End-to-End Latency Distribution in Future Mobile Communication Networks

Philipp Schulz*†, Lyndon Ong‡, Bashar Abdullah‡, Meryem Simsek†*, and Gerhard Fettweis*†
*Technische Universität Dresden, Dresden, Germany. Email: {firstname.lastname}@ifn.et.tu-dresden.de.
†International Computer Science Institute, Berkeley, California, USA. Email: {pschulz, simsek}@icsi.berkeley.edu.
‡Ciena, 7035 Ridge Road Hanover, MD 21076 E-mail: {lyong, babdulla}@ciena.com.

***Abstract*—Future applications for next generation mobile networks demand extremely low end-to-end latency and high reliability. Besides new concepts to achieve low latency, there is also a need to prove that the requirements are met and high reliability can be guaranteed under the influence of randomly behaving traffic. Due to the complexity and the ultra-low failure rates, simulations or testbeds are cumbersome or even unfeasible. Thus, this paper aims for a flexible mathematical model that is able to capture the randomness and to analyze the end-to-end latency distribution in networks. The distribution provides not only mean values but also percentiles, which are of key importance for critical communications. Moreover, such a model may help to optimize the network, e. g., by selecting the involved nodes and improving the routing strategy.**

***Index Terms*—latency, modelling, queuing networks, 5G, URLLC.**

I. Introduction

The fifth generation of mobile networks (5G) is envisioned to enable a variety of new applications. Apart from the ever increasing data rate demands, 5G networks need to tackle extremely low end-to-end (E2E) latency and ultra-high reliability requirements [1], which is denoted as ultra-reliable low-latency communications (URLLC). The simultaneous support of multiple use cases with different constraints is expected to be realized in a flexible and autonomous network enabled by software defined networks (SDN) and network function virtualization (NFV) [2]. The E2E analysis of the required key performance indicators (KPIs) demands the joint consideration of the Radio Access Network (RAN), the fronthaul, and the network including switches, routers, and other nodes, which is highly complex and its evaluation is extremely time-consuming. Thus, an evaluation tool is introduced to analyze the E2E latency distribution of any network topology. It is designed to be universal in the sense that the model can be applied to the RAN, the fronthaul, any other part of the network individually, or the entire network.

The proposed tool is based on queuing theory and provides a mathematical model to analyze the processing time of individual nodes under given load conditions, capacities, service rates, and routing probabilities. In particular, the models build on queuing networks [3], which denote a class of stochastic models extensively used to analyze resource sharing systems such as communication and computer systems, e. g., [4], [5]. In contrast to other work, which is often limited to the derivation of the mean and variance, this article aims to find the entire latency distribution. Thereby, percentiles can be derived as well, which are of key importance for URLLC.

In the introduced framework, any network topology with any number of nodes, e. g., switches, routers, access points, etc., with any possible connection between them can be captured. Here, every node can differ from each other and can have different capacities and processing attributes. The connections reflect the (wired or wireless) links along which data is forwarded according to certain routing probabilities. Hence, the network topology can be defined by the number of nodes and the connections, which can be flexibly adjusted based on the selected probabilities, i. e., a probability of zero reflects no connection. Given the SDN and NFV features of 5G networks, these probabilities can be updated in an adaptive and autonomous manner, such that the proposed model serves as a tool for autonomous and intelligent node selection and traffic forwarding in both a centralized and distributed manner.

The main contributions are as follows. A mathematical framework to evaluate the E2E latency distribution of a network with any given topology is provided. For a realistic E2E latency evaluation, possible mutual dependencies of the service times of nodes are incorporated. Wherever available, analytical expressions, e. g., known results such as the exemplary queuing models used in this paper, can be implemented. However, other models are also applicable in the proposed general framework. If no analytic results are available, the model is still flexible by applying numerical results. For instance, distributions obtained from real data could be applied as well. The proposed model is validated through an evaluation of a basic RAN scenario containing a realistic model for Ethernet switches and comparison to simulation results. Here, $M/D/1$ queues are considered as representative models for machine type communication (MTC) traffic.

II. System Model

A. Notation

In this work, a random variable (RV) is depicted by a capital letter X, its realization by a small letter x, and its probability density function (pdf) and cumulative distribution function (cdf) by f_X and F_X, respectively. The operator $*$ denotes the convolution. The indicator function of a set A is denoted as $\mathbf{1}_A$, being one if the argument is in A and zero otherwise.

B. The Queuing Network

Each node, e. g., a base station (BS), an Ethernet switch (ES), or an edge cloud server (CS) in a RAN, is modeled as a queuing system with different properties, mainly characterized by their (external) arrival process, service time distribution, capacity, and scheduling. A detailed mapping is discussed in Sec. II-D. Accordingly, the terms node and queue are used interchangeably in this paper.

Let $M \in \mathbb{N}$ denote the number of nodes Q_i, $i \in \mathcal{M}$ in the network and $\mathcal{M} = \{1, \ldots, M\}$ be the set of all node indices. The queues are connected and form a queuing network [3], i. e., after being processed, each output of a queue Q_i will be forwarded to the queue Q_j with probability p_{ij} or will leave the network with probability p_{i0}, $i, j \in \mathcal{M}$ (cf. Fig. 1). External packets arrive at the network with a mean rate $\alpha > 0$ and enter at the queue Q_j with probability p_{0j}, $j \in \mathcal{M}$. For now, the arrival process is not further specified. The probabilities are collected in the routing matrix $P = (p_{ij})_{i,j \in \mathcal{M}}$ and the vectors $p_{0\cdot} = (p_{0j})_{j \in \mathcal{M}}$ and $p_{\cdot 0} = (p_{i0})_{i \in \mathcal{M}}$, respectively. Together, P, $p_{0\cdot}$, and $p_{\cdot 0}$ completely describe the topology of the network. It is worth noting that by setting certain elements in the routing matrix to either zero or one, one can achieve disconnection or deterministic routing between two queues, respectively.

Let λ_i denote the effective arrival rate at Q_i, $i \in \mathcal{M}$ and $\lambda = (\lambda_i)_{i \in \mathcal{M}}$. Then, presuming a stable network without dropped traffic, the following equations of traffic conservation hold

$$\lambda_j = \alpha p_{0j} + \sum_{i \in \mathcal{M}} p_{ij} \lambda_i, \quad j \in \mathcal{M}, \tag{1}$$

or, equivalently,

$$(I - P)\lambda = \alpha p_{0\cdot}, \tag{2}$$

where $I \in \mathbb{R}^{M \times M}$ is an identity matrix. For non-degenerated networks, (2) can be solved for λ. The regularity of $(I - P)$ is shown in [6]. Further, let μ_i denote the mean service rate of the queue Q_i, $i \in \mathcal{M}$. Again, the service time distribution is not further specified for now, since different distributions can be applied. The load of Q_i is denoted as $\rho_i = \lambda_i / \mu_i$.

The current state of Q_i, i. e., the number $k \in \mathbb{N}$ of packets currently waiting and being in service at Q_i, at time t is denoted as the RV $X_i(t)$ and let $\pi_{ik}(t) = \mathbb{P}[X_i(t) = k]$ be the probability of Q_i being in that state at time t. Further, let $\pi^{\mathrm{A}}_{ik}(t) = \mathbb{P}[X_i(t) = k \mid \text{Arrival at } t]$ be the conditioned probability upon arrival. In systems, where the PASTA property (Poisson Arrivals See Time Averages [7]) does not hold, typically $\pi_{ik}(t) \neq \pi^{\mathrm{A}}_{ik}(t)$. If exists, let

$$\pi_{ik} = \lim_{t \to \infty} \pi_{ik}(t), \quad \pi^{\mathrm{A}}_{ik} = \lim_{t \to \infty} \pi^{\mathrm{A}}_{ik}(t), \quad i \in \mathcal{M}, k \in \mathbb{N} \tag{3}$$

denote the steady state probabilities, which are used to measure long-term performance. Let $X(t) = (X_i(t))_{i \in \mathcal{M}}$ be the multivariate RV, containing the states of each queue, and $\pi_x(t) = \mathbb{P}[X(t) = x]$ and $\pi_x = \lim_{t \to \infty} \pi_x(t)$ denote the transient and steady state probabilities of X, respectively. Here, the steady state, i. e., the asymptotic behavior is of interest.

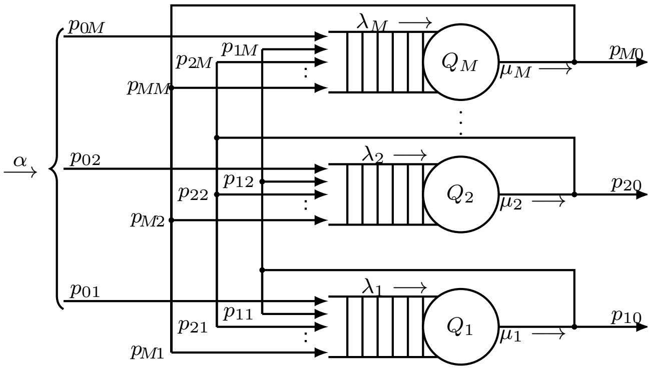


Fig. 1. A general queuing network. Traffic at queue Q_i arrives with an effective rate λ_i, is being processed with a service rate μ_i, and forwarded with probability p_{ij} to the queue Q_j or leaves the system with probability p_{i0}. External packets arrive with a rate of α and are split according to the probabilities p_{0j}, $j \in \mathcal{M}$.

C. Waiting Time, Service Time, and Sojourn Time

The latency analysis starts with the delay at each queue. The RV J_i denotes the sojourn time at the queue Q_i, i. e., the time a packet spends for waiting and service, and is presented as follows:

$$J_i = W_i + S_i + D_i, \quad i \in \mathcal{M}, \tag{4}$$

where the RVs W_i and S_i denote the waiting time and the service time at Q_i, respectively. The RV D_i is introduced to accommodate any additional delays that do not refer to queuing theory but occur in a realistic scenario, such as processing or propagation time. In particular, D_i may be deterministic. However, it is set to zero in this work. The waiting time W_i depends on the conditioned waiting times W_{ik}, which a packet experiences after arriving in state k, and so the following holds for the RV and its pdf

$$W_i = \sum_{k \in \mathbb{N}} \mathbf{1}_{\{X_i = k\}} W_{ik}, \qquad f_{W_i} = \sum_{k \in \mathbb{N}} \pi^{\mathrm{A}}_{ik} f_{W_{ik}}. \tag{5}$$

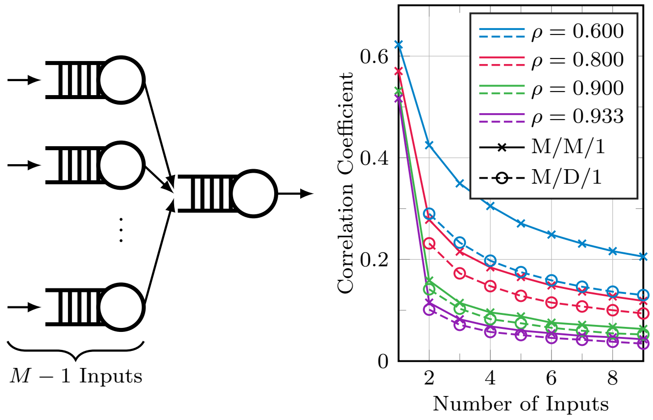
For this work, the following assumption is stated.

Assumption 1 (Independence)**.** The waiting times W_i at the queues are assumed to be independent from each other.

Assumption 1 appears to be a very strong requirement, since the queues are mutually coupled. However, according to Kleinrock's independence approximation (cf. Sec. 3.6.1 in [6]), independence can be assumed for sufficiently dense networks. Indeed, the observations depicted in Fig. 2 as well as our results presented and compared to simulations have shown that this assumption leads to good approximations, particularly for two or more inputs per queue and for high network load.

Since the analysis is cumbersome, there exist only a few specific works on dependent queuing systems, e. g., [8]. For now, the dependency is left for future studies.

In contrast, the dependency of the service time S_i between the queues is more crucial. If the service time is packet dependent, i. e., not independently drawn at each queue from the service time distribution, it has a non-negligible effect on



(a) (b)

Fig. 2. Illustration of Kleinrock’s independence approximation. A simple network (a) consisting of M queues is studied. The correlation of the waiting time at one of the $M-1$ input queues and the output queue is shown in (b) for different network loads and queuing models.

the overall latency. Thus, the service time S_i at the queue Q_i is assumed to be a scaled version, of an initially drawn packet size S_0:

$$S_i = \mu_i^{-1} S_0, \quad i \in \mathcal{M}. \tag{6}$$

This approach requires that the queues have identical service time distributions, at least up to scale.

Now, let $q = (q_1, \ldots, q_n) \in \mathcal{M}^n$ be a path in the network. Then, the overall waiting time W_q, overall service time S_q and overall additional delay D_q along the path q is the sum of the single waiting times, service times, and delays respectively

$$W_q = \sum_{i=1}^{n} W_{q_i}, \qquad S_q = \sum_{i=1}^{n} S_{q_i}. \qquad D_q = \sum_{i=1}^{n} D_{q_i}. \tag{7}$$

and the overall latency or sojourn time is given by

$$J_q = \sum_{i=1}^{n} J_{q_i} = W_q + S_q + D_q. \tag{8}$$

Eq. (6) leads to

$$S_q = \left(\sum_{i=1}^{n} \mu_{q_i}^{-1} \right) S_0 =: \mu_q^{-1} S_0, \tag{9}$$

and with Assumption 1, Eqs. (7) can be expressed as pdfs as

$$f_{W_q} = f_{W_{q_1}} * f_{W_{q_2}} * \cdots * f_{W_{q_n}} =: \mathop{\Large *}_{i=1}^{n} f_{W_{q_i}}, \tag{10}$$

$$f_{S_q}(t) = \mu_q f_{S_0}(\mu_q t), \tag{11}$$

which leads to the pdf of the overall sojourn time along path q

$$f_{J_q}(t) = \mathop{\Large *}_{i=1}^{n} f_{W_{q_i}}(t) * \mu_q f_{S_0}(\mu_q t) * f_{D_q}(t). \tag{12}$$

Depending on the considered distributions, Eq. (12) may be solved analytically, or, in the more general case, numerically.

D. A Queuing Model of an Access Network

In the following, the model is mapped to a realistic RAN. Its components, i. e., BSs, ESs, and CSs can be modeled as queuing systems to evaluate the delay behavior of each of them but for now only the ESs are in focus. Fig. 3 (a) shows the structure of a realistic ES, which is a key component in RANs. Preceding nodes are connected to the input ports (ingress) on the left. Each input is equipped with a meter that drops traffic exceeding any predefined service level agreement (SLA). In this work, the metering is not modeled explicitly, but is implemented by limiting the arriving traffic accordingly.

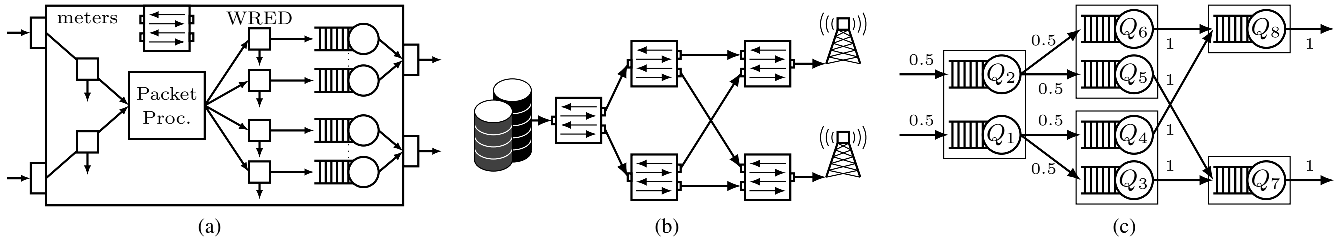
The incoming packets are processed at the packet processing unit that determines to which port at the egress side they will be forwarded. The packet processing is designed such that it is capable to process the accepted packets with a deterministic delay that depends only on packet size, which can be captured by the RV D_i. Each output port has multiple first-in first-out (FIFO) buffers for different service priorities, so packets may have to wait for others. The model evaluates the occurring waiting and service times W_i and S_i. Typically, in Common Public Radio Interface (CPRI)-based networks [9], at least three buffers per output are used for data, control and management, and timing synchronization data. For the sake of simplicity, only the data traffic and, thus, only one buffer per port is considered here. It is assumed that the other traffic is assigned a fixed amount of resources. Each buffer applies weighted random early discard (WRED [10]), i. e., randomly dropping packets to avoid congestion based on the instantaneous queue utilization, configured weights and thresholds. The described behavior is modeled by representing the ESs by multiple queues. Fig. 3 (c) shows this principle by inserting the ES structure (a) into the topology (b).

III. Exemplary Application Scenario

The scenario, shown in Fig. 3 (b), consists of two BSs, one CS, and five ESs. This work focuses on the ESs. Two traffic models are studied and described in the following sections. In [11] it is shown how general traffic models can be analyzed. It should be noted that the choice of the traffic model is crucial as shown in [12].

A. A Scenario with Exponentially Distributed Service Times

Here, each of the ESs is modeled as an $M/M/1$ queue with FIFO scheduling (cf. Sec. II-D). Other scheduling schemes, e. g., egalitarian processor sharing (EPS), could be applied, too. External packets arrive with a rate $\alpha = 0.1$ and the service rates are set to $\mu_i = 0.15$ per time unit for $i \in \mathcal{M}$. The routing probabilities are set symmetrically. For $M/M/1$ queues, Jackson’s Theorem [3] provides the joint steady state probabilities as $\pi_x = \prod_{i=1}^{M} \left[\rho_i^{x_i} (1 - \rho_i) \right]$. In addition, the packet size is assumed as $S_0 \sim \operatorname{Exp}(1)$ and S_i is set according to Eq. (6). The components of the waiting time are $W_{i0} = 0$ and $W_{ik} \sim \Gamma(k, \mu_i)$, since they are the sum of k independent exponentially distributed RVs. In particular, the remaining service time of the already running packet at the first position is still exponentially distributed, due to the memoryless nature



(a)

(b)

(c)

Fig. 3. Modeling of ES behavior in a queuing network. (a) ES structure. Input ports (left) are equipped with meters that drop traffic which violates SLAs. Each output port (right) has its own buffers. (b) The investigated network. (c) The queuing network, when (a) is put into (b) and only one buffer per output.

of the exponential distribution. It should be noted that, strictly spoken, the used $M/M/1$ results only provide approximations, since the service times at each queue are coupled by Eq. (6).

B. A Scenario with Deterministic Service Times

Here, the service times at the outputs of the ESs are assumed to be deterministic, i. e., $S_i = \mu_i^{-1}$. In contrast, the arrival process to the overall network is still set to be Poisson with rate α. Thus, the network consists of pure $M/D/1$ FIFO queues at the ingress, but each queue which is fed by another queue of the system has a general arrival process and is therefore $G/D/1$. State probabilities for $M/D/1$ queues are given by [13]

$$\pi_{i0} = 1 - \rho_i, \qquad \pi_{i1} = (1-\rho_i)(\mathrm{e}^{\rho_i} - 1), \tag{13}$$

$$\pi_{ik} = (1-\rho_i)\left(\mathrm{e}^{k\rho_i} + \sum_{j=1}^{k-1} \mathrm{e}^{j\rho_i}(-1)^{k-j}\left[\frac{(j\rho_i)^{k-j}}{(k-j)!} + \frac{(j\rho_i)^{k-j-1}}{(k-j-1)!}\right]\right),$$

for $k \geqslant 2$. Eqs. (13) do not hold in the $G/D/1$ case. For such a queue Q_j, the probabilities are approximated by a discrete Markov chain as follows. At each time step one packet is served, but each preceding queue Q_i adds a packet with the probability $(1 - \pi_{i0})p_{ij}$ of being active and forwarding to Q_j. Thus, transition rates can be obtained iteratively and the state probabilities are obtained from the resulting equilibrium equation.

The waiting time W_i at each queue Q_i is obtained from Eq. (5) as follows. $W_{i0} = 0$, since there is no waiting at an empty queue, and $W_{i1} \sim \mathcal{U}(0, \mu^{-1})$ with $\mathcal{U}(a, b)$ being the uniform distribution on the interval (a, b), since for the packet currently being served all remaining times between 0 and μ^{-1} are equally likely. Finally,

$$W_{ik} = (k-1)\mu^{-1} + W_{i1}, \tag{14}$$

or equivalently

$$W_{ik} \sim \mathcal{U}\left((k-1)\cdot\mu^{-1}, k\mu^{-1}\right) \tag{15}$$

for $k > 1$.

IV. Numerical Evaluation

For validation, a discrete event simulator was implemented that creates packets at time instances from the arrival distribution and samples sizes from the distribution of S_0. When being in service, the remaining time is determined through the packet size and the service rate of the queue. The routing is handled according to the routing probabilities. A FIFO scheduler lets a packet wait for all others that arrived before.

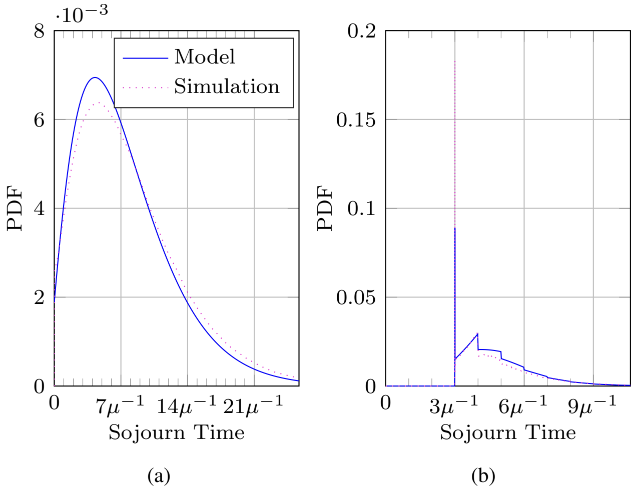


Fig. 4. The modeled and simulated latency along (Q_1, Q_3, Q_7) for (a) exponential and (b) deterministic service. The legend applies to both figures.

A. Model Validation and Results

Fig. 4 depicts the latency distribution for both traffic models along the path (Q_1, Q_3, Q_7). It turns out that the model curve approximates the simulation results well. The remaining gap between both curves is due to the approximation error of Assumption 1 and the approximations of the state probabilities.

One useful result of the model is the fact that latency guarantees can be generated directly from the percentiles of the resulting distribution. Thus, statements such as “*90 % of the packets experience a latency less than $7\mu^{-1}$ in the deterministic scenario*” can be made, which is very important in the context of URLLC, where certain latency bounds have to be guaranteed.

The results depicted in Fig. 4 show that the choice of the service time distribution has a crucial impact. Even though both models implement the same mean arrival and service rates, the deterministic service rates result in a much more concentrated distribution, since the absence of any variation in the service times is beneficial for the system. However, the deterministic service times also lead to a distinct minimum

TABLE I
COMPUTATIONAL EFFORT. COMPARISON OF SIMULATION AND MODEL.

	Simulation		Model	
Part	Compl. $O(\cdot)$	Comp. [s]	Compl. $O(\cdot)$	Comp. [s]
Sim.	$N_\mathrm{F} M$	103.72	–	–
(5)	–	–	$n N_\mathrm{G} k_\mathrm{max}$	12.28
Pdf	$N_\mathrm{F} N_\mathrm{G}$	41.94	$n N_\mathrm{G} \log N_\mathrm{G}$	5.28
Total	$N_\mathrm{F}(N_\mathrm{G} + M)$	145.65	$n N_\mathrm{G}(\log N_\mathrm{G} + k_\mathrm{max})$	18.16

Remark: The computation times are averaged over 100 runs performed on an Intel® Xeon® CPU E5-2699A v4, 2.40GHz for exponential service.

service time of $3\mu^{-1}$, where a peak indicates the probability of experiencing this minimum in a completely empty system.

B. Computational Effort

Obtaining reliable results from a simulation requires a certain minimum simulation time to capture all possible constellations that can occur within the system, which increases with the system complexity. For instance, a failure rate of 10^{-6} requires a significantly higher number than 10^6 of instances to be detected. Extremely low failure rates of 10^{-9}, as required by some URLLC use cases, become infeasible for simulations. Here, only $N_\mathrm{F} = 10^6$ packets were generated.

A mathematical model typically does not suffer from such limitations. However, some of the computations have to be performed numerically. Thus, accuracy as well as computation time both scale with the number of cells N_G in the underlying grid of the involved functions. The numerical effort is governed by the calculation of Eq. (5), which linearly scales with N_G and the maximum number of considered states k_max, and the convolution in Eq. (10) that scales as $N_\mathrm{G} \log N_\mathrm{G}$ and the length of the investigated path n. For the presented results, a grid with $N_\mathrm{G} = 40{,}000$ cells was used.

Table I shows the differences in complexity and computation time. In the example the simulation took 146 s, compared to the significantly lower 18 s for the model evaluation.

V. CONCLUSION AND OUTLOOK

A mathematical framework to study E2E latency in future mobile networks is presented. The queuing models are kept simple for now to prove feasibility and accuracy, but may be replaced in future by more complex ones. However, since Assumption 1 is based on dense networks, it is expected that the approximation improves for more sophisticated setups.

The model provides insights into the latency along different paths in a network. Hence, it can be used to find optimal paths within a network for URLLC traffic. Further, it provides a valuable foundation to optimize the network. Traffic can be steered by adjusting the routing matrix P. Suitable optimization algorithms are left for further studies.

REFERENCES

[1] P. Schulz *et al.*, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, February 2017.

[2] M. Simsek *et al.*, "On the Flexibility and Autonomy of 5G Wireless Networks," *IEEE Access*, vol. 5, 2017.

[3] J. R. Jackson, "Networks of Waiting Lines," *Operations Research*, vol. 5, no. 4, pp. 518–521, Feb. 1957.

[4] M. Mashaly and P. J. Kuehn, "Modeling and Analysis of Virtualized Multi-Service Cloud Data Centers with Automatic Server Consolidation and Prescribed Service Level Agreements," in *2016 IEEE 41st Conference on Local Computer Networks Workshops*, Nov 2016, pp. 9–16.

[5] Y. Xu *et al.*, "Impact of Flow-level Dynamics on QoE of Video Streaming in Wireless Networks," in *IEEE INFOCOM*, April 2013.

[6] D. Bertsekas and R. G. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1992.

[7] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, 1982.

[8] S. Calo, "Message delays in repeated-service tandem connections," *IEEE Trans. on Communications*, vol. 29, no. 5, pp. 670–678, May 1981.

[9] eCPRI Specification V2.0, "Common public radio interface: ecpri interface specification," CPRI initative, Tech. Rep., May 2019.

[10] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, Aug. 1993.

[11] P. Schulz *et al.*, "End-to-End Latency Analysis in Wireless Networks with Queuing Models for General Prioritized Traffic," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2019, pp. 1–6.

[12] M. Laner *et al.*, "End-to-end delay in mobile networks: Does the traffic pattern matter?" in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, 8 2013, pp. 1–5.

[13] K. Nakagawa, "On the Series Expansion for the Stationary Probabilities of an M/D/1 queue," *Journal of the Operations Research Society of Japan*, vol. 48, pp. 111–122, 2005.