SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2024
ORGANIZATION AND MANAGEMENT SERIES NO. 214

WAITING TIME BEHAVIOR IN A SERVICE MODEL WITH A MULTI-TYPE PROCESSING

Wojciech M. KEMPA[1*], Viacheslav KOVTUN[2]

[1] Silesian University of Technology, Faculty of Applied Mathematics, Department of Mathematical Methods in Technics and Informatics, Gliwice, Poland; wojciech.kempa@polsl.pl, ORCID: 0000-0001-9476-2070
[2] Vinnytsia National Technical University, Faculty of Intelligent Information Technologies and Automation, Department of Computer Control Systems, Vinnytsia, Ukraine; kovtun_v_v@vntu.edu.ua, ORCID: 0000-0002-7624-7072
* Correspondence author

Purpose: In many real-life service systems an arriving customer can obtain different-type processing in dependence on its preferences or requirements. In the paper, a model of a service system with finite waiting room and various types of processing is proposed.
Findings: Analytic method based on integral equations and matrix approach is applied to find a representation for the Laplace transform of the customer waiting time distribution conditioned by the number of customers accumulated in the waiting room initially. Numerical examples are attached as well.
Research limitations/implications: The current study focuses on the analysis of a queueing system with a finite waiting room and multi-type service characteristics under specific assumptions, such as Poisson input streams and hyper-exponential service times. The generalizability of the results may be limited to scenarios adhering to these conditions. Additionally, the transient state analysis primarily addresses fixed initial conditions, which could restrict its applicability to dynamically changing systems or environments. Further research could explore extensions incorporating alternative arrival patterns, varying buffer sizes, and adaptive service mechanisms to improve real-world applicability.
Practical implications: The findings of this study offer valuable insights for the design and optimization of service systems with limited waiting room capacities, such as healthcare facilities, production lines, and network routers. By understanding the transient waiting time distribution, system administrators can better predict potential bottlenecks and implement strategies to minimize customer delays and service interruptions. The mathematical framework can guide decision-making regarding buffer sizing, service process allocation, and workload distribution to improve overall service quality and operational efficiency. Additionally, the approach may aid in the development of real-time control mechanisms for systems experiencing fluctuating demand or unexpected surges in traffic.
Social implications: Efficient queue management is crucial in various sectors that directly impact society, such as healthcare, public transportation, and digital services. By minimizing waiting times, the study contributes to enhancing customer satisfaction, reducing frustration, and improving overall well-being. In healthcare, optimized waiting room capacities can lead to faster service delivery, potentially improving patient outcomes. Similarly, in public services and transportation, reducing delays helps to maintain social trust and ensure equitable access.

The study's findings can also support sustainable development by reducing resource wastage and enhancing system resilience in critical infrastructures, benefiting communities at large.
Originality/value: This study presents a novel analytical approach to evaluating the transient waiting time distribution in finite-buffer queueing systems, employing integral equations and a matrix-based solution framework. Unlike many existing studies focused primarily on steady-state analysis, this research addresses the system's transient behavior, offering a more comprehensive understanding of queue dynamics during non-equilibrium conditions. The consideration of multi-type service times modeled by hyper-exponential distributions further enhances the model's applicability to diverse real-world service processes. The findings provide practical tools for optimizing queue performance in various industries, highlighting the study's contribution to both theoretical advancements and practical implementations.
Keywords: from Poisson stream; queue; service station; time-dependent analysis; waiting time.
Category of the paper: Research paper.

1. Introduction

Nowadays, the customer service process is becoming more and more personalized. This is of course related to the conditions of market competition in which companies from the service industry operate. The desire to attract as many customers as possible forces these companies to make the service process much more flexible and adapt it to certain specific requirements and preferences of customers. As a consequence, the servicing of individual customers may vary significantly, in particular, these differences often relate to the duration of the service itself.

Queueing systems are a very convenient mathematical tool that allows for practical modeling of the behavior of many real service stations. Monitoring of phenomena typical for models related to customer service, such as the accumulation of customers waiting for service, the status of the queue of customers waiting for service, the waiting time for a single customer to start its service, is possible by analyzing the stochastic characteristics of the relevant queueing models, which approximate the behavior of the real system. Of particular importance here are queueing systems with limited capacities of the so-called waiting rooms, i.e. buffers accumulating customers waiting for service. Indeed, in practice, the size of the waiting room is limited in most real service systems. This is, for example, in the case of patients waiting for a doctor's appointment, in the case of components awaiting processing at a specific point in the production line, or in the case of network switch buffers (e.g. Internet routers) in IoT traffic.

One of the most essential stochastic characteristics of each queueing model is the probability distribution of the customer waiting time to start its service (queueing delay, virtual waiting time). This characteristic is crucial from the point of view of ensuring the appropriate level of customer service quality (QoS), and affects the cost of system operation, because keeping customers waiting for service in the buffer is often expensive. Moreover, long waiting times can lead to losses of customers due to buffer overflow.

The literature on queueing models and their practical use in modeling real service systems is huge an,d constantly growing. This applies, in particular, to works devoted to the distribution of customer waiting time. Therefore, the following literature review has been prepared taking into account the most important, in the author's opinion, items, research directions and applications in this field published in recent years.

In (Bellomo, Brezzi, 2020) a discussion on new trends and challenges in traffic, crowds, and dynamics of self-organized particles can be found, in which queueing theory results can be essentially used. A new approach for computing waiting time distribution in finite-capacity queueing models with Markov arrivals can be found in Chaudhry et al. (2023). In Kim (2020) a priority queue with Poisson arrivals is analyzed. Waiting time distributions for different class of customers are investigated there. The problem of bus delays is studied in Sun et al. (2015) by using queueing theory and Markov chain approach. The problem of estimating entropy production from the point of view of waiting time distributions is discussed in Skinner, Dunkel (2021). In Lee et al. (2020) queueing delay distribution in the discrete-type multi-server queue with batch arrivals of customers is investigated. The problem of the impact of skewness of interarrival and service times on the waiting time distribution is studied in Romero-Silva et al. (2020). In Walraevens et al. (2022) asymptotics of the waiting time distributions in the accumulating queue with priority is considered. A rarely used LIFO processing discipline with the auto-correlated input stream in MAP/G/1/N-type model is considered in Dudin et al. (2017), where the representation for the stationary queueing delay distribution is found. Waiting time distributions in an M/G/1 retrial queue with two classes of customers and in a correlated model with exponential interarrival and service times are analyzed in Kim, Kim (2017, 2018), respectively. In Baek et al. (2016) explicit-form representations for transient waiting time distribution in the M/D/1 queue can be found. The waiting time distribution in the D-BMAP/G/1 queueing model is investigated in Samanta (2020). In Bratiychuk, Kempa (2003) a new approach for studying transient characteristics, e.g. waiting time distribution, in a general-type batch-arrival queueing model is proposed. The method is based on the factorization technique and integral equations. Stationary analysis of key performance measures in M/G/n-type model with bounded capacity and packet dropping is done in Tikhonenko, Kempa (2016). In Kempa (2010) the compact-form representation for the actual waiting time in the GI/G/1-type model with batch arrivals is obtained in the transient state of the system operation. A model of a wireless sensor network node operation with a modified threshold-type energy saving mechanism is proposed in Kempa (2019). A proposal of a weight queue active queue management which is based on dynamic monitoring of the current queue size can be found in (Baklizi, 2020) as a tool for reduction congestions at router buffers. In Xie et al. (2023) the problem of adapting of queueing systems to changing model conditions, such as e.g. fluctuations in the number of devices or message sizes, is discussed in the context of using IoT edge computing. The so-called message queues being a way of asynchronous communication for software components or applications by using a shared buffer are

investigated in Maharjan et al. (2023). The mechanism of Active Queue Management and its impact on queueing characteristics is considered in Marek et al. (2022). In Solaiappan et al. (2023) an interesting proposal of using of signal distribution control algorithm in minimizing the vehicle queue waiting time can be found, which fits the hot topic of smart cities. In Dimakou et al. (2015) the problem of the estimation of the waiting time distribution in public health care is discussed. Queueing delay distribution for dynamic pickup and delivery problems is analyzed in Vonolfen, Affenzeller (2016). In Bounkhel et al. (2020) a model with server breakdowns is investigated. In Arita, Schadschneider (2015) an interesting queueing model in a microscopic level is considered.

Most of the results obtained for stochastic characteristics of queueing models concern the stationary (steady) state of the system (the case in which time parameter t tends to infinity). In practice, however, the steady state does not always describe the functioning of the system well. For example, each time the service station fails, the system has to stabilize again, and the indicator of the system operation is then the transient (non-stationary) state. The same applies to the analysis of the system operation just after its start-up or just after changing the traffic control mechanism, or the size of the accumulating buffer.

The article analyzes the non-stationary (transient, at a fixed moment t) probability distribution of the customer waiting time in a queueing model with a Poisson input stream of customers and limited waiting room capacity. The traditional, classic FIFO service discipline is applied, according to which customers are served in the order in which they appear in the system. Customers are offered various "conditions" of service, hence the service time for a single customer is modeled using a hyper-exponential distribution with fixed parameters. In Section 2, the considered queueing model is described in detail mathematically. In Section 3, we construct the system of integral equations governing the transient waiting time distribution conditioned by the initial state of the system, i.e. the number of customers waiting for service at the opening ($t = 0$). In Section 4, we write a system of linear equations for Laplace transforms corresponding to the original one, represent it in a matrix form and state the formula for its general solution. Some supplementary results can be found in Section 5. Section 6 contains results of numerical experiments illustrating sensitivity of the transient waiting time distribution on key predefined model parameters.

2. Mathematical model

In the paper, we study a single-channel queueing model with a multi-type service process. Customers arrive into the service station according to a Poisson process with given rate a. The system is equipped with a finite-capacity buffer (waiting time) for accumulating entering customers which must wait for start the service process. The buffer capacity (volume) equals

N-1 places that is a non-random (fixed) value, so the maximum number of customers which are allowed to be present in the system simultaneously equals N (the buffer capacity plus one place for the customer being processed). In the case an arriving customer finds the waiting time being full, it is being lost (it leaves the system immediately without service). The processing is organized according to the natural FIFO (First-In-First-Out) service discipline.

In dependence on their preferences or requirements the entering customers may obtain different-type processing. So, we assume that the service time of a single customer is hyper-exponentially distributed with parameters

$$(b_1, p_1), \ldots, (b_k, p_k), \tag{1}$$

where $b_i > 0, p_i \geq 0$ for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} p_i = 1$.

Thus, the service time of an arriving customer has exponential distribution with mean b_i^{-1} with probability b_i^{-1}, where $i = 1, \ldots, k$, i.e. a customer can obtain k different types of service in the considered model, where k is fixed.

In consequence, the CDF (cumulative distribution function) and the PDF (probability density function) of single customer service time are defined, respectively, as follows:

$$B(t) \stackrel{def}{=} \sum_{i=1}^{k} p_i (1 - e^{-b_i t}) \tag{2}$$

and

$$b(t) \stackrel{def}{=} \sum_{i=1}^{k} p_i b_i e^{-b_i t}, \tag{3}$$

where $t > 0$.

One of the key characteristics of any queueing system is the so-called queueing delay defined for any fixed time t. The queueing delay (also known as the virtual waiting time for service) at fixed time t expresses the time that a customer appearing in the system at exactly time t would have to wait for the start of the service. Obviously, the moment t need not be a real customer arrival moment, hence the term "virtual".

Denoting by $\tau(t)$the virtual waiting time (queueing delay) at time t, let us introduce the following notation:

$$T_n(t, x) \stackrel{def}{=} \mathbf{P}\{\tau(t) > x \mid \text{initialbufferstate} = n\}, \tag{4}$$

where $t > 0$, $x > 0$ and $n \in \{0, \ldots, N\}$.

Indeed, $T_n(t, x)$ stands for the probability that the waiting time of a customer arriving at time t exceeds x on condition that the system initially (at time $t = 0$) contains n customers accumulated in the buffer (waiting room) exactly. Evidently, for fixed t the probability $T_n(t, x)$is dependent on n essentially.

3. Time-dependent equations

In this section, we establish a system of Volterra-type integral equations for conditional virtual delay distribution defined in (4) utilizing Markov moments in the evolution of the considered queueing system. Indeed, due to memoryless property of interarrival times, consecutive departure moments (time epochs at which customers complete their processing and leave the system) are renewal (Markov) moments in the evolution of the system.

Let us begin with the case of the system being empty at the opening (at time $t = 0$). Denoting by y the first arrival moment after the starting of the system, we obtain the following equation:

$$T_0(t,x) = a\int_0^t e^{-ay}T_1(t-y,x)dy. \tag{5}$$

Indeed, if the first arrival moment y precedes t then, obviously, the probability that the waiting time of a "virtual" customer entering exactly at time t exceeds x is equal to the analogous probability but for the system beginning its operation with one customer present and calculated at time t-y. If the first customer arrives after time t then the probability that the waiting time at time t exceeds x equals 0 (at time t the system is still empty).

Similarly, if the accumulating buffer contains at least one customer at the opening epoch, denoting by y the first service completion epoch (that is a renewal moment in the system evolution) and applying the formula of total probability, we get

$$\begin{aligned} T_n(t,x) &= \textstyle\sum_{j=0}^{N-n-1}\int_0^t T_{n+j-1}(t-y,x)\frac{(ay)^j}{j!}e^{-ay}\sum_{i=1}^k p_i b_i e^{-b_i y}dy \\ &+ \sum_{j=N-n}^{\infty}\int_0^t T_{N-1}(t-y,x)\frac{(ay)^j}{j!}e^{-ay}\sum_{i=1}^k p_i b_i e^{-b_i y}dy + \theta_n(t,x), \end{aligned} \tag{6}$$

where

$$\theta_n(t,x) = \sum_{k=0}^{N-n-1}\frac{(at)^k}{k!}e^{-at}\int_t^{\infty}\overline{B}^{(n+j-1)\star}(x-y+t)\sum_{i=1}^k p_i b_i e^{-b_i y}dy. \tag{7}$$

In the formula (7) we use the notation $\overline{B}^{i\star}(u) = 1 - B^{i\star}(u)$, where $B^{i\star}(\cdot)$ stands for the i-fold Laplace-Stieltjes convolution of the CDF $B(\cdot)$ with itself which is defined as follows:

$$B^{0\star}(t) = 1, B^{1\star}(t) = B(t), B^{j\star}(t) = \int_0^t B^{(j-1)\star}(t-u)dB(u), \tag{8}$$

where $j \geq 2$.

The first summand on the right side of (6) refers to the situation in that the first customer leaves the system at time $y < t$ and, simultaneously, just before the moment (y) there is at least one free place in the accumulation buffer. The second summand describes a similar situation

with the difference that just before y the buffer is completely saturated. The last summand on the right side of (6) relates to the case in that the first customer departs at time $y > t$. In this situation, if the number of arrivals before t is equal $j \leq N - n - 1$, then the probability that the waiting time of a customer arriving to the system at time t is greater than x is equal to the probability that the total service time of $n + j - 1$ customers exceeds $x - y + t$ (the component $\overline{B}^{(n+j-1)\star}(x - y + t)$. If the buffer is saturated at time t, the "virtual" customer entering at this moment is lost and hence we assume that its waiting time equals 0.

4. Linear system of equations for Laplace transforms and its matrix-form solution

In this section, we establish a system of linear equations corresponding to (5)-(6) written for Laplace transforms of the conditional waiting time distribution. Next we transform this system into a matrix form and obtain the representation for the solution.

So, introduce the following notation:

$$\hat{T}_n(s,x) \overset{def}{=} \int_0^\infty e^{-st} T_n(t,x)dt \tag{9}$$

where $\mathrm{Re}(s) > 0$.

Due to the fact that

$$\begin{aligned} &a\int_{t=0}^{\infty} e^{-st}dt \int_{y=0}^{t} e^{-ay} T_1(t-y,x)dy \\ &= a\int_{y=0}^{\infty} e^{-(a+s)y}dy \int_{t=y}^{\infty} e^{-s(t-y)} T_1(t-y,x)dt = \frac{a}{a+s}\hat{T}_1(s,x), \end{aligned} \tag{10}$$

we obtain from (5) the following equation:

$$\hat{T}_0(s,x) = \frac{a}{a+s}\hat{T}_1(s,x). \tag{11}$$

Let us observe that, changing the order of integration, the following representation is true (compare to the right side of (6)):

$$\begin{aligned} &\int_{t=0}^{\infty} e^{-st}dt \int_{y=0}^{t} T_r(t-y,x)\frac{(ay)^j}{j!} e^{-ay} p_i b_i e^{-b_i y}dy \\ &= p_i b_i \int_{y=0}^{\infty} e^{-(a+b_i+s)y}\frac{(ay)^j}{j!}dy \int_{t=y}^{\infty} e^{-s(t-y)} T_r(t-y,x)dt \\ &= \frac{p_i b_i}{a+b_i+s}\left(\frac{a}{a+b_i+s}\right)^j \hat{T}_r(s,x). \end{aligned} \tag{12}$$

Denoting

$$\beta_j(s) \overset{def}{=} \sum_{i=1}^{k} \frac{p_i b_i}{a+b_i+s}\left(\frac{a}{a+b_i+s}\right)^j, \tag{13}$$

and (see (7))

$$\hat{\theta}_n(s,x) \overset{def}{=} \int_0^\infty e^{-st}\theta_n(t,x)dt, \tag{14}$$

we can rewrite equations (6) in terms of Laplace transforms as follows:

$$\hat{T}_n(s,x) = \sum_{j=0}^{N-n-1} \beta_j(s)\hat{T}_{n+j-1}(s,x) + \hat{T}_{N-1}(s,x)\sum_{j=N-n}^{\infty} \beta_j(s) + \hat{\theta}_n(s,x), \tag{15}$$

where $n \in \{1,\ldots,N\}$.

Now let us transform the system of linear equations (11) and (15) into a matrix form by defining appropriate functional matrices.

Let start with introducing a functional square matrix $\mathbf{B}(s) = \left(b_{i,j}(s)\right)$ of size $(N+1)\times(N+1)$ of coefficients of the system (11) and (15).

Successive entries of the first row of this matrix we define as follows:

$$b_{1,j}(s) \overset{def}{=} \begin{cases} 1, & \text{for } j=1, \\ -\dfrac{a}{a+s}, & \text{for } j=2, \\ 0, & \text{for } j \in \{3,\ldots,N+1\}. \end{cases} \tag{16}$$

Next, for $i = 2,\ldots,N-1$let us denote

$$b_{i,j}(s) \overset{def}{=} \begin{cases} -\beta_0(s), & \text{for } j=i-1, \\ 1-\beta_1(s), & \text{for } j=i, \\ \beta_{j-i+1}(s), & \text{for } j \in \{i+1,\ldots,N-1\}, \\ -\sum\limits_{k=N-i+1}^{\infty} \beta_k(s), & \text{for } j=N, \\ 0, & \text{for } j=N+1. \end{cases} \tag{17}$$

The penultimate row of the matrix $\mathbf{B}(s)$ has the following entries:

$$b_{i,j}(s) \overset{def}{=} \begin{cases} -\beta_0(s), & \text{for } j=i-1, \\ 1-\beta_1(s), & \text{for } j=i, \\ \beta_{j-i+1}(s), & \text{for } j \in \{i+1,\ldots,N-1\}, \\ -\sum\limits_{k=N-i+1}^{\infty} \beta_k(s), & \text{for } j=N, \\ 0, & \text{for } j=N+1. \end{cases} \tag{18}$$

Finally, let us define entries of the last row as follows:

$$b_{N+1,j}(s) \overset{def}{=} \begin{cases} 0, & \text{for } j \in \{1,\dots,N-1\}, \\ -\sum_{k=0}^{\infty} \beta_k(s), & \text{for } j = N, \\ 1, & \text{for } j = N+1. \end{cases} \tag{19}$$

In consequence, the functional matrix of coefficients has the following shape:

$$\begin{bmatrix} 1 & -\frac{a}{a+s} & 0 & 0 & \dots & 0 & 0 & 0 \\ -\beta_0(s) & 1-\beta_1(s) & -\beta_2(s) & -\beta_3(s) & \dots & -\beta_{N-2}(s) & -\sum_{k=N-1}^{\infty} \beta_k(s) & 0 \\ 0 & -\beta_0(s) & 1-\beta_1(s) & -\beta_2(s) & \dots & -\beta_{N-3}(s) & -\sum_{k=N-2}^{\infty} \beta_k(s) & 0 \\ 0 & 0 & -\beta_0(s) & 1-\beta_1(s) & \dots & -\beta_{N-4}(s) & -\sum_{k=N-3}^{\infty} \beta_k(s) & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -\beta_0(s) & 1-\sum_{k=1}^{\infty} \beta_k(s) & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -\sum_{k=0}^{\infty} \beta_k(s) & 1 \end{bmatrix}. \tag{20}$$

Let us define one-column matrices of unknown functions and free terms with $N+1$ rows in the following way:

$$\mathbf{T}(s,x) \overset{def}{=} \left[\hat{T}_0(s,x),\dots,\hat{T}_N(s,x)\right]^T \tag{21}$$

and

$$\mathbf{K}(s,x) \overset{def}{=} \left[\hat{\theta}_0(s,x),\dots,\hat{\theta}_N(s,x)\right]^T, \tag{22}$$

respectively, where we assume additionally that $\hat{\theta}_0(s,x) = 0$ (compare the right side of the equation (11)).

Referring to (20)-(22), the linear system of equations (11), (15) can be written as follows:

$$\mathbf{B}(s)\mathbf{T}(s,x) = \mathbf{K}(s,x). \tag{23}$$

The representation for the solution of the system (23) can be given in the following matrix form:

$$\mathbf{T}(s,x) = \mathbf{B}(s)^{-1}\mathbf{K}(s,x). \tag{24}$$

5. Supplementary results

In this section, we present some supplementary results related to the considered model, namely the explicit form formula for the Laplace transform of the waiting time distribution in the case of the system without waiting room, the representation for the stationary waiting time distribution and its mean value.

5.1. The model without waiting room

A special case of the considered queueing model is that without waiting time in which the arriving customers can wait for starting their processing, i.e. $N = 1$. Then we get the following simplified formulae for key functional matrices:

$$\mathbf{B}(s) = \begin{bmatrix} 1 & -\dfrac{a}{a+s} \\ -\displaystyle\sum_{k=0}^{\infty} \beta_k(s) & 1 \end{bmatrix} \tag{25}$$

and

$$\mathbf{K}(s,x) \overset{def}{=} \left[\hat{\theta}_0(s,x), \hat{\theta}_1(s,x)\right]^T = [0,0]^T. \tag{26}$$

Because

$$|\mathbf{B}(s)| = 1 - \frac{a}{a+s}\sum_{k=0}^{\infty} \beta_k(s) \neq 0, \tag{27}$$

the only solution of the system (11), (15) is the zero solution. Indeed, if an arriving customer find the server busy with processing it is lost, so its waiting time equals 0. Similarly, an arriving customer that finds the system empty is being processed without waiting, hence its waiting time is 0, too.

5.2. Stationary waiting time distribution

Obviously, since the considered queueing model has finite buffer capacity, the stationary waiting time distribution exists and, moreover, it is independent on the initial buffer state, i.e. the number of customers accumulated in the waiting room at the starting epoch. The formula for the stationary waiting time distribution can then be found by using the Tauberian theorem.

Denoting

$$T(x) \overset{def}{=} \lim_{t\to\infty} T_n(t,x) = \lim_{t\to\infty} \mathbf{P}\{\tau(t) > x \mid \text{initial buffer state} = n\}, \tag{28}$$

we obtain

$$T(x) = \lim_{s \downarrow 0} s \cdot \hat{T}_n(s,x) = \lim_{s \downarrow 0} s \int_0^\infty e^{-st} \mathbf{P}\{\tau(t) > x \mid \text{initial buffer state} = n\} dt, \qquad (29)$$

for arbitrary $x \geq 0$.

5.3. Mean waiting time in equilibrium

Denoting by τ the waiting time in the stationary state (equilibrium), its mean value can be calculated as follows:

$$\mathbf{E}(\tau) = \int_0^\infty \mathbf{P}\{\tau > x\} dx = \int_0^\infty T(x) dx, \qquad (30)$$

where the formula for $T(x)$ is given in (29).

6. Numerical results

In this section, we present the results of numerical experiments in which the impact of the input parameters of the considered system on the distribution of the waiting time for service, such as the intensity of customer arrivals, parameters determining the distribution of the service time of a single customer, as well as the initial state of the system. Four different scenarios were considered, in which $N = 2$ and the assumption that the service station offers three different types of service described with exponential distributions with different probabilities, so values

$$(b_1, p_1), (b_2, p_2), (b_3, p_3)$$

are predefined.

The graphs presented below show the behavior of the function

$$T_n(t,x) = \mathbf{P}\{\tau(t) > x \mid \text{initialbufferstate} = n\}$$

in dependence on time parameter t for selected values of the argument x and different initial states n of the accumulative buffer at the opening of the system. Moreover, appropriate stationary waiting time probabilities are found in each case.

6.1. Scenario 1

In Scenario 1, we take into consideration the model in which $a = 2$ and the hyper-exponential service time distribution is defined by the following parameters:

$$b_1 = 1, p_1 = 0.5, b_2 = 2, p_2 = 0.3, b_3 = 3, p_3 = 0.2,$$

So 50% of customers are offered service with an average duration 1, for 30% of customers the mean service duration equals 0.5, and for the remaining 20% it is equal to 0.3 time unit.

It is easy to check that the offered load ρ for such a model equals

$$\rho \overset{def}{=} (\text{Arrival rate}) \times (\text{Mean service time}) = 1.43,$$

so the system is overloaded.

In Figures 1 and 2 the visualization of probabilities

$$T_n(t,1) = \mathbf{P}\{\tau(t) > 1 \mid \text{initialbufferstate} = n\}$$

and

$$T_n(t,0.2) = \mathbf{P}\{\tau(t) > 0.2 \mid \text{initialbufferstate} = n\}$$

are presented, respectively, for $n = 0$ (solid line), $n = 1$ (dashed line) and $n = 2$ (dotted line). The same convention is adopted for all other figures.

Stationary probabilities are the following:

$$\mathbf{P}\{\tau > 1\} = 0.078, \mathbf{P}\{\tau > 0.2\} = 0.223.$$

Let us note that around these values, the curves in each of the graphs stabilize (which illustrates that stationary probabilities do not depend on the initial buffer state). As it can be easily noted, the transient waiting time distribution depends essentially on the initial buffer state n. This dependence is especially noticeable for small values of t.

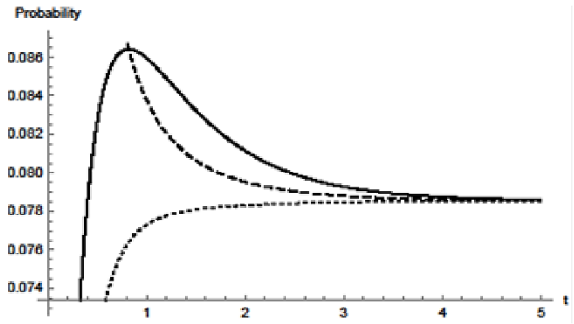

Figure 1. Visualization of probabilities $T_n(t,1)$ for Scenario 1 and $n = 0, 1, 2$.

Source: Authors' own.

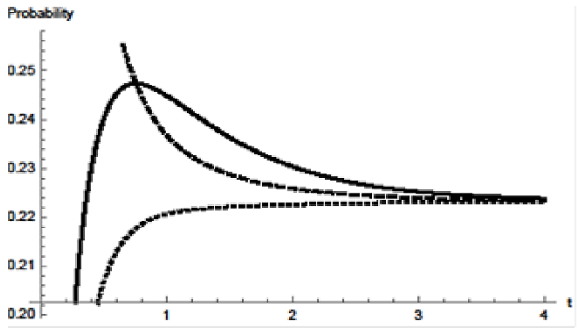

Figure 2. Visualization of probabilities $T_n(t,0.2)$ for Scenario 1 and $n = 0, 1, 2$.

Source: Authors' own.

6.2. Scenario 2

In this scenario, we take into consideration the same probability distribution of the service time as in Scenario 1 but we take smaller arrival intensity, namely $a = 1$. In consequence we have $\rho = 0.717$.

Similarly to Scenario 1, we present in Figures 3 and 4 transient probabilities $T_n(t, 1)$ and $T_n(t, 0.2)$ respectively, for $n = 0, 1$ and 2.

Appropriate probabilities in the equilibrium of the system are the following ones:

$$\mathbf{P}\{\tau > 1\} = 0.082, \mathbf{P}\{\tau > 0.2\} = 0.227.$$

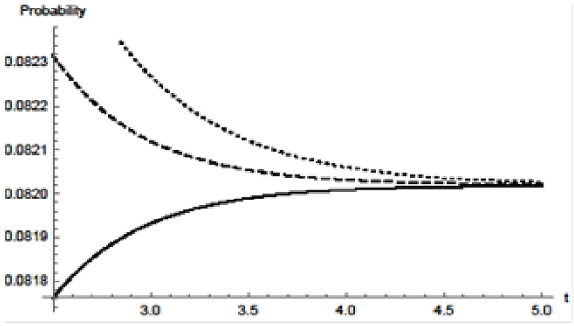

Figure 3. Visualization of probabilities $T_n(t, 1)$ for Scenario 2 and $n = 0, 1, 2$.
Source: Authors' own.

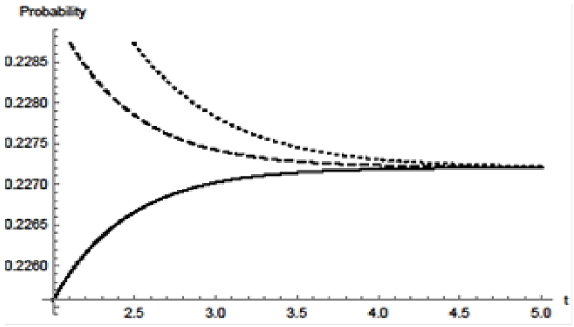

Figure 4. Visualization of probabilities $T_n(t, 0.2)$ for Scenario 2 and $n = 0, 1, 2$.
Source: Authors' own.

Let us note that in the case of a lower offered load (as in Figures 3-4), transient distributions converge slower to the stationary ones. For example, for $t = 4$, the values of probabilities for different values of n differ markedly, while in the case of a greater offered load (Figures 1-2) they are almost imperceptible.

6.3. Scenario 3

In this scenario, the service time distribution is defined by the following parameters:

$$b_1 = 1, p_1 = 0.2, b_2 = 2, p_2 = 0.3, b_3 = 3, p_3 = 0.5,$$

So 50% of customers are offered service with the smallest average duration 0.3, for 30% of customers the mean service duration equals 0.5, and for the remaining 20% it equals 1 time unit.

For such a model we take the arrival rate $a = 3$ and hence the offered load $\rho = 1.550$, so the system is overloaded and in Scenario 1.

Probabilities in the steady state of the system are following:

$$\mathbf{P}\{\tau > 1\} = 0.047, \mathbf{P}\{\tau > 0.2\} = 0.194.$$

Visualizations of transient probabilities $T_n(t, 1)$ and $T_n(t, 0.2)$ are shown in Figures 5 and 6, respectively.

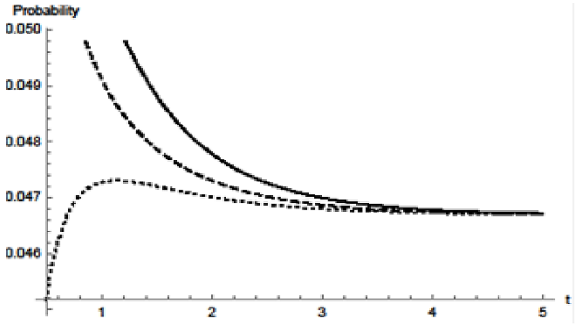

Figure 5. Visualization of probabilities $T_n(t, 1)$ for Scenario 3 and $n = 0, 1, 2$.

Source: Authors' own.

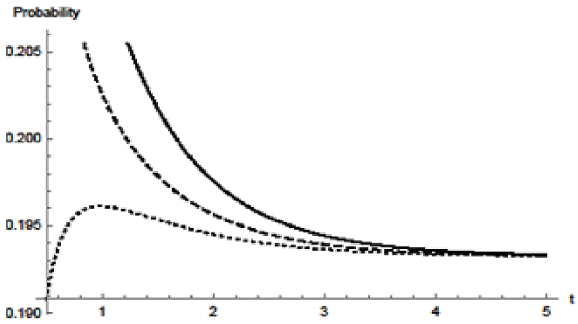

Figure 6. Visualization of probabilities $T_n(t, 0.2)$ for Scenario 3 and $n = 0, 1, 2$.

Source: Authors' own.

6.4. Scenario 4

In the last scenario we take the same service time distribution as defined for Scenario 3 but we take smaller intensity of customer arrivals, namely $a = 2$. In this case we then have $\rho = 1.033$, so the offered load is smaller as in Scenario 3 essentially.

For Scenario 4 we obtain

$$\mathbf{P}\{\tau > 1\} = 0.051, \mathbf{P}\{\tau > 0.2\} = 0.206.$$

Transient probabilities $T_n(t, 1)$ and $T_n(t, 0.2)$ are presented in Figures 7 and 8, respectively.

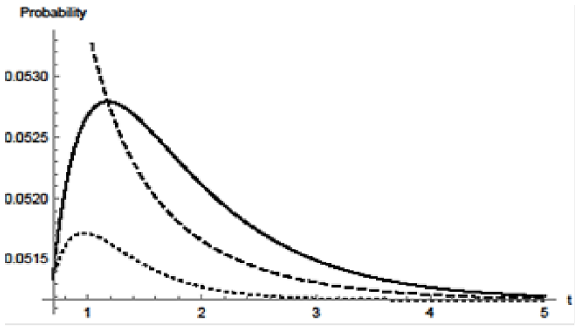

Figure 7. Visualization of probabilities $T_n(t,1)$ for Scenario 4 and $n = 0, 1, 2$.
Source: Authors' own.

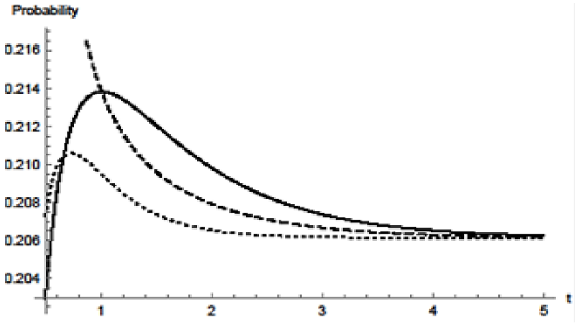

Figure 8. Visualization of probabilities $T_n(t,0.2)$ for Scenario 4 and $n = 0, 1, 2$.
Source: Authors' own.

Let us note that the service time distribution defined for Scenarios 3-4 is "opposite" to that defined for Scenarios 1-2: firstly the smallest mean service time is the rarest one, in Scenarios 3-4 it is the most frequent. In consequence, "roles" of successive curves representing transient probabilities change too: the behavior of curves obtained for $n = 0$ in Scenarios 1-2 is similar to the behavior of curves obtained for $n = 2$ in Scenarios 3-4, and vice versa.

References

1. Arita, Ch., Schadschneider, A. (2015). Exclusive queueing processes and their application to traffic systems. *Mathematical Models and Methods in Applied Sciences, 25(3),* 401-422.
2. Baek, J.W., Lee, H.W., Ahn, S., Bae, Y.H. (2016). Exact time-dependent solutions for the M/D/1 queue. *Operations Research Letters, 44(5)*, 692-695.
3. Baklizi, M. (2020). Weight queue dynamic active queue management algorithm. *Symmetry 12(12),* 2077.
4. Bellomo, M., Brezzi, F. (2015). Traffic, crowds, and dynamics of self-organized particles: New trends and challenges. *Mathematical Models and Methods in Applied Sciences, 25(3),* 395-400.
5. Bounkhel, M., Tadj, L., Hedjar, R. (2020). Entropy analysis of a flexible Markovian queue with server breakdowns. *Entropy, 22(9),* 979.

6. Bratiychuk, M.S., Kempa, W.M. (2003). Application of the superposition of renewal processes to the study of batch arrival queues. *Queueing Systems, 44(1),* 51-67.
7. Chaudhry, M., Banik, A.D., Barik, S., Goswami, V. (2023). A novel computational procedure for the waiting-time distribution (in the queue) for bulk-service finite-buffer queues with Poisson input. *Mathematics, 11(5),* 1142.
8. Dimakou, S., Dimakou, O., Basso, H.S. (2015). Waiting time distribution in public health care: empirics and theory. *Health Economics Review, 5(1),* 25.
9. Dudin, A., Klimenok, V., Samouylov, K. (2017). Stationary distribution of waiting time in MAP/G/1/N queueing system with LIFO service discipline. *Lecture Notes in Computer Science, 10372,* 50-61.
10. Kempa, W.M. (2010). Some results for the actual waiting time in batch arrival queueing systems. *Stochastic Models, 26(3),* 335-356.
11. Kempa, W.M. (2019). Analytical model of a wireless sensor network (WSN) node operation with a modified threshold-type energy saving mechanism. *Sensors, 19(14),* 3114.
12. Kim, B., Kim, J. (2017). Waiting time distributions in an M/G/1 retrial queue with two classes of customers. *Annals of Operations Research, 252(1),* 121-134.
13. Kim, B., Kim, J. (2018). The waiting time distribution for a correlated queue with exponential interarrival and service times. *Operations Research Letters, 46(2),* 268-271.
14. Kim, K. (2020). Delay cycle analysis of finite-buffer M/G/1 queues and its application to the analysis of M/G/1 priority queues with finite and infinite buffers. *Performance Evaluation, 143*, 102133.
15. Lee, S., Kim, B., Kang, J. (2020). Delay analysis in the discrete-time multi-server queue with batch arrivals of packets having deterministic length. *Electronics Letters, 56(23),* 1250-1253.
16. Maharjan, R., Chy, M.S.H., Arju, M.A., Cerny, T. (2023). Benchmarking message queues. *Telecom, 4(2),* 298-312.
17. Marek, D., Szyguła, J., Domański, A., Domańska, J., Filus, K., Szczygieł, M. (2022). Adaptive Hurst-sensitive Active Queue Management. *Entropy, 24(3),* 418.
18. Romero-Silva, R., Shaaban, S., Marsillac, E., Hurtado-Hernández, M. (2020). Studying the effects of skewness of inter-arrival and service times on the probability distribution of waiting times. *Pesquisa Operacional, 40*, 1-29.
19. Samanta, S.K. (2020). Waiting-time analysis of D-BMAP/G/1 queueing system. *Annals of Operations Research, 284(1),* 401-413.
20. Skinner, D., Dunkel, J. (2021). Estimating entropy production from waiting time distributions. *Physical Review Letters, 127(19),* 198101.
21. Solaiappan, S., Kumar, B.R., Anbazhagan, N., Song, Y., Joshi, G.P., Cho, W. (2023). Vehicular traffic flow analysis and minimize the vehicle queue waiting time using signal distribution control algorithm. *Sensors, 23(15),* 6819.

22. Sun, F., Sun, L., Sun, S.-W., Wang, D.-H., Shen, Y. (2015). Study on the calculation models of bus delay at bays using queueing theory and Markov chain. *Computational Intelligence and Neuroscience*, 750304-9.
23. Tikhonenko, O., Kempa, W.M. (2016). Performance evaluation of an M/G/n-type queue with bounded capacity and packet dropping. *International Journal of Applied Mathematics and Computer Science, 26(4),* 841-854.
24. Vonolfen, S., Affenzeller, M. (2016). Distribution of waiting time for dynamic pickup and delivery problems. *Annals of Operations Research, 236(2),* 359-382.
25. Walraevens, J., Van Giel, T., De Vuyst, S., Wittevrongel, S. (2022). Asymptotics of waiting time distributions in the accumulating priority queue. *Queueing Systems, 101(3-4),* 221-244.
26. Xie, Z., Ji, Ch., Xu, L., Xia, M., Cao, H. (2023). Towards an optimized distributed message queue system for AIoT edge computing: A reinforcement learning approach. *Sensors, 23(12),* 5447.