. The queueing network analyzer, proposed by Whitt in [4], [5] and later expanded upon in [6], is able to approximate various metrics within complex queueing networks, such as congestion prediction and mean steady-state performance for each of the queues. In their book [7] Shortle *et al.* discuss several important fundamentals for the analysis of queuing networks, from parametric decomposition, over superimposed processes, to the computation of departure processes. In more recent publications, the authors of [8] present a modeling approach for multi-component queuing networks that allows for the concatenation of queueing components, as well as splitting and superposition of random processes. When it comes to clock-regulated queueing models, several works discuss both clocked arrivals [9], [10] and service units [11]–[13]. Thereby, service events – or arrivals – do not occur based on stochastic processes, but according to a clock-synchronized mechanism. Note that this differs from a deterministic service process, as the service times are not necessarily deterministic. Instead, the time between departures is strictly regulated by multiples of the clock cycle.

When it comes to the performance evaluation of 5G mobile networks, several works propose different modeling approaches for components within the 5G ecosystem. In [14], [15], the authors propose a model for massive Machine-Type Communications (mMTC) traffic in 5G networks. To this end, they analyze the expected Interarrival Times (IATs) for the traffic arriving at base-stations as well as at the User Plane Function (UPF) within the core network. For traffic and UE generation, the authors relied heavily on the UERANSIM emulation tool. However, the study lacks the usage of real hardware, as well as the inclusion of the radio channel, to investigate its impact on key performance metrics. Especially, as previous studies have shown that the radio channel has a significant and measurable impact on the traffic characteristics observed at the UPF [16]. In [17] the authors propose a performance model to optimize system configurations based on the different states a UE can assume and how these affect the performance of the 5G network. However, this work does not take into account the behavior of gNBs, as well as the impact and behavior of different traffic streams. The authors of [18] develop a GI/GI/1 queueing model for the 5G fronthaul, modeling the performance of the uplink after data has been processed by the gNB. Opposed to the authors' assumptions, our model shows that the traffic stream leaving the gNB is not a general independent stochastic process, but exhibits strong burstiness due to the batching behavior of the gNB. Finally, [19], [20] present simulators of the 5G transmission channel. The simulators cover key NR features, including flexible frame structures, support for multiple numerologies, and bandwidth part management for dynamic resource allocation. The latter of the two is used in this work to validate the predictions made by our proposed model. In previous works, we have already begun to investigate the impact of different gNB configurations and implementations on the overall performance of 5G campus networks [1]. Furthermore, in [16], we have analyzed the discretization behavior of different gNB implementations, and highlighted that the traffic leaves the gNB in batches determined by the system configuration parameters. None of these works, however, present a generalizable model of the Radio Access Network (RAN) components and their influence on the traffic characteristics at the gNB egress.

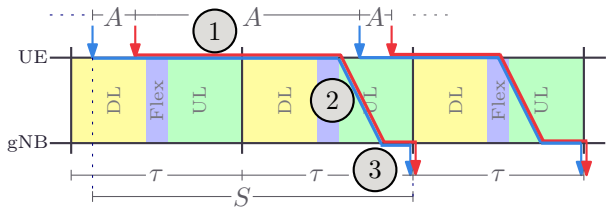


Fig. 1. Schematic representation of the 5G radio channel.

To address this gap in the literature, we propose a multi-stage clocked discrete-time model that enables a thorough analysis of the impact of gNB implementations and 5G NR configurations on the packets' sojourn time distribution and the batch size of reassembled packets at the gNB egress. Thus, critical key performance metrics, such as the expected delay in the radio uplink, can be captured.

III. Technical System Description

We base our model on a standards-compliant [21] 5G NR transmission channel as shown in Figure 1. Thereby, the transmission process begins at the UE, where packets are prepared for delivery by the protocol stack, encapsulated with headers, and queued in the buffer for transmission. Here, we assume an arrival process of packets determined by their interarrival time A. These packets await their allocated time slot on the radio channel (1), determined by the scheduler at the gNB, which orchestrates uplink and downlink transmissions based on channel conditions and resource availability. During their designated time slot within a frame, the UE transmits the packets over the air interface. For Time Division Duplex (TDD) configurations, which we focus on in this work, the frame structure alternates between downlink and uplink periods separated by a flex period that acts as a guard band and can be used as uplink or downlink. This pattern repeats after

each TDD period τ. Packets are transmitted during the uplink period, but can also use parts of the flex period (2). Upon reception at the gNB, the transmitted signals are demodulated, decoded, and error-checked to reconstruct the original packets. These packets are then reassembled in sequence at the end of each TDD pattern, completing the end-to-end delivery process (3). The total time between packet arrival at the UE and reassembly at the gNB is referred to as the sojourn time S. In the following, we provide additional technical details on the baseline technical system. To increase the flexibility of the 5G radio channel, the standard includes new, dynamically adjustable parameters to extend the features already introduced in LTE [22]. Next, we outline the 5G NR frame structure as well as available resources in time and frequency domains.

The overall structure of the 5G radio channel is determined by the duplex mechanism. Frequency-Division Duplex (FDD) makes use of dedicated frequency bands to facilitate simultaneous uplink and downlink communication, and is generally used for lower bands. Time-Division Duplex (TDD), on the other hand, is the standard for bands of $3.5\,\mathrm{GHz}$ and higher, and uses discrete-time slots to schedule uplink and downlink traffic within the same frequency band. When using TDD, the radio channel is organized in frames along the time axis, and subcarriers along the frequency axis.

Figure 2 depicts two exemplary frame structures using different TDD configurations. In any configuration, the largest time unit is an NR frame with a duration of $10\,\mathrm{ms}$. Each frame consists of 10 subframes with a duration of $1\,\mathrm{ms}$ that each contains a set number of transmission slots defined by the used numerology [21]. In the examples provided here as well as for all evaluations conducted in this work, we use a numerology of $\mu = 1$, resulting in a subcarrier spacing of $30\,\mathrm{kHz}$ whereas each subframe consists of two transmission slots. Each slot, therefore, has a duration of $0.5\,\mathrm{ms}$ and can carry up to 14 OFDM symbols, or 12 for extended cyclic prefix [21].

Within a given numerology, it is possible to configure different TDD patterns and periods for fine granular channel control. Thereby, the TDD pattern describes the number of slots that are being allocated to either downlink or uplink, and how many symbols must be reserved for either direction within the flex slot. The TDD periodicity specifies the time interval after which the chosen pattern repeats. Note that the periodicity is constrained by the maximum duration of the NR frame as an upper limit, while a chosen period must fit exactly 1 to *n* times into a radio frame without exceeding the $10\,\mathrm{ms}$ duration. Operators can leverage these parameters to precisely define how many uplink or downlink slots should be contained within one NR frame, and thus tailor their configurations to the chosen use case.

A configuration with a 20 slot TDD pattern is shown in Figure 2a. In this scenario, the pattern has a periodicity of 10 ms and thus fills an entire frame. Within each frame, 9 slots are reserved for downlink (yellow) and 10 for uplink (green), respectively. The remaining 20th flex slot (blue) is split evenly between downlink and uplink, separated by a guard time (red). Figure 2b shows an alternative configuration with a TDD periodicity of 5 ms, meaning the configured pattern repeats two times in each radio frame. In addition, the slot allocation has changed, and more slots are allocated to downlink than uplink, while the flex slot contains more uplink symbols than downlink symbols. This fine-grained control of the radio resources enables high flexibility in the time domain to adapt to individual use cases.

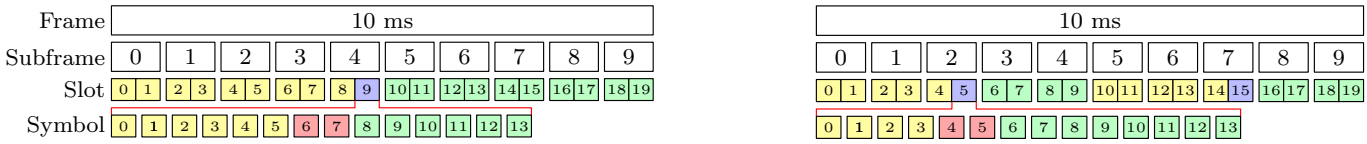
In addition to multiplexing in the time domain using TDD, 5G allocates resources along the frequency domain as well. Thereby, multiple orthogonal, and hence non-interfering, OFDM subcarriers are spread across the used band and form the smallest resource unit in the frequency domain. A Resource Block (RB) aggregates 12 subcarriers into a schedulable unit in the frequency domain. These available resources in the radio channel are assigned to individual UEs by allocating a number of slots in the time domain and a number of RBs in the frequency domain. The resulting time-frequency matrix is called the resource grid and contains all resources available for 5G radio transmissions. This means that, if more than one device is present, UEs compete for the resources in both time and frequency domains. Similarly, if no competing devices are present, a single UE could be assigned all available RBs and slots. Furthermore, the way in which these resources are assigned to the UEs depends on the implementation of the gNB. At the time of writing, the most common open-source gNB solutions make use of either round-robin or proportional fair share, to distribute their resources [23]. Note that our model uses the number of allocated RBs as an input, and hence assumes scheduling has already happened.

Finally, once resources have been assigned and payload data has been transmitted across the radio channel from a connected UE to the gNB, packet data is reassembled into IP packets, encapsulated into the GPRS Tunneling Protocol (GTP) tunnel, and forwarded towards the UPF. This reassembly step is happening at discrete points in time, for the most prominent open-source implementations, this occurs at the end of each TDD pattern [1], [16], i.e., every 10 ms and 5 ms in the examples shown in Figure 2.

IV. System Model

This section presents the system model and abstractions for data transmission in the 5G RAN. Based on these, we develop a discrete-time queueing model to evaluate key metrics like packet sojourn time from generation at the UE to gNB egress.

The main components of the system are displayed in Figure 3. The system is characterized by a packet arrival process from the UE's application layer whose interarrival times follow an arbitrary distribution A, as well as two unbounded queues that represent buffers that store individual symbols at the UE and gNB, respectively. Transmissions between UE and gNB, as well as emissions of departing reassembled packets at the gNB, follow a clocked regime whose frequency is determined by the duration τ of the TDD pattern. In particular, the UE can transmit up to a maximum number β of symbols per TDD pattern, while the gNB is assumed to reassemble all complete packets at the end of a pattern.



(a) 20 slot TDD pattern repeats every full frame (10 ms), (b) 10 slot TDD pattern repeats every half frame (5 ms).

Fig. 2. Exemplary TDD frame structure for numerology $\mu = 1$. Colors indicate usage: downlink (yellow), uplink (green), flex (blue), guard time (red).

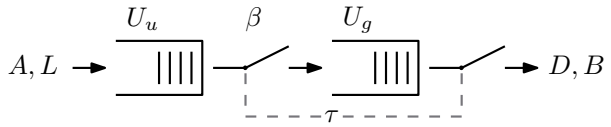


Fig. 3. Model overview.

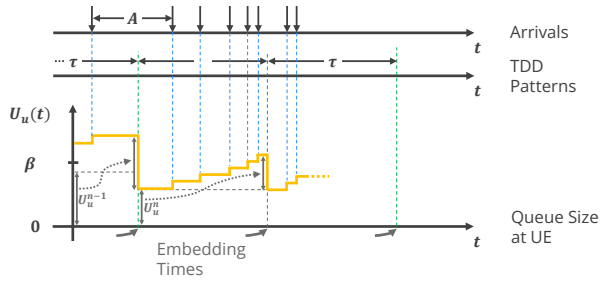


Fig. 4. Exemplary state development of the model.

TABLE I
RANDOM VARIABLES (RV) USED IN THE DISCRETE-TIME MODEL.

Variable	Description
Input parameters of the discrete-time model	
τ	TDD pattern duration [ms] (const.).
β	Maximum number of symbols that can be transmitted from UE to gNB during one pattern duration [symbols] (const.).
A	Packet interarrival time [ms] (RV).
L	Packet size [symbols] (RV).
$X_{t,a}$	Number of arrivals (RV) whose interarrival time is distributed according to a during an interval whose length is distributed according to t [24]. If the interval duration is a constant, we implicitly apply the deterministic distribution with probability mass 1 at t.
Random variable at n-th embedding point	
U_u^n	Unfinished work (queue size) at the UE after n-th embedding time [symbols].
Random variables in steady state	
U_u	Unfinished work at the UE after embedding times [symbols].
T	Amount of transmitted symbols from UE to gNB during a TDD pattern [symbols].
U_g	Unfinished work (queue size) at the gNB after embedding times [symbols].
B	Batch size of reassembled packets at the gNB egress [packets].
O	Relative position (order) of an arrival within a TDD pattern.
U_u^A	Unfinished work (queue size) at the UE at packet arrival times [symbols].
S	Sojourn time of packets from arrival at the UE to reassembly at the gNB egress [ms].

In the following, we first outline conventions regarding notation and illustrate additional details of the analytical model. Based on this, we provide the computational steps for solving an embedded Markov chain using fixed-point iteration to extract steady-state distributions of system characteristics such as queue sizes, as well as derived metrics such as the sojourn time S. Table I provides an overview of the random variables and distributions that are used throughout this paper. The top part lists model inputs, whereas the bottom part consists of auxiliary variables and outputs. We denote random variables (RVs) with uppercase letters.

In the proposed model, we characterize the system state by the queue sizes observed at the UE and gNB immediately after the end of a TDD pattern, and reason about the evolution of these quantities between consecutive embedding times. We use RVs with superscript indices such as U_u^n to refer to RVs at specific n-th embedding times, and omit the indices when talking about the steady state. To account for the fact that packets are broken down into symbols and hence might get transferred over the course of multiple consecutive TDD patterns, we consider queue sizes in terms of symbols rather than packets. We present an exemplary sequence of events and their impact on the queue size at the UE, as observed by arrivals U_u^A and at embedding times U_u^n, in Figure 4. On the one hand, packet arrivals increase the queue fill level at the UE by the size of the corresponding packets. On the other hand, the UE can transmit up to β symbols to the gNB every TDD pattern duration τ. The value of β depends on configuration parameters such as the bandwidth, the number of uplink slots, the number of uplink symbols in the special slot, and the modulation scheme. While this transmission actually occurs continuously, we focus on the system state at embedding times and drain the UE at the end of each pattern. Since we consider the uplink direction, the UE can not immediately start transmitting symbols of packets that just arrived, but needs to inform the gNB about its intent. Thus, the number of transmitted symbols from UE to gNB during a TDD pattern is determined by β and the number of symbols in the UE queue at the *previous* embedding instance U_u^{n-1}, as indicated by the dashed arrows in the figure. The transmitted symbols arrive at the gNB, which follows a clocked approach to reassemble

batches of complete packets every τ, leaving symbols that do not amount to a full packet in the queue.

Based on this process, we perform a fixed-point iteration by initializing the system state with $U_u^0 = 0$ and $U_g^0 = 0$, i.e., empty queues at UE and gNB. Then, we compute the distributions at consecutive embedding times based on preceding ones until we reach steady state, i.e., $U_u = \lim_{n\to\infty} U_u^n$. Using the steady state distributions of different quantities at embedding times, we can finally derive additional system characteristics, such as the queue sizes seen by individual packet arrivals and the resulting packet sojourn times.

A. Non-stationary Analysis: Fixed-point Iteration

In the following, we provide details on the iterative calculation of the steady state distributions of U_u, U_g, B, and T. We justify the memorylessness of U_u and U_g, due to arrival of symbols in the system and the batching process being indepent and identically distributed. First, we can determine the queue size at the UE immediately after the $(n+1)$-st embedding time based on the outlined system description, i.e., subtracting the up to β symbols that remained in the queue after the previous transmission U_u^n and adding the number of symbols introduced by $X_{\tau,a}$ packet arrivals over the course of a pattern duration according to Equation 1.

$$U_u^{n+1} = \max(U_u^n - \beta, 0) + X_{\tau,a} \cdot L \tag{1}$$

Next, we obtain the number of transmitted symbols from UE to gNB, T^n, and the number of reassembled packets at the gNB, B^n, during the n-th TDD pattern by means of Equations 2 and 3.

$$T^n = \min(U_u^{n-1}, \beta) \tag{2}$$

$$B^n = \left\lfloor \frac{U_g^n + T^n}{L} \right\rfloor \tag{3}$$

With these auxiliary RVs, we calculate the queue size at the gNB after the $(n+1)$-st embedding instance as

$$U_g^{n+1} = \begin{cases} 0 & \text{if } T^n < \beta, \\ U_g^n + T^n - B^n \cdot L & \text{otherwise.} \end{cases} \tag{4}$$

Since all symbols of a packet arrive simultaneously at the UE, cases with fewer than β transmitted symbols between the UE and the gNB indicate that the last transmitted symbol marks the end of a packet, and therefore, the gNB queue is drained completely. Otherwise, a transmission might end mid-packet, leaving symbols in the gNB queue.

We apply the power method and perform the fixed-point iteration until convergence is reached, i.e., until the distributions of the discussed RVs between consecutive iterations only differ within a predefined numerical accuracy threshold. Unless otherwise stated, the convergence criteria used in this work is a total absolute difference of less than 10^{-5}. Note that this convergence is only reached for parameter combinations that result in $\rho < 1$. Thereby, we compute $\rho = \frac{L/A}{\beta/\tau}$. Based on Lindley's Integral Equation and FIFO queueing, $\rho < 1$ is sufficient to ensure convergence. For $\rho >= 1$, the model simply does not converge, as the queue size at the UE diverges against $+\infty$.

B. Derived Metrics

With the steady-state distributions of the queue and batch sizes, we can proceed to compute the sojourn time S of packets, covering the time from their generation at the UE to their reassembly at the gNB. To this end, we first take the perspective of an arbitrary packet arrival, reason about the quantities that are required to compute its sojourn time, and provide steps for obtaining them.

When a packet is generated at the UE, there may already be a number of symbols in the UE queue, meaning that there may be one or more size-β UE-to-gNB transmissions ahead of it. We also observe that while U_u gives us the amount of unfinished work immediately after embedding times, i.e., after TDD pattern borders, we need to know the amount of work seen by packet arrivals, U_u^A.

To obtain U_u^A, we need to consider that depending on the relative position of an arrival within a pattern, it will encounter a different amount of unfinished work at the UE. Specifically, the first arrival in a pattern will encounter U_u unfinished work, the second will encounter $U_u + L$, and the $i-$th arrival will encounter $U_u + (i-1) \cdot L$. The distribution of the relative position across all arrivals can be derived from $X_{\tau,a}$ and is represented by the RV O.

$$U_u^A = U_u + (O-1) \cdot L \tag{5}$$

Given U_u^A, we can determine N, the number of UE-to-gNB transmissions that will happen before the symbols of an arriving packet can be transmitted. Since the last symbol of a packet needs to reach the gNB before reassembly, we add $L-1$ to U_u^A and use β to compute the number of UE-to-gNB transmissions that will happen before it is this arrival's turn, cf. Equation 6. For the sojourn time calculation, this quantity is later multiplied by τ.

$$N = \left\lfloor \frac{U_u^A + L - 1}{\beta} \right\rfloor \tag{6}$$

Additionally, the packet's arrival time relative to the TDD pattern borders affects its time until eventual reassembly. We define Δ as the time between a packet arrival and the next TDD pattern border and compute it using similar considerations as with U_u^A: depending on the number of arrivals in a pattern and an arrival's relative position in that sequence of arrivals, the time until the next pattern border equals $\tau - R_a$ for the first arrival and $\tau - R_a - (i-1) \cdot A$ for the $i-$th arrival. In this context, R_a denotes the forward recurrence time of A. As per Equation 7, the probability for each outcome can be computed using the auxiliary RV O.

$$\Delta = \tau - R_a - (O-1) \cdot A \tag{7}$$

With the quantities derived so far, we can determine the total sojourn time S. Note that both N and Δ are correlated with the relative position of an arrival within its TDD pattern O, i.e., early arrivals encounter less unfinished work in the UE queue,

TABLE II
EVALUATED PARAMETER CONFIGURATIONS ACROSS BOTH OAI AND SRS.

IAT Distribution	Packet Size [B]	TDD Period	DL:UL Ratio	DL Slots	DL Symbols	UL Slots	UL Symbols	ρ
Geom, Poisson[a], Deterministic[a]	100, 700, 1400	20 slots	2:1	13	5	6	7	0.6, 0.7, 0.8, 0.9, 0.95, 0.98
Geom, Poisson[a], Deterministic[a]	100, 700, 1400	10 slots	2:1	6	8	3	4	0.6, 0.7, 0.8, 0.9, 0.95, 0.98
Geom, Poisson[a], Deterministic[a]	100, 700, 1400	5 slots	2:1	3	5	1	7	0.6, 0.7, 0.8, 0.9, 0.95, 0.98

[a]Investigated with hardware measurements

but tend to have to wait longer until the end of the next pattern and vice versa. To account for this and respect the dependency, we condition both random variables on O when calculating the sojourn time, yielding the following expression.

$$S|o = \max(N|o, 1) \cdot \tau + \Delta|o \tag{8}$$

Note that the distribution of the overall sojourn time S can be derived from the conditional $S|o$ by iterating over all possible realizations of O. Furthermore, the first part of the expression, $\max(N|o, 1) \cdot \tau$, covers the specific case of the uplink direction where arrivals that find a small enough queue at the UE still have to wait for a full TDD pattern duration before being transmitted due to scheduling. Finally, a packet's relative position in the batch of reassembled packets adds an offset to its sojourn time. Since gNBs are typically connected with wired links with significantly higher capacities than that of the radio channel, we consider this offset to be negligible.

V. METHODOLOGY

In the following section, we describe the methodology for verifying and validating the model. Verification is done via a discrete-event simulation that replicates the model. Validation compares model results with a detailed OMNeT++ simulation [20], [25] and measurements from a licensed 5G campus deployment with physical UEs.

A. Simulation Frameworks

Before evaluating the validity of our proposed model, we verify that the model is, in fact, working as described in the previous section. To this end, we develop and implement a discrete-event simulation that reproduces the behavior described by the model. We call this the *basic simulation* from here on out. To accurately reflect the behavior of the model, the accompanying simulation includes the same abstractions and assumptions. Specifically, the simulation works off of the same input parameters and abstracts the radio channel identically. As described in the previous section, we abstract the radio channel by computing the number of symbols that can be transmitted from the UE to the gNB in the span of a single TDD pattern. On the one hand, this allows for a fine granular configuration of various TDD patterns, on the other hand, it allows for easy extension in the future towards non-optimal or even fluctuating channel conditions. Currently, we assume the number of transmitted symbols T to be constant, however, neither the model nor the validating simulation is limited to a constant value and can accept a distribution instead.

In order to evaluate the impact of the above abstractions, we further compare the model results to a more detailed simulation on the basis of the *OMNeT++*, *5GTQ*, and *Simu5G* framework, as presented in [25]. In a previous work [20], this simulation was extended to also include the TDD patterns and the ability to accurately model signaling in the radio channel. We use this implementation to evaluate the impact of our assumptions on the model output and show that the effects observed in our model and resulting simulation are not an artifact of our approach. We call this second simulation the detailed simulation from here on out.

B. Testbed Description

To evaluate the validity of our proposed model, we implemented a 5G testbed using off-the-shelf components and open-source software. Since we are investigating the QoS performance of the 5G network, especially in regard to delay and IATs, having synchronized clocks is crucial. To this end, we ensure comparable timestamps by connecting a Quectel RM520N-GL modem via a USB M.2 carrier board to the same host running the gNB implementation, using an Ettus USRP B210 as radio frontend. This ensures that the two measurement points, shown in Figure 5, are based on the same clock and are hence comparable. As gNB solutions, we make use of two of the most prominent open-source implementations, *Openairinterface* [26] (v2.2.0; 03946cd47b) and *srsRAN* [27] (24.04.0; 51e44a642). Both of these allow us the configuration flexibility that was described above, and thus enable us to investigate the behavior described in our model.

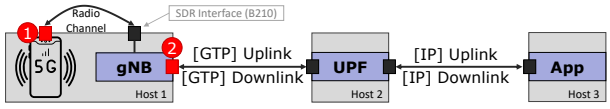


Fig. 5. Hardware testbed for campus 5G measurements. Measurement points (1) Radio Interface of UE and (2) Backhaul Interface of gNB, both use the clock of Host 1.

C. Measurement Scenarios

A full description of the investigated scenarios can be seen in Table II. Throughout our whole investigation, we keep the numerology set to $\mu = 1$ and use the frequency band n79 as well as a modulation scheme of QAM256 at a bandwidth of 20 MHz. The 12 data symbols for the flex slot is taken from the 3GPP standard [28]. Note that in practice, additional overhead may exist due to additional DRMS and PDCCH

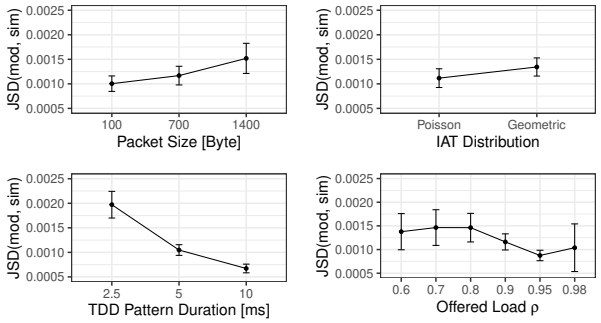


Fig. 6. Main effects on JSD between the sojourn time distributions obtained via our model and the basic simulation.

symbols. We differentiate between different arrival processes, that we use to define the packet stream from the UE towards the gNB. For this study, we generated the IATs using three different distributions: deterministic, geometric distribution, and Poisson distribution. We make use of the parameters describing the distributions, to adjust the expected load ρ from 0.6 up to 0.98. Furthermore, we use varying packet sizes, going from $100\,\mathrm{Bytes}$ over $700\,\mathrm{Bytes}$, to a maximum of $1400\,\mathrm{Bytes}$. Lastly, we leverage the configuration parameters described earlier and investigate different TDD pattern settings. For the TDD period, we use three different values: 5 slots, 10 slots, and 20 slots. For these configurations, we adjust the number of downlink and uplink slots and symbols, to, as closely as possible, reach a downlink to uplink symbols ratio of 2:1.

VI. Evaluation

In the following. we first compare our model's data with simulations and measurements. Then, a case study examines how configuration parameters affect sojourn time (uplink delay) and the batch size of packets exiting the gNB, highlighting 5G-induced burstiness.

A. Verification and Validation

Figure 6 shows the main effect plot for the Jensen-Shannon Divergence (JSD) between the resulting sojourn time distributions of our proposed model and the basic simulation. First, we investigate the impact of the packet size, which we varied between $100\,\mathrm{B}$, $700\,\mathrm{B}$, and $1400\,\mathrm{B}$. Increasing the packet size leads to a bigger approximation error for our model, even though only the confidence intervals for $100\,\mathrm{B}$ and $1400\,\mathrm{B}$ do not overlap. The observed error, due to numerical inaccuracies, gets lower the more packets are included in an embedding interval, and hence in each TDD pattern, as effects tend to average out with a larger number of events. Next, we look into the IAT distribution and the offered load and their impact on the JSD. For both parameters, all the confidence intervals overlap, indicating that they have no significant impact on the approximation error. Lastly, the impact of the TDD pattern duration is investigated, which also represents the number of slots within the pattern. We can see a clear improvement in the approximation error with an increase in the pattern duration. For increasing pattern durations, there is a shift in the number of symbols per pattern interval we can transmit ranging from 21 over 46 to 91 symbols, for $2.5\,\mathrm{ms}$, $5\,\mathrm{ms}$ and, $10\,\mathrm{ms}$, respectively. Similarly to the effect of the packet size, the higher number of events within each embedding interval improves the prediction accuracy of the model. Overall, the JSD values are small across the board, highlighting the close fit between our model and the basic simulation.

Next, in Figure 7, we investigate the calculated sojourn time of our model compared to the state-of-the-art OMNeT++ simulation framework that was described above. To this end, we look into the CDF for three exemplary configurations, as noted in the plot facets. The blue line represents the sojourn time of our model, in red is the one obtained from the detailed simulation, and lastly, green is the sojourn time of our model slightly adjusted to fit the constraints of the OMNeT++ simulation. Specifically, our model assumes all packets arriving in the previous TDD pattern are scheduled for transmission in the current interval by neglecting the condition that the frame is only scheduled in case it was already buffered before the last transmission of a buffer status report. This buffering process usually takes about 1 slot (0.5 ms), leading to packets arriving in the last slot of a pattern not being scheduled for the subsequent transmission period. Furthermore, in the detailed simulation, a packet is only forwarded after the acknowledgement is sent, resulting in an additional slot of delay (0.5 ms). These two abstractions lead to the total offset of 1 ms observed in the figure that can be accounted for by simply shifting the model prediction by 1 ms towards the right. The shifted model matches the experimental data from the detailed simulation closely.

Finally, we compare our model output to measurements obtained in the physical testbed described before. To this end, Figure 8 shows the ECDF of the batch interdeparture time at the gNB egress. This coincides with the time between packet assembly steps, as mentioned before. Our model assumption is that this happens after every TDD pattern, meaning it should happen every $2.5\,\mathrm{ms}$, $5\,\mathrm{ms}$ and, $10\,\mathrm{ms}$ for patterns with 5, 10 and 20 slots, respectively. The measurement data shows that the majority of batch interdepartures are spaced exactly as expected. However, the figure also shows that there are additional occurrences at multiples of the expected interdeparture time. This behavior is more prominent for SRS than for OAI, but the trend is clearly noticeable. This is a clear difference to the model that predicts a deterministic interdeparture time that is in line with the TDD pattern duration. The reason for this discrepancy is the neglect of radio channel conditions in our model, where we assume no transmission errors to occur. In the testbed, however, channel conditions are not perfect and retransmissions on L2 can occur.

This behavior is visualized as a time series in Figure 9, depicting the one-way-delay and the gNB interdeparture time for our empirical measurements SRS and OAI as well as the detailed OMNeT++ simulation for which we used the included indoor hotspot model to reflect the expected channel degradation [29]. For all three deployments, identical configuration parameters have been chosen, with a packet size of $700\,\mathrm{B}$, a

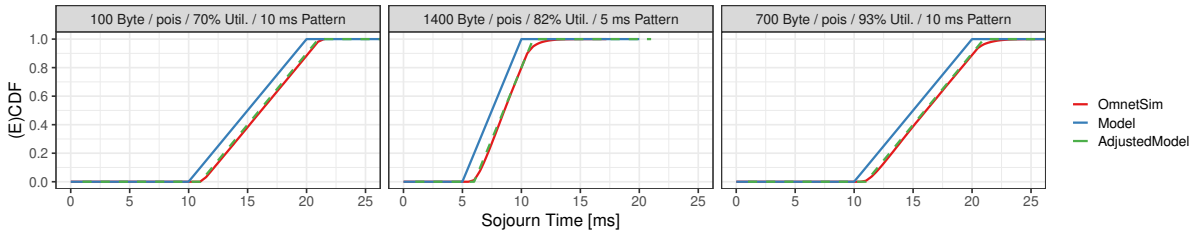


Fig. 7. Sojourn time distributions in different scenarios obtained via our model and the OMNeT++ simulator. The adjusted model distributions are obtained by shifting the original model output by 1 ms to account for differences in the way resources in time domain are allocated.

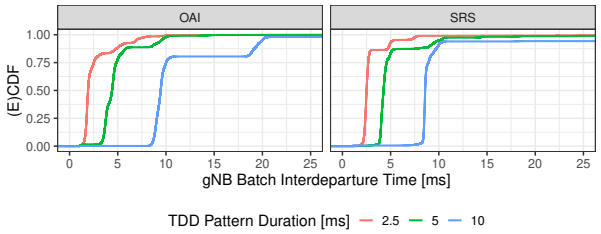


Fig. 8. ECDF of the batch interdeparture time for different TDD pattern durations for both OAI and SRS.

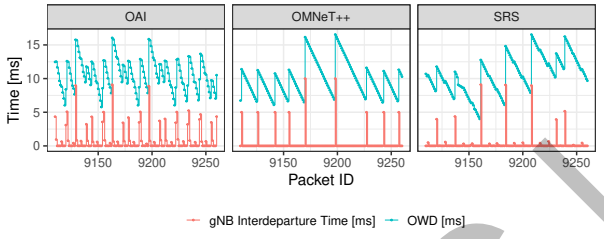


Fig. 9. Time series showcasing the one-way-delay and its relationship to the gNB interdeparture time. One-way-delay in blue, with gNB interdeparture time in red.

utilization of 0.7, a pattern duration of 5 ms, and deterministic interarrival times. The one-way-delay is shown in blue, with the gNB batch interdeparture time in red. For the one-way-delay, all three facets exhibit the typical saw-tooth pattern that has already been discussed in previous works [1], with sporadic jumps in the delay, whenever a retransmission occurs due to deteriorating channel quality. The gNB inter departure time, exhibits burstiness, with most of the packets leaving in batches, every 5 ms, and therefore at the end of the TDD pattern. Some batch interdeparture times exhibit multiples of the pattern duration, these correlate with the jumps in the one-way-delay, which can be explained by the gNB waiting on the re-transmission of the missing frame before re-assembling the packet, thus adding one additional pattern duration to the batch interdeparture time. For the empirical measurement studies, these values slightly differ, due to testbed and implementation limitations, but the general trend remains.

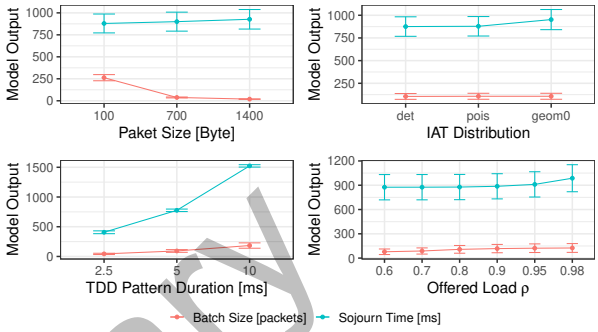


Fig. 10. Main effects on mean sojourn time $E[S]$ (blue) and mean batch size at gNB egress $E[B]$ (red) based on model predictions.

B. Case Study

Finally, to showcase the application of the proposed model, a main effects plot to visualize the impact of different configuration parameters on the performance of the 5G uplink channel is presented in Figure 10. The mean sojourn time $E[S]$ is depicted in blue, while the mean batch size at the gNB egress $E[B]$ is drawn in red. Focusing on the batch size, the data shows that, due to the overlapping confidence intervals, neither the offered load nor the IAT distribution has a significant impact on the mean batch size. For the TDD pattern duration and packet size, there is a clear trend visible. As larger packets decrease the batch size, since intuitively fewer packets fit into a single transmission window, an increasing pattern duration has the opposite effect, as longer TDD patterns provide more resources to transmit packets before the next reassembly happens, leading to larger batches. For the sojourn time, the data shows only the TDD pattern duration to have a significant impact, with longer patterns leading to higher sojourn times. As longer TDD patterns directly lead to a larger distance between packet reassembly events at the gNB, packets, on average, have to wait longer before being reassembled and transmitted toward the UPF. In the real world, there is likely an efficiency tradeoff between reassembling more often with short patterns at the cost of increased computational effort. However, more detailed evaluations are required to confirm or deny this assumption. Overall, our model is capable of predicting several KPIs of the 5G New Radio uplink channel and can be used for both dimensioning and the identification of optimal configuration parameters.

VII. Conclusion

With many degrees of freedom for configuring 5G New Radio components to meet the requirements of a wide range of heterogeneous use cases, identifying the main factors and quantifying their effect on desired key performance indicators poses a significant challenge. In this work, we present a discrete-time model of the 5G New Radio channel, specifically focusing on the delay behavior of the uplink channel from packet generation at the UE to packet reassembly at the gNB.

The model allows assessing the impact of configuration parameters such as TDD period and pattern configuration as well as traffic-related characteristics like packet size, interarrival time distribution, and packet rate on one-way delays and the burstiness of the gNB departure process. We validate the model using simulations that operate at different degrees of detail as well as measurements from a licensed 5G campus deployment using physical UEs, showing a high degree of agreement both qualitatively and quantitatively.

By applying the proposed model to a wide range of configuration parameters, we identified the key influencing factors on the sojourn time, the one-way-delay in the uplink direction, and the traffic characteristics of the traffic stream at the gNB egress. The data shows that shorter TDD patterns generally produce lower delay values, as packet reassembly is happening more frequently. The model generally predicts mean sojourn times that are roughly 1.5 times the pattern duration with 3.76 ms, 7.51 ms, and 15 ms for patterns of length 2.5 ms, 5 ms, and 10 ms, respectively. At the same time, the shorter intervals generate a less bursty traffic stream at the gNB egress. Future studies should quantify the tradeoff between computational overhead and reduced delay. Our model currently omits radio channel deterioration, which may cause retransmissions, but its flexibility allows for incorporating this effect by modeling transmitted symbols as a random variable whose distribution follows that of existing channel loss models, such as the one we applied in the detailed simulation. Additionally, we plan to expand the model for multiple UEs, and therefore our testbeds and simulations. These extensions and further investigations remain for future work.

Acknowledgments

This work was partially funded by the Research Council of Norway through the SFI Norwegian Centre for Cybersecurity in Critical Sectors (NORCICS) project no. 310105 as well as ORIGAMI project from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101139270.

References

[1] S. Raffeck, S. G. Grøsvik, S. Lange, T. Hossfeld, T. Zinner, and S. Geissler, "Parameterizing 5G New Radio: A Comparative Measurement Study on Throughput and Delay," in *International Conference on Network and Service Management (CNSM)*. IEEE, 2024.

[2] K. Bharath-Kumar, "Discrete-Time Queueing Systems and Their Networks," *IEEE Transactions on Communications*, 1980.

[3] H. Bruneel, "Comments on "Discrete-Time Queueing Systems and Their Networks"," *IEEE Transactions on Communications*, 1983.

[4] W. Whitt, "The Queueing Network Analyzer," *The Bell System Technical Journal*, 1983.

[5] ——, "Performance of the Queueing Network Analyzer," *The Bell System Technical Journal*, 1983.

[6] W. Whitt and W. You, "A robust queueing network analyzer based on indices of dispersion," *Naval Research Logistics (NRL)*, 2022.

[7] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of queueing theory*. John Wiley & Sons, 2018.

[8] S. Geißler, S. Lange, G. Hasslinger, P. Tran-Gia, and T. Hoßfeld, "Discrete-Time Analysis of Multi-Component Queuing Networks under Renewal Approximation," in *34th International Teletraffic Congress (ITC 34), IEEE*, 2022.

[9] O. E. Percus and J. K. Percus, "Some results concerning clock-regulated queues," *ACM SIGARCH Computer Architecture News*, 1988.

[10] C. P. Kruskal, M. Snir, and A. Weiss, "The distribution of waiting times in clocked multistage interconnection networks," *IEEE Transactions on Computers*, 1988.

[11] T. Katayama, "Waiting time analysis for a queueing system with time-limited service and exponential timer," *Naval Research Logistics (NRL)*, 2001.

[12] O. J. Boxma and W. P. Groenendijk, "Waiting times in discrete-time cyclic-service systems," *IEEE Transactions on Communications*, 1988.

[13] O. E. Percus and J. K. Percus, "Time series transformations in clocked queueing networks," *Communications on pure and applied mathematics*, 1991.

[14] E. Goshi, F. Mehmeti, T. F. La Porta, and W. Kellerer, "Modeling and Analysis of mMTC Traffic in 5G Core Networks," *Transactions on Network and Service Management*, 2024.

[15] F. Mehmeti and T. F. La Porta, "Modeling and Analysis of mMTC Traffic in 5G Base Stations," in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, 2022.

[16] S. Raffeck, S. Geissler, and T. Hossfeld, "Towards Modeling the Impact of 5G New Radio on Dataplane Traffic Characteristics," in *KuVS Fachgespräch - Würzburg Workshop on Modeling, Analysis and Simulation of Next-Generation Communication Networks (WueWoWAS)*, 2024.

[17] W. Zhan, C. Xu, X. Sun, and J. Zou, "Toward Optimal Connection Management for Massive Machine-Type Communications in 5G System," *IEEE Internet of Things Journal*, 2021.

[18] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *Journal of optical communications and networking*, 2018.

[19] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, 2019.

[20] L. Becker and W. Kellerer, "P5G-TSN: A Private 5G TSN Simulation Framework," 2024.

[21] 3rd Generation Partnership Project (3GPP), "3GPP TS 38.211 V18.4.0 (2024-09): NR; Physical channels and modulation (Release 18)," Tech. Rep., 2024.

[22] D. Hui, S. Sandberg, Y. Blankenship, M. Andersson, and L. Grosjean, "Channel coding in 5G new radio: A tutorial overview and performance comparison with 4G LTE," *IEEE Vehicular Technology Magazine*, 2018.

[23] R.-M. Ursu, A. Papa, and W. Kellerer, "Experimental Evaluation of Downlink Scheduling Algorithms using OpenAirInterface," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022.

[24] S. Gebert, T. Zinner, S. Lange, C. Schwartz, and P. Tran-Gia, "Discrete-Time Analysis: Deriving the Distribution of the Number of Events in an Arbitrarily Distributed Interval," Tech. Rep., 6 2016.

[25] G. Nardini, D. Sabella, G. Stea, P. Thakkar, and A. Virdis, "Simu5G–An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks," *IEEE Access*, 2020.

[26] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A Flexible Platform for 5G Research," *SIGCOMM Comput. Commun. Rev.*, Oct. 2014.

[27] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, "srsLTE: an open-source platform for LTE evolution and experimentation," in *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*, 2016.

[28] 3rd Generation Partnership Project (3GPP), "3GPP TS 38.306 V18.4.0 (2024-12): NR; Physical channels and modulation (Release 18)," Tech. Rep., 2024.

[29] Q. Zhu, C.-X. Wang, B. Hua, K. Mao, S. Jiang, and M. Yao, "3gpp tr 38.901 channel model," in *the wiley 5G Ref: the essential 5G reference online*. Wiley Press Hoboken, NJ, USA, 2021.