Adaptive management of communication resource allocation in high-load 5G infrastructures: a queuing-based approach*

Viacheslav Kovtun[1,*,†], Oksana Kovtun[2,†], Tetiana Gryshchuk[3,†], Maria Yukhimchuk[3,†]

[1] Institute of Theoretical and Applied Informatics Polish Academy of Sciences, Bałtycka Str., 5, Gliwice, 44-100, Poland
[2] Vasyl' Stus Donetsk National University, 600-richchya Str., 21, Vinnytsia, 21000, Ukraine
[3] Vinnytsia National Technical University, Khmelnytske shose, 95, Vinnytsia, 21021, Ukraine

Abstract

The rapid evolution of 5G networks demands advanced methodologies for optimizing communication resource allocation, particularly under high-load conditions with fluctuating traffic patterns. This paper presents a novel adaptive model for managing the distribution of communication resources in 5G infrastructures, utilizing a queuing system with delay. The proposed approach accounts for subscriber mobility, traffic irregularities, and peak load conditions, enabling real-time optimization of base station utilization. By integrating probability distribution functions with delay, the model enhances service quality by reducing waiting times and minimizing energy consumption. The study provides analytical formulations for key performance metrics, including queue waiting time, base station utilization, and variation coefficients. Experimental validation confirms the efficiency of the model in comparison with classical queuing approaches, demonstrating its potential for intelligent traffic management in next-generation networks.

Keywords

5G networks, intelligent traffic management, communication resource allocation, queuing systems with delay, base station utilization, adaptive network optimization, service quality enhancement

1. Introduction

The ongoing digital transformation requires a communication infrastructure capable of ensuring high data transmission speeds, low latency, and scalability for millions of devices operating simultaneously within the network. In this context, 5G technology has become a cornerstone in the development of telecommunication systems, enabling the management of high-load networks with diverse use-case scenarios such as the Internet of Things, augmented reality, autonomous vehicles, and smart cities [1-5]. However, as the number of subscribers and devices continues to grow, challenges emerge in effectively managing utilization, particularly under conditions of traffic irregularities, high subscriber mobility, and variable input request density.

Ensuring connection stability under peak loads is particularly crucial, as traditional communication resource management models exhibit limited efficiency [6]. Neglecting behavioral patterns of network load, such as the delay between the arrival of incoming requests and their processing, may lead to a critical decline in service quality for subscribers. This justifies the relevance of developing adaptive models capable of maintaining high performance and reliability of communication systems even under extreme operating conditions.

Intelitsis'25: The 6th International Workshop on Intelligent Information Technologies & Systems of Information Security, April 04, 2025, Khmelnytskyi, Ukraine
* Corresponding author.
† These authors contributed equally.
vkovtun@iitis.pl (V. Kovtun); o.kovtun@donnu.edu.ua (O. Kovtun); thryshuk@gmail.com (T. Gryshchuk); umcmasha@gmail.com (M. Yukhimchuk)
0000-0002-7624-7072 (V. Kovtun); 0000-0002-9139-8987 (O. Kovtun); 0000-0002-4308-2654 (T. Gryshchuk); 0000-0002-8131-9739 (M. Yukhimchuk)

Managing the distribution of utilization in high-load 5G infrastructures is a critical task for ensuring connection stability and efficient utilization of network resources [7, 8]. In response to this challenge, the article presents a corresponding concept formalized within the framework of queuing theory [9-12]. To assess the uniqueness and effectiveness of the proposed approach, existing counterparts were examined to identify their limitations.

The $\frac{M}{\frac{M}{1}}$-type queuing system model is widely used for analyzing network systems [13, 14]. It assumes an exponential distribution of intervals between incoming requests and their service duration, which significantly simplifies calculations. However, it has serious limitations: it does not account for traffic irregularities and peak loads, does not allow for effective delay prediction under variable loads, and its sensitivity to increasing base station utilization leads to an exponential rise in the average waiting time for accepted service requests. Additionally, the $\frac{M}{\frac{M}{1}}$-model does not consider subscriber mobility, which is a key factor in 5G networks. These limitations make it ineffective for real-world networks with dynamic traffic and uneven spatial distribution of subscribers.

The more generalized $\frac{G}{\frac{G}{1}}$-model allows for arbitrary distributions of incoming request arrival times and service durations, significantly improving adaptation to real-world information and communication scenarios [12, 15, 16]. However, more complex analytical and numerical methods are required for parameter evaluation. It does not account for the specific characteristics of 5G infrastructure, where significant load fluctuations can occur due to subscriber mobility and dynamic changes in traffic density. Additionally, the complexity of the $\frac{G}{\frac{G}{1}}$-model makes its real-time implementation impractical, as it necessitates computing a large number of probabilistic characteristics.

Some researchers propose using machine learning (ML) to predict traffic and optimize resource allocation in 5G networks [17, 18]. The main advantages of this approach include automated model training based on operational logs of target network infrastructures, enabling proactive adjustments in communication resource management, and facilitating rapid adaptation of the framework to changing load conditions. However, ML-based approaches require significant computational resources and prior model training, making real-time implementation challenging. Additionally, the effectiveness of ML methods depends on the availability of high-quality data, which can be difficult to obtain in high-load 5G networks.

A distinct role in structuring 5G infrastructures is played by Network Slicing (NS) technology [19, 20], which enables the allocation of utilization among different categories of subscribers based on their specific needs (eMBB, URLLC, mMTC). However, complex resource allocation management between slices under variable load conditions and the high costs associated with implementing and maintaining such an architecture are its primary drawbacks. Additionally, NS technology relies on complex optimization algorithms, which can significantly increase latency and lead to inefficient network utilization.

Thus, existing solutions have significant limitations in the context of communication resource management in high-load 5G infrastructures. The article presents a concept that eliminates these limitations, based on the formalization of the research object, goal, and tasks, formulated considering the results of a critical analysis of existing approaches.

Research Object: The process of managing the utilization distribution in a high-load 5G infrastructure.

Research Goal: Formalizing a mathematical framework for efficient communication resource management in 5G infrastructure by reducing latency.

Research Tasks:

1. Identify behavioral patterns of network load, including changes in subscriber density and their mobility between base stations.

2. Develop probability distribution functions to describe the intensity of incoming requests and service duration, considering latency.
3. Formalize mathematical expressions for key model parameters, particularly for the average waiting time of an accepted incoming request, variation coefficients, and other characteristics.
4. Compare the developed mathematical framework with a classical counterpart.
5. Investigate the impact of latency and base station utilization on service quality in a high-load 5G infrastructure.

2. 2. Models and methods

2.1. Research statement

Let us focus on the process of managing the utilization distribution in a densely populated (high-load) 5G infrastructure. Considering the specificity of such an object of study, an adequate description of the process is possible only if behavioral patterns of network load are considered – particularly changes in subscriber density during peak hours and the rapid movement of mobile subscribers between base stations.

To account for these network load behavioral patterns, we employ a recurrent modeling framework with delay [13, 14]. This mathematical framework enables proactive resource demand forecasting and adaptive redistribution in real-time. The recurrent approach helps minimize management delays, ensure seamless connectivity during subscriber mobility, and optimize energy consumption, which is critical for maintaining the high efficiency of 5G networks.

We introduce the concept of a recurrent flow, defined by a set of probability distribution functions $C_1(t) = C_2(t) = \cdots C_k(t) = C(t)$ between incoming subscriber requests. Consider a scenario in which the quality-of-service system manages two flows, each defined by probability distribution functions of type $C(t) = \begin{cases} 1 - exp\big(-\varphi(t - t_0)\big) \forall t \geq t_0, \\ 0 \forall 0 \leq t < t_0, \end{cases}$ with identical delays equal to t_0, where φ is the distribution rate parameter in the system. This parameter determines how rapidly the probability accumulation changes after the delay t_0 moment.

The studied process of communication resource management is defined as a queuing system, where subscriber requests arrive at the input, with stochastic intervals between them determined by a probability distribution function of the form:

$$\alpha(t) = \begin{cases} \mu\, exp\big(-\mu(t - t_0)\big) \forall t \geq t_0, \\ 0 \forall 0 \leq t < t_0, \end{cases} \quad (1)$$

where μ is a parameter that characterizes the arrival intensity of subscriber requests at the system's input.

In turn, the service duration of accepted requests is determined by a probability distribution function of the form:

$$\beta(t) = \begin{cases} \eta\, exp\big(-\eta(t - t_0)\big) \forall t \geq t_0, \\ 0 \forall 0 \leq t < t_0, \end{cases} \quad (2)$$

where η is a parameter that characterizes the service intensity of accepted subscriber requests within the system.

The probability distribution functions (1) and (2) are shifted to the right relative to the zero reference point by the magnitude of the delay t_0. These functions are exponential, with controlled parameters (μ, t_0) and (η, t_0), respectively. Moreover, $\mu < \eta$. We now analyze the dynamic properties of key qualitative parameters of the studied system, particularly the arrival intervals of subscriber requests $\alpha(t)$ and the service duration of accepted requests $\beta(t)$.

2.2. The concept of managing the utilization distribution in a high-load 5G infrastructure

To determine the numerical characteristics of the qualitative parameters $\alpha(t)$ and $\beta(t)$, we apply the Laplace transform to functions (1) and (2):

$$A^*(z) = \frac{\mu\, exp(-t_0 z)}{z+\mu}, \tag{3}$$

$$B^*(z) = \frac{\eta\, exp(-t_0 z)}{z+\eta}, \tag{4}$$

where z is a variable used in the Laplace transform to transition from a time-domain function to a frequency-domain function. The first derivative of function $A^*(z)$ is given by $\frac{dA^*(z)}{dz} = \frac{(-\mu t_0(z+\mu)-\mu)\, exp(-t_0 z)}{(z+\mu)^2}$, $\left.\frac{dA^*(z)}{dz}\right|_{z=0} = \frac{\mu^2 t_0+\mu}{\mu^2} = \frac{1}{\mu} + t_0$. Based on these analytical expressions, we formalize the mathematical expectation of the arrival interval of subscriber requests as:

$$\bar{T}_\mu = \frac{1}{\mu} + t_0. \tag{5}$$

In the context of expression (5), it can be stated that the arrival intensity of subscriber requests μ' in the studied queuing model is determined through the parameters of the probability distribution function (1):

$$\mu' = \frac{\mu}{(1+\mu t_0)}. \tag{6}$$

We derive analogous expressions to (5) and (6) for the service duration of accepted requests (4):

$$\bar{T}_\eta = \frac{1}{\eta} + t_0, \tag{7}$$

$$\eta' = \frac{\eta}{(1+\eta t_0)}, \tag{8}$$

where expression (8) is obtained from expression (7) through the parameters of the probability distribution function (2).

In the context of arrival intensities (6) and service intensities (8) of incoming subscriber requests, it is observed that the utilization u of a base station in the cluster of the studied 5G infrastructure has increased by a factor $\frac{(1+\eta t_0)}{(1+\mu t_0)}$ relative to the corresponding characteristic of the $\frac{M}{\frac{M}{1}}$-type system:

$$u = \frac{\mu'}{\eta'} = \frac{\mu(1+\eta t_0)}{\eta(1+\mu t_0)}. \tag{9}$$

The utilization u can also be interpreted as $u = \frac{\bar{T}_\eta}{\bar{T}_\mu}$. This fact allows the use of the controlled parameters from expressions (5) and (7) as input variables for the studied system. This approach enables adaptive resource management, where arrival intensity and service intensity dynamically adjust based on real-time network conditions, optimizing communication resource distribution in the high-load 5G infrastructure.

We introduce the variance of intervals for the probability distribution function (1): $\sigma_\mu^2 = \frac{1}{\mu^2}$, and express the coefficient of variation $CV_\mu = \frac{\sqrt{\sigma_\mu^2}}{\bar{T}_\mu}$ through it. The latter, considering expression (5), can be represented in the form:

$$CV_\mu = \frac{1}{(1+\mu t_0)}. \tag{10}$$

Similarly, for the probability distribution function (2), we introduce the variance $\sigma_\eta^2 = \frac{1}{\eta^2}$ and define the coefficient of variation as:

$$CV_\eta = \frac{\sqrt{\sigma_\eta^2}}{\bar{T}_\eta} = \frac{1}{(1+\eta t_0)}. \tag{11}$$

Further, we take into account that for $t_0 > 0$, $\mu > 0$, $\eta > 0$, the values of the coefficients of variation from expressions (10) and (11) are less than one. This implies that the distribution of arrival intervals and service durations exhibits low variability, meaning that the process is relatively stable

and predictable, which is critical for ensuring efficient resource allocation in a high-load 5G infrastructure.

Based on the characteristics of the studied system, such as (5), (7), (10), and (11), several assumptions can be made:

1. Considering the delay in time leads to an increase in the utilization of the base station cluster in the studied 5G infrastructure by a factor of $\frac{(1+\eta t_0)}{(1+\mu t_0)}$ relative to the corresponding metric in the classical $\frac{M}{\frac{M}{1}}$-type system;
2. Since the coefficient of variation values from expressions (10) and (11) are less than one, it can be concluded that the model of the studied system is non-Markovian. Thus, for the same value of (9), the average waiting time for an incoming subscriber request in the studied system should be shorter than the corresponding metric in the $\frac{M}{\frac{M}{1}}$-type system;
3. Unlike the $\frac{M}{\frac{M}{1}}$-type system, the use of probability distribution functions (1) and (2) allows for approximating the distributions of controlled parameters at the level of their first two moments.

In further studies of the recurrent model with delay defined in Section 2.1, we focus on its convergence to the $\frac{G}{\frac{G}{1}}$-type system. In such a queuing system, the states at the moment t depend on previous system states, which can be formalized using recurrent equations with delay. The states of the $\frac{G}{\frac{G}{1}}$-type system at the moment t are uniquely characterized by known integral equations of spectral decomposition [21, 22], which are related through the Laplace transform to Lindley's integral equation:

$$F(x) = \begin{cases} \int_{-\infty}^{x} F(x-w)d\,R(w) \forall x \geq 0, \\ 0, x < 0, \end{cases}. \quad (12)$$

where $F(x)$ is the probability distribution function of x, which represents the waiting time of an accepted incoming request in the buffer; w is a stochastic variable characterizing the service duration or the time a request spends in the system after being accepted; $R(w)$ is the probability distribution function of the boundary stochastic value $W = \lim_{k\to\infty} W_k = s_k - t_{k+1}$, where s_k represents the service duration of an incoming request R_k, and t_{k+1} is the interval between the arrival of requests R_k and R_{k+1} at the system's input. This formulation enables a precise mathematical description of dynamic processes in the high-load 5G infrastructure, contributing to adaptive communication resource management.

The solution of equation (12) using the spectral method results in the product

$$A^*(-\tau)B^*(\tau) - 1, \quad (13)$$

which is a rational function [15, 16], where the complex variable τ is used in the Laplace transform to represent probability distribution functions in the frequency domain. Thus, to determine the distribution function from (12), we need to find the spectral decomposition of the form:

$$. A^*(-\tau)B^*(\tau) - 1 = \frac{\Phi_+(\tau)}{\Phi_-(\tau)}, \quad (14)$$

where $\Phi_+(\tau)$, $\Phi_-(\tau)$ represents fractional-rational functions, which must satisfy the following conditions:

- $\forall\, Re(\tau) > 0$, the function $\Phi_+(\tau)$ is analytic and has no zeros in this half-plane, and the equality $\lim_{|\tau|\to\infty, \mathrm{Re}(\tau)>0} \Phi_+(\tau)/\tau = 1$ must hold.

- $\forall Re(\tau) > \sigma^2$, the function $\Phi_-(\tau)$ is analytic and has no zeros in this half-plane, and the equality $\lim_{|\tau|\to\infty, \mathrm{Re}(\tau)<\sigma^2} \frac{\Phi_+(\tau)}{\tau} = 1$ must hold, where σ^2 is determined by the condition $\lim_{t\to\infty} \frac{\alpha(t)}{exp(-\sigma^2 t)} < \infty$.

We represent the probability distribution functions (1) and (2) on the basis of expression (14), taking into account the Laplace transforms (3) and (4), respectively. The solution of equation (12) for this case takes the form:

$$\frac{\Phi_+(\tau)}{\Phi_-(\tau)} = \frac{\mu\, exp(-t_0\tau)}{\mu-\tau}\frac{\eta\, exp(-t_0\tau)}{\eta-\tau} - 1 = \frac{\tau(\tau-\mu-\eta)}{(\mu-\tau)(\eta+\tau)}. \quad (15)$$

As we can see, during the transformation process of functions (1) and (2) into the form (15), the delay was lost. At the same time, it is worth noting that function (15) also serves as a spectral decomposition for the solution of the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system. However, this identity is purely formal since, unlike the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, the parameters μ and η in the queuing system with delay, defined in Section 2.1, are not interpreted as the arrival intensity and service intensity of incoming requests, respectively.

We now unveil the essence of the controlled parameters μ and η within the framework of the queuing system with delay, as presented in Section 2.1. For further transformations, we define $\Phi_+(\tau) = \frac{\tau(\tau+\eta-\mu)}{\eta+\tau}$, $\Phi_-(\tau) = \mu - \tau$. These functions do not have zeros or poles in their respective domains $Re(\tau) > 0$, $Re(\tau) < \mu$. In analytical form, the Laplace transform of the probability distribution function $F(x)$, introduced in (12), takes the form $\Psi_+(\tau) = \frac{V}{\Phi_+(\tau)}$, where the constant V is defined as $V = \lim_{\tau\to 0}\frac{\Phi_+(\tau)}{\tau} = \lim_{\tau\to 0}\frac{\tau+\eta-\mu}{\tau+\eta} = 1 - \frac{\mu}{\eta}$. In turn, the parameters μ, η are determined by equations (5) and (7), respectively, while the ratio $\frac{\mu}{\eta}$ characterizes the utilization u (similar to the analysis of the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system). Thus, the input parameters of the studied queuing system with delay are the mathematical expectations $\bar{\mathrm{T}}_\mu$, $\bar{\mathrm{T}}_\eta$ (see expressions (5) and (7), respectively), and coefficients of variation CV_μ, CV_η (see expressions (10) and (11), respectively).

Considering that $\Psi_+(\tau) = \frac{V}{\Phi_+(\tau)} = \frac{\left(1-\frac{\mu}{\eta}\right)(\eta+\tau)}{\tau(\tau+\eta-\mu)}$, we define the probability density function $F^*(\tau)$ as:

$$F^*(\tau) = \tau\Phi_+(\tau) = \frac{\left(1-\frac{\mu}{\eta}\right)(\eta+\tau)}{\tau+\eta-\mu}. \quad (16)$$

The derivative of function (16) is given by:

$$\frac{dF^*(\tau)}{d\tau} = \frac{\left(1-\frac{\mu}{\eta}\right)(\tau+\eta-\mu)-\left(1-\frac{\mu}{\eta}\right)(\tau+\eta)}{(\tau+\eta-\mu)^2} = -\frac{\mu\left(1-\frac{\mu}{\eta}\right)}{(\tau+\eta-\mu)^2}. \quad (17)$$

For $\tau = 0$, expression (17) takes the form $\left.\frac{dF^*(\tau)}{d\tau}\right|_{\tau=0} = \frac{\mu\left(1-\frac{\mu}{\eta}\right)}{(\tau-\mu)^2}$, from which the average waiting time for an accepted incoming subscriber request in the buffer of the queuing system with delay, as presented in Section 2.1, is given by:

$$\bar{F} = \frac{\left(\frac{\mu}{\eta}\right)}{(\eta-\mu)}. \quad (18)$$

To complete the formalization of the concept of managing communication resource distribution in a high-load 5G infrastructure, we formulate the methodology for calculating unknown parameters of the studied queuing system with delay. For this purpose, we define the qualitative parameters μ, η, t_0 based on the analytical expressions (5), (7), (10), and (11), which are oriented toward the

calculation of numerical characteristics of the probability distribution functions (1) and (2). We introduce a system of equations of the form:

$$\begin{cases} \frac{1}{\mu} + t_0 = \bar{T}_\mu, \\ \frac{1}{(1+\mu t_0)} = CV_\mu, \\ \frac{1}{\eta} + t_0 = \bar{T}_\eta, \\ \frac{1}{(1+\eta t_0)} = CV_\eta. \end{cases} \quad (19)$$

We use the numerical characteristics on the right-hand side of the system of equations (19) to determine the desired qualitative parameters μ, η, t_0. The system (19) exhibits redundancy, which we overcome by introducing input parameters $\bar{\mathrm{T}}_\mu$, $\bar{\mathrm{T}}_\eta$, CV_μ, CV_η. From the first equation of system (19), we express the qualitative parameter t_0 as

$$t_0 = \bar{T}_\mu - \frac{1}{\mu}. \quad (20)$$

From the second equation, we derive $CV_\mu = \frac{1}{(1+\mu\bar{T}_\mu - 1)}$. From the last expression, we obtain:

$\mu = \frac{1}{(\bar{T}_\mu CV_\mu)}$. (21)

From the third equation, we express $\bar{\mathrm{T}}_\eta$: $\bar{T}_\eta = \bar{T}_\mu + \frac{1}{\eta} - \frac{1}{\mu}$. Generalizing the obtained results, we write:

$$\eta = \frac{1}{\left(\bar{T}_\eta - \bar{T}_\mu(1-CV_\mu)\right)}. \quad (21)$$

Finally, we process the fourth equation, taking into account the previously derived analytical constructions:

$$CV_\eta = \frac{1}{\left(1+\frac{\bar{T}_\mu(1-CV_\mu)}{\bar{T}_\eta - \bar{T}_\mu(1-CV_\mu)}\right)} = 1 - \frac{1-CV_\mu}{u} = 1 - u\left(1 - CV_\mu\right), \quad (22)$$

where $u = \frac{\bar{T}_\eta}{\bar{T}_\mu}$. Thus, we have expressed the unknown parameters μ, η, t_0 of the studied queuing system with delay through a set of parameter values $\bar{\mathrm{T}}_\mu$, $\bar{\mathrm{T}}_\eta$, CV_μ, CV_η.

Overall, Section 2 presents a mathematical model of a queuing system with delay, which replicates the process of managing communication resource distribution in a high-load 5G infrastructure. The foundation of the model is the introduction of probability distribution functions with delay (expressions (1) and (2)) to describe behavioral patterns of incoming request flows and their service durations. Key analytical expressions are introduced, including request arrival intensity μ' (expression (6)) and service intensity η' (expression (8)), average waiting time $\bar{F}$ (expression (18)), coefficients of variation CV_μ, CV_η (expressions (10) and (11)), and spectral approach for determining waiting time distribution functions (expression (15)). The model accounts for dynamic network load variations, enabling efficient resource demand forecasting, latency reduction, energy consumption optimization, and uninterrupted connectivity. Due to its flexibility, the proposed approach is well-suited for analyzing and managing resources in complex 5G infrastructure scenarios, particularly under variable subscriber density and high subscriber mobility conditions.

3. Results and discussion

To demonstrate the practical value of the proposed communication resource management concept from Section 2, we analyze its application in a 5G infrastructure under three operating modes: Low-load mode, Medium-load mode, and High-load mode.

Assume that as a result of censored observation of the target 5G infrastructure, the following values of the system's input parameters in a queuing system with delay have been determined: $\bar{T}_\mu =$

10, $\bar{T}_\eta = 1$, $CV_\eta = 0.5$, corresponding to a Low -load mode. Then, for utilization $u = 0.1$, expression (23) yields $CV_\mu = 0.95$. Considering the known input parameters, expression (21) allows us to calculate the parameter μ: $\mu = \frac{1}{9{,}5=\frac{2}{19}}$, while expression (22) determines the parameter η: $\eta = \frac{1}{(1-10\cdot 0.05)} = 2$. For the target 5G infrastructure, determine the average waiting time $\bar{F}$ for the accepted incoming subscriber request in the buffer using expression (18): $\bar{F} = \frac{\left(\frac{1}{19}\right)}{\left(\frac{36}{19}\right)=\frac{1}{36}}$.

It should be noted that for the classical $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, considering the same input parameter values (including $u = 0.1$, $\eta = 1$), the average waiting time $\bar{F}$ equals $\bar{F} = \frac{\left(\frac{u}{\eta}\right)}{(1-u)} = \frac{1}{9}$, which is four times higher than the recently calculated value for the queueing system with delay. However, if parameters $\mu = \frac{2}{19}$ and $\eta = 2$ are interpreted as the arrival rate of incoming requests and the service rate, respectively, then for the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, we obtain $u = \frac{1}{19} < 0.1$ and $\bar{F} = \frac{1}{36}$. As we can see, with identical parameter μ, η values for both the queueing system with delay and the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, the average waiting time $\bar{F}$ turns out to be the same.

Now, let us characterize the operation of the target 5G infrastructure in a Medium-load mode, which is defined by the following input parameter values: $\bar{T}_\mu = 2$, $\bar{T}_\eta = 1$, $CV_\eta = 0.5$. Then for $u = 0.5$, using expression (23), we obtain $CV_\mu = 0.75$, and according to expressions (21) and (22), we obtain $\mu = \frac{2}{3}$, $\eta = 2$. The average waiting time $\bar{F}$ is determined using expression (18): $\bar{F} = \frac{\left(\frac{1}{3}\right)}{\left(\frac{4}{3}\right)=\frac{1}{4}}$. For the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, under the same input parameter values and considering $u = 0.5$, we obtain $\bar{F} = \frac{\left(\frac{0.5}{1}\right)}{0.5=1}$. Meanwhile, for the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, with an incoming request arrival rate $\mu = \frac{2}{3}$ and a service rate $\eta = 2$, the utilization $u = \frac{1}{3}$ and the average waiting time $\bar{F} = \frac{\left(\frac{1}{\frac{3}{2}}\right)}{\left(\frac{1}{3}\right)} = \frac{1}{4}$ are obtained accordingly.

Finally, we characterize the operation of the target 5G infrastructure in a High-load mode, which is defined by the following input parameter values: $\bar{T}_\mu = \frac{10}{9}$, $\bar{T}_\eta = 1$, $CV_\eta = 0.5$. Then, for $u = 0.9$, using expression (23), we obtain $CV_\mu = 0.55$, and according to expressions (21) and (22), we obtain $\mu = \frac{19}{11}$, $\eta = 2$. The average waiting time $\bar{F}$ is determined using expression (18): $\bar{F} = \frac{\left(\frac{\frac{18}{11}}{2}\right)}{\left(2-\frac{18}{11}\right)=\frac{9}{4}}$. For the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, under the same input parameter values and considering $u = 0.9$, we obtain $\bar{F} = \frac{\left(\frac{0.9}{1}\right)}{(1-0.9)=9}$. Meanwhile, for the $\frac{\mathrm{M}}{\frac{\mathrm{M}}{1}}$-type system, with an incoming request arrival rate $\mu = \frac{18}{11}$ and a service rate $\eta = 2$, the utilization $u = \frac{1}{3}$ and the average waiting time $\bar{F} = \frac{9}{4}$ are obtained accordingly.

Let us explore the potential of the mathematical framework proposed in Section 2 by conducting a series of studies analyzing the above parametrically defined 5G infrastructure. All subsequent studies presented here focus on evaluating the process of managing the utilization distribution within the target 5G infrastructure (hereinafter referred to as the studied process), based on both the queuing system with the delay model introduced in Section 2 (hereinafter referred to as the delay model) and the classical $\frac{M}{\frac{M}{1}}$-type queuing system model (hereinafter referred to as the classical model).

In a real 5G infrastructure, the quality of subscriber service is significantly affected by the uneven flow of incoming requests and the high utilization of base stations. Unlike the classical model, the delay model accounts for these factors when managing the utilization distribution, ensuring connection stability even when the target 5G infrastructure operates under high-load conditions. Let us compare the results of describing the studied process using the delay model and the classical model, focusing on the impact of base station utilization u and the coefficient of variation of incoming request flow CV_μ on the average waiting time $\bar{F}$ (see Fig. 1).

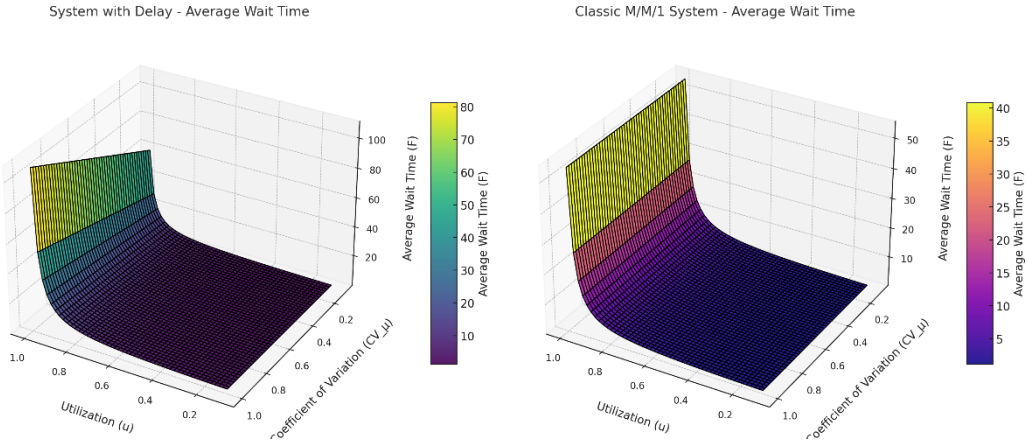

Figure 1: Results of the parameterization of the studied process in the form of the dependency $\bar{F} = f(u, CV_\mu)$.

The results presented in Fig. 1 reveal a significant difference in the representation of the studied process by the delay model and the classical model. The delay model ensures a stable dynamic of the average waiting time $\bar{F}$ even under a high coefficient of variation CV_μ and substantial utilization u. This indicates the model's ability to adapt to the variability of the incoming request flow effectively. Regardless CV_μ, the classical model allows for a rapid increase in average waiting time as utilization approaches the critical level $u \to 1$, which limits its efficiency. Therefore, unlike the classical model, the delay model accounts for the uneven distribution of incoming traffic, reducing service delays for accepted incoming requests, particularly under low or moderate utilization $u < 0.7$.

The analysis of average waiting time conducted in the previous study provided insights into how changes in request structure and utilization are accounted for by the studied process in managing subscriber service quality. However, this alone is insufficient for a comprehensive description of the target 5G infrastructure's operation, as the waiting time for an incoming request in the buffer depends on the efficiency of base station resource utilization. Examining the dependency $u = f(\mu, CV_\mu)$ allows for an assessment of how utilization levels fluctuate with varying intensity and unevenness of incoming request flow, which is crucial for maintaining a balance between efficient communication resource usage and minimizing waiting time. The calculated $u = f(\mu, CV_\mu)$ dependency variants for describing the studied process using the delay model and the classical model are presented in Fig. 2.

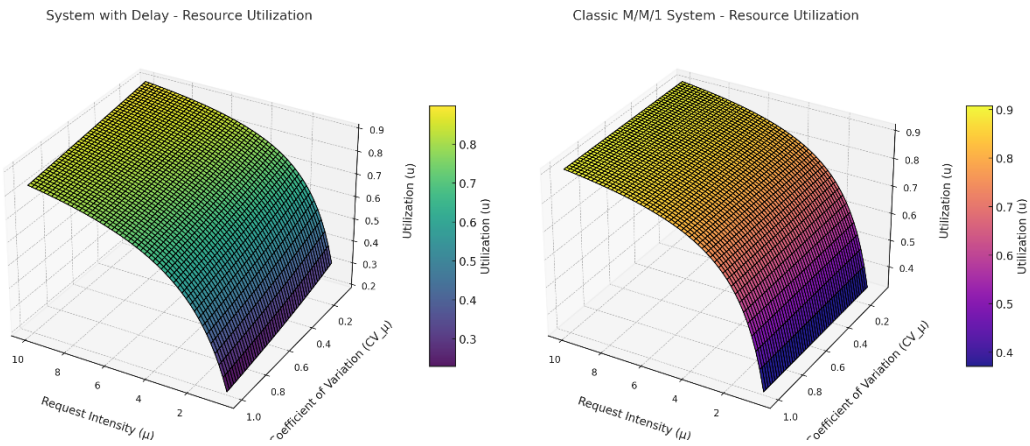

Figure 2: Results of the parameterization of the studied process in the form of the dependency $\boldsymbol{u = f(\mu, CV_\mu)}$.

From Fig. 2, it is evident that the results of communication resource management using the delay model differ significantly from those demonstrated by the classical model. The delay model ensures a smoother adjustment of base station utilization u in response to increasing incoming request intensity μ and coefficient of variation CV_μ, highlighting its ability to adapt to the unevenness of incoming traffic flow. The classical model linearizes the dependency of utilization on μ and exhibits insensitivity u to the impact of the coefficient of variation CV_μ, which reduces the efficiency of communication resource utilization in cases of high traffic flow irregularity. The graph on the left shows that even at high values of CV_μ, utilization increases in a controlled manner. In contrast, the classical model exhibits a significant rise in u only in response to a substantial increase in intensity μ. This confirms the advantage of the delay model, which ensures the stability of communication resource management under complex operating conditions of the 5G infrastructure.

For a real 5G infrastructure, delay is one of the key factors negatively affecting subscriber service quality, especially under high base station utilization. This fact highlights the relevance of studying the impact of delay t_0 and utilization u on the average waiting time $\bar{F}$. The mathematical framework presented in Section 2 provides the necessary functionality to represent the studied process within this basis. The obtained results, including those for the classical model, are presented in Fig. 3.

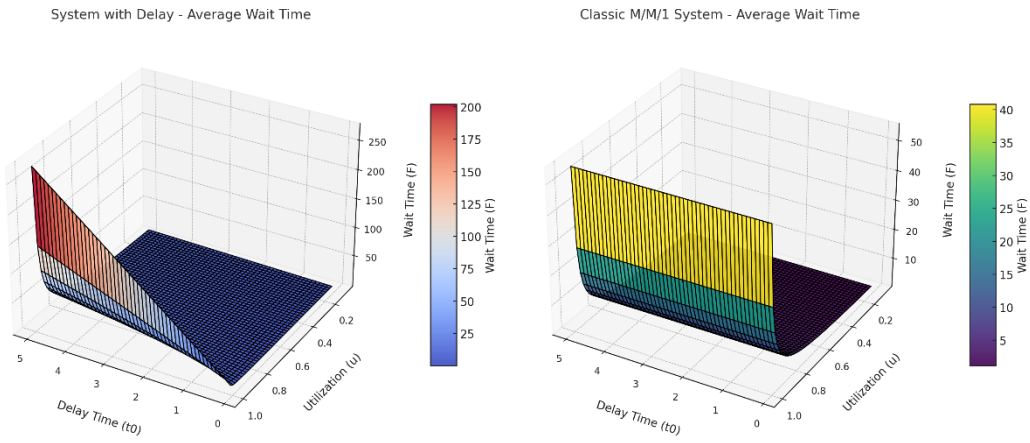

Figure 3: Results of the parameterization of the studied process in the form of the dependency $\boldsymbol{\bar{F} = f(t_0, u)}$.

The results presented in Fig. 3 clearly demonstrate that the delay model and the classical model account for the impact of delay t_0 duration on the average waiting time $\bar{F}$ in different ways. The

graph on the left distinctly shows that the delay model allows for a significant increase in $\bar{F}$ in synchronization with the growth of t_0, especially at high utilization levels: $u \to 1$. This highlights the importance of considering the delay factor when studying the operation of a 5G infrastructure under high-load conditions, where even a slight increase in t_0 can significantly impact subscribers' service quality. The graph on the right illustrates that within the framework of the classical model, $\bar{F}$ it depends only on utilization and increases linearly with u, ignoring the impact of delay duration. This reduces the computational complexity of the analysis but simultaneously limits the applicability of the classical model in real-world conditions where delay is inevitable. Thus, the delay model provides a more accurate representation of the studied process and enables a more precise determination of the moment when the impact of t_0 on $\bar{F}$ may become critical.

4. Conclusions

The article presents a concept for managing the utilization distribution in a high-load 5G infrastructure. This concept is based on the queuing system with a delay model. Analytical expressions have been derived to calculate the characteristics of such a system under variation coefficients of inter-arrival periods of incoming requests $CV_\mu < 1$ and specific additional constraints on the system's input parameters. Based on the obtained results, the following conclusions can be drawn:

1. Considering the delay t_0 in modeling the process of communication resource management in a high-load 5G infrastructure significantly affects the determination of utilization u, which exceeds the corresponding value for the classical $\frac{M}{\frac{M}{1}}$--type system by a factor of $\frac{(1+\eta t_0)}{(1+\mu t_0)}$;
2. The stability of the studied queuing system with delay is largely determined by this value. This is confirmed by the analytical form of expressions (10) and (11), and subsequently, expression (15);
3. When the variation coefficients reach $CV_\mu < 1$, $CV_\eta < 1$, the studied queuing system with delay loses its Markovian properties. In this case, the average waiting time $\bar{F}$ for an accepted incoming subscriber request in the system's buffer becomes lower than the corresponding parameter in the classical $\frac{M}{\frac{M}{1}}$- -type system under identical utilization u values;
4. The use of probability distribution functions (1) and (2) in the proposed queuing system with delay enabled the approximation of input parameter distributions at the level of the first two moments, in contrast to the classical $\frac{M}{\frac{M}{1}}$- -type system;
5. The obtained queuing system with delay can be applied to model a wide range of processes. Furthermore, adapting the system for an adequate description of the target process can be achieved by introducing a primary regulatory element in the form of the $\frac{M}{\frac{M}{1}}$- type system, with parameters defined according to expressions (21) and (22).

It should be noted that in the theoretical part of the article when characterizing the studied queuing system with delay, we focused on the average waiting time of an accepted incoming subscriber request in the buffer. The remaining informative system parameters, such as the average queue length, the average number of accepted subscriber requests, and others, are derived from the parameter $\bar{F}$.

Future research will focus on improving the mathematical framework presented in the article. In particular, a promising direction is the analysis of the impact of dynamic changes in delay and variation coefficients CV_μ, CV_η on service quality under conditions of uneven incoming request flow. Additionally, it is advisable to develop adaptive algorithms that account for variable subscriber mobility and fluctuations in subscriber density during peak load periods. Special attention will be

given to integrating machine learning technologies for predicting communication resource demands and optimizing energy consumption, ensuring the stability of the target infrastructure under complex operational scenarios in 5G networks.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish its results.

Funding

This research is part of the project No. 2022/45/P/ST7/03450 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. A. I. Ahmad et al., Emerging 5G technology: A review of its far-reaching implications for communication and security, World Journal of Advanced Research and Reviews 21 (2024) 2474–2486. doi:10.30574/wjarr.2024.21.1.0346.

[2] K. Obelovska, K. Pelekh, Y. Pelekh, E. Benova, R. Liskevych, An approach towards automate models construction and research of wireless local area networks based on state transition diagram, WSEAS Transactions on Information Science and Applications 20 (2023) 390–397. doi:10.37394/23209.2023.20.41/.

[3] W. Owoko, Exploring the technological advancements and security issues of 5G, World Journal of Advanced Research and Reviews 23 (2024) 812–846. doi:10.30574/wjarr.2024.23.2.2367.

[4] V. Kovtun, O. Kovtun, Service-oriented model for handling mMTC subscribers' traffic in a 5G cluster, in: Proceedings of the 5th International Workshop on Intelligent Information Technologies & Systems of Information Security, CEUR-WS 3675 (2024) 236–246.

[5] R. Girau et al., Definition and implementation of the cloud infrastructure for the integration of the human digital twin in the social internet of things, Computer Networks 251 (2024) 110632. doi:10.1016/j.comnet.2024.110632.

[6] M. Ahmadi, M. Hadi, M. R. Pakravan, Power-efficient joint dynamic resource allocation in virtualized inter-data center elastic optical networks, IEEE Access 12 (2024) 75599–75609. doi:10.1109/access.2024.3406206.

[7] A. Sarah, G. Nencioni, Md. M. I. Khan, Resource allocation in multi-access edge computing for 5G-and-beyond networks, Computer Networks 227 (2023) 109720. doi:10.1016/j.comnet.2023.109720.

[8] A. A. Khan et al., Secure remote sensing data with blockchain distributed ledger technology: A solution for smart cities, IEEE Access 12 (2024) 69383–69396. doi:10.1109/ACCESS.2024.XXXXXXX.

[9] V. Kovtun, O. Kovtun, The concept of efficient utilization of the uplink frequency resource of a smart factory 5G cluster by IIoT devices, in: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume I: Machine Learning Workshop, CEUR-WS 3664 (2024) 273–283.

[10] J. A. J. Alsayaydeh, Irianto, M. Zainon, H. Baskaran, S. G. Herawan, Intelligent interfaces for assisting blind people using object recognition methods, International Journal of Advanced Computer Science and Applications 13 (2022).
[11] D. M. C. K., H. M., D. Selvamuthu, P. Kalita, Energy efficiency in a base station of 5G cellular networks using M/G/1 queue with multiple sleeps and N-policy, Methodology and Computing in Applied Probability 25 (2023). doi:10.1007/s11009-023-10026-1.
[12] A. M. Anwar et al., Handoff scheme for 5G mobile networks based on Markovian queuing model, Journal of Advanced Research in Applied Sciences and Engineering Technology 30 (2023) 348–361. doi:10.37934/araset.30.3.348361.
[13] I. Keramidi et al., Analytical modelling of a vehicular ad hoc network using queueing theory models and the notion of channel availability, AEU - International Journal of Electronics and Communications 170 (2023) 154811. doi:10.1016/j.aeue.2023.154811.
[14] S. Vladov, L. Scislo, N. Szczepanik-Ścisło, A. Sachenko, V. Vysotska, Neural network method of controllers' parametric optimization with variable structure and semi-permanent integration based on the computation of second-order sensitivity functions, Applied Sciences 15 (2025) 2586. doi:10.3390/app15052586.
[15] H. B. Garcia e Silva, R. M. N. Santos, M. Ricardo, Mitigating information asymmetry in 5G networks, Internet Policy Review 13 (2024). doi:10.14763/2024.2.1765.
[16] D. Merit C. K., H. M., Analysis of multiple sleeps and N-policy on a M/G/1/K user request queue in 5G networks base station, The Scientific Temper 14 (2023) 375–382. doi:10.58414/scientifictemper.2023.14.2.21
[17] W. Ding, M. Shikh-Bahaei, HARQ delay minimization of 5G wireless network with imperfect feedback, arXiv (2023). doi:10.48550/ARXIV.2305.02948.
[18] S. Abubakar, A. R. Mohd Shariff, S. I. Fadilah, Numerical approximation of 5G-V2V physical sidelink shared channel (PSSCH) capacity limit using M/G/1 queuing and Newton–Raphson method, Alexandria Engineering Journal 71 (2023) 201–207. doi:10.1016/j.aej.2023.03.045.
[19] L. Mochurad, V. Babii, Y. Boliubash, Y. Mochurad, Improving stroke risk prediction by integrating XGBoost, optimized principal component analysis, and explainable artificial intelligence, BMC Medical Informatics and Decision Making 25 (2025). doi:10.1186/s12911-025-02894-z.
[20] M. Talal et al., A comprehensive systematic review on machine learning application in the 5G-RAN architecture: Issues, challenges, and future directions, Journal of Network and Computer Applications 233 (2025) 104041. doi:10.1016/j.jnca.2024.104041.
[21] X. Li et al., Network slicing for 5G: Challenges and opportunities, IEEE Internet Computing 21 (2017) 20–27. doi:10.1109/mic.2017.3481355.
[22] K. Obelovska, Y. Snaichuk, O. Liskevych, S.-A. Mitoulis, R. Liskevych, Mitigation of risks associated with distrustful routers in OSPF networks—An enhanced method, Computers 14 (2025) 43. doi:10.3390/computers14020043.