Article

Optimal Hysteresis Control via a Queuing System with Two Heterogeneous Energy-Consuming Servers

Ciro D’Apice 1, Maria Pia D’Arienzo 1, Alexander Dudin 2 and Rosanna Manzo 3,*

1 Dipartimento di Scienze Aziendali—Management and Innovation Systems, University of Salerno, 132, Via Giovanni Paolo II, 84084 Fisciano, SA, Italy; cdapice@unisa.it (C.D.); mdarienzo@unisa.it (M.P.D.)
2 Department of Applied Mathematics and Informatics, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudin@bsu.by
3 Dipartimento di Scienze Politiche e della Comunicazione, University of Salerno, 132, Via Giovanni Paolo II, 84084 Fisciano, SA, Italy
* Correspondence: rmanzo@unisa.it

Abstract: A queuing system having two different servers is under study. Demands enter the system according to a Markov arrival process. Service times have phase-type distribution. Service of demands is possible only if the fixed number of energy units, probably different for two servers, is available in the system at the potential service beginning moment. Energy units arrive in the system also according to a Markov arrival process and are stored in a stock (battery) of a finite capacity. Leakage of energy units from the stock can occur. Demands waiting in the infinite buffer are impatient and can leave the buffer after an exponentially distributed waiting time. One server is the main one and permanently provides service when the buffer is not empty and the required number of energy units is available. The second server is the assistant server and is switched on or off depending on the availability of energy units and queue length according to the hysteresis strategy defined by two thresholds. The assistant server is switched on when the queue length is not less than the greater threshold and is switched off when the queue length becomes smaller than the smaller threshold. The use of the assistant server has to be paid. Thus, the problem of the optimal selection of the thresholds defining the control strategy naturally arises. To solve this problem, the study of the behavior of the system under any fixed values of the parameters of the control strategy is necessary. Such a study is given in this paper. Numerical results are presented. They illustrate the feasibility of computer realization of the developed algorithms for computation of the stationary distribution of the system states and the main key performance indicators as well as the result of solving one of the possible optimization tasks.

Keywords: hysteresis control; heterogeneous servers; Markov arrival process; phase-type distribution; energy harvesting

MSC: 60K25; 60K30; 68M20; 90B22

1. Introduction

This paper contains the analysis of a queuing system that may be potentially useful for the optimal design of a variety of real-world systems. In brief, besides the consideration of arrival processes, which are essentially more general than the stationary Poisson, and service times having become more complicated than those popular in the literature’s exponential distribution, the distinguishing features of the considered system are (i) the necessity of the use of energy units (e.u. for brevity) for the service of demands; (ii) the generation, consumption, and leakage of e.u. during the system operation with their temporal accumulation stored in the battery (stock) with limited capacity; (iii) the availability of two heterogeneous service devices having different parameters of service time distribution and different consumption of e.u.; (iv) the opportunity of the dynamic activation...
and deactivation of the server, which consumes more energy according to the hysteresis strategy, depending on the current queue length and amount of available energy; (v) the account of impatience pertaining to waiting demands. The simultaneous presence of all the listed features guarantees the novelty of the analyzed system. A very brief overview of the literature about the systems characterized by some of the above features is as follows.

1.1. Short Literature Overview

1.1.1. Single-Server Queues with Variable Service Rates

The simplest kind of management of the service or arrival rate and the number of active service devices in queuing models is the threshold-type strategy. It is defined by a number called a threshold. Parameters of the system operation are changed when the number of demands in the system crosses up or down this threshold. It is considered quite a general single-server system with the batch Markov Arrival Process (MAP), the general service time distribution, and the multi-threshold control by the service time all being considered in [1]. A disadvantage of the threshold strategy of control is the possibility of many changes in the service rate (oscillation) when the number of demands in the system admits values around the threshold. This may be undesirable in situations when the change in the rate is charged or is implemented non-instantaneously. The hysteresis-type control suggests that a threshold is replaced with two thresholds. The use of the higher service rate begins when the queue length increases up to the larger threshold. The use of the slower service rate begins when the queue length drops below the smaller threshold. When the length of the queue remains between two thresholds, no change in the current service rate is performed. Such a type of control gives a possibility to reduce the number of rate switching compared to the threshold-type control. The hysteresis-type control using the queuing system operation was introduced in 1960 and is often applied to different queuing systems. As early works in this direction, the following papers can be mentioned: [2–4]. More references can be found, e.g., in [5–8] and the references therein.

1.1.2. Multi-Server Queues with a Variable Number of Homogeneous Active Service Devices

In the context of control via the number of active service devices, the following results can be mentioned. In [9], a two-service device queue with identical service devices and a threshold control of activation of an additional service device is analyzed. In [10,11], a more general system with many identical service devices and hysteresis control via the number of active service devices is considered. A similar model but with significantly less restrictive assumptions about the arrival and service processes was investigated in [12]. A more detailed survey of the relevant research is given there. In [13], hysteresis-type control via service device reservation was applied to a multi-server priority queuing system describing the operation of a cognitive radio system.

1.1.3. Multi-Server Queues with a Variable Number of Heterogeneous Service Devices

The all above mentioned multi-server queues have identical service devices. The analysis of systems with distinct servers is essentially more complicated than the analysis of systems with identical servers. This is evidently explained as follows. In the case of identical servers, the service process in the system at an arbitrary moment is defined by the number of active servers. When service time distribution has exponential distribution, the number of active servers completely defines the total service rate at a given moment. In the case of non-identical servers, the service process at an arbitrary moment is not defined by the number of active servers. Keeping track of which servers namely are active and which are idle is required to build the Markov chain describing the behavior of the system. Consideration of a multi-dimensional process (number of active servers and indicators of which servers are active) instead of a one-dimensional process (just the number of active servers) is clearly a more difficult task.
The situation in the study of systems with distinct service devices is as follows. In [14], a system with two heterogeneous service devices and a threshold policy of control via the number of used service devices is considered. The practical motivation of model consideration is given and the invariant distribution of the system states is calculated. In [15], the optimality of the threshold-type strategy, which is used in [14], is proven. Non-optimality of such a strategy in the case of more than two service devices is forecasted. In [16], the task of finding the optimal thresholds defining the hysteresis-type strategy for the two-server system providing service to the demands arriving in batches is solved. In [17], the optimality of the threshold-type control in a system with many heterogeneous service devices is proven and a numerical study of the model is implemented. The results from [18–21] are obtained as further development of the results from [17]. In [11,22], the application of the hysteresis strategy in a system with many heterogeneous service devices is also discussed.

1.1.4. Queuing Systems with Energy Consumption and Harvesting

In the above mentioned papers, no aspects related to the consumption of the e.u. during service of demands are touched. Meanwhile, this aspect is very important in many real-world systems because more e.u. can be required to provide a higher service rate and it is desirable to reduce its use due to economical and ecological reasons. In particular, this aspect is very important in the management of the operation of wireless sensor networks. Sensors designed for monitoring the state of some objects can be installed in hard-to-reach areas and they can collect and transfer information only if they succeed in receiving (harvesting) enough e.u. The e.u. can be earned from sources, such as solar panels, and is accumulated in a small battery whose low charge makes it impossible for the transmission of information. Such situations are considered using queuing models with energy harvesting, see, e.g., [23–29]. In such models, it is assumed that e.u. arrive in the system during its operation according to a stationary Poisson process. E.u. are stored in a finite stock and service of any demand can start only in the presence of a minimal required number of e.u. in the stock. Similar queues have been studied in the literature for about seventy years and they are called assembly-like systems; queues with paired demands; taxi/passenger queues; matching systems; double-ended queues; systems with passive service devices; systems with demands’ arrival in sessions, etc. Queuing/inventory models (see, e.g., [30–32]) very close to such models are popular now. The distinction is that these models usually assume some standard replenishment policies for increasing the number of e.u. (inventory units), but not the arrival of these units in some random input flow. Aspects relating to the service rate control, depending on the amount of the harvested e.u., are considered very rare and, as such, they deserve attention. We can mention the paper [33] where the threshold strategy for service rate selection depending on the available amount of e.u. is analyzed. When the number of available e.u. increases (or decreases), higher (or lower) service rates can be used.

1.1.5. Queuing Systems with Demand Impatience

An important peculiarity of the model under study is that it takes into account the impatience of demands. Such impatience is typical in many real-world systems including communications systems, call centers, etc. Impatience means that a demand waiting for service in a buffer can renego from the system if the service is not started during some period. In the context of sensor networks, impatience is caused by the useless transmission of a signal to the coordinating node of the network due to the obsolescence of information contained in this signal. Lists of related research on queues with impatient demands can be found, e.g., in recent papers [34–39]. Account of the impatience frequently causes essential complications in the analysis because the behavior of the random process describing the dynamics of the system becomes level-dependent and the corresponding process does not belong to the classes of QBD processes or $M/G/1$- and $GI/M/1$-type structures that Markov chains were used to exhaustively investigate by M. Neuts, see [40,41].
1.2. Contributions of the Paper

As far as we know, this paper is the first one where the system with e.u. harvesting and heterogeneous service devices are considered. Two service devices differ by the service rate and the number of e.u. required for service providing. We suggest the hysteresis-type control via the activation and deactivation of the service device that consumes more energy. This service device is activated if the queue length is not below the larger threshold and enough e.u. is available. The service device is deactivated if the queue length becomes less than the smaller threshold or the number of available e.u. is not sufficient to begin service. Arrivals of demands and e.u. are described by the MAPs, which allows us to catch the possible correlation and high variability of inter-arrival times typical for many real-world systems. We assume the service times by both service devices have phase-type (PH) distribution that allows it to fit well with an arbitrary distribution, see [42], or, at least, several initial moments of service time distribution. Possible leakage of e.u. from the stock and the reneging of waiting demands from the buffer are taken into account.

The dynamic of the system is determined via a multidimensional continuous-time Markov chain that is a level-dependent QBD (Quasi-Birth-and-Death) process. The generator of this QBD is written down as the block matrix with sizes of blocks depending on the relation of the level of the chain with the thresholds of the control policy and the relation of the component of the chain describing both the number of e.u. and the number required for a demand service. Ergodicity conditions of the chain in the cases when demands are impatient and absolutely patient are presented and the invariant distribution of the chain is computed. Expressions for the computation of various performance measures are derived. The feasibility of the obtained algorithms for the computation of invariant distribution and performance measures is illustrated by several examples.

1.3. Possible Applications of the Model

Additional items required for demands’ service, which are interpreted in the paper as e.u., may have a meaning of any other resource, e.g., computer memory, service assistants, any other kind of resources (workpiece, place in a warehouse, service device, money, etc.). Therefore, the analyzed model can be used for the performance evaluation and optimization of operation of any real-world object where service of demand requires the use, besides the service device, of some additional resources and available service devices require a different amount of resources, provide different service rates, and have a different cost. In particular, the model fits for the description of the work of the node of a sensor e.u. harvesting network, which has the main, low cost, and low e.u. consuming, transmitter, and reserve transmitter that should be used only in situations when there is enough energy, and there is also a risk of information loss due to its obsolescence. The results can be used for managing the system with the use of the best values of the thresholds of control policy by verifying the ability of the system to provide the required quality of service or the desired revenue in the case of the optimal choice of control strategy.

1.4. Structure of the Text

The mathematical model under study is fully described in Section 2. A multidimensional level-dependent Markov chain (MC) with floating cardinality of blocks that determines the dynamics of the system is introduced in Section 3. The infinitesimal generator of this chain is presented. The ergodicity condition for the system and the procedure for the computation of its stationary distribution in the case of absolute patient demands are presented in Section 4. Analogous results in the case of impatient demands are given in Section 5. Expressions for the computation of the performance indicators of the system based on the invariant distribution of MC are reported in Section 6. Section 7 deals with the results of some numerical experiments and a short illustration of the possibility of a solution to the optimization problem. Appendix A contains a brief explanation of the blocks of the generator.
2. Mathematical Model

We consider a queuing model having two heterogeneous service devices and an infinite buffer for temporal storing of the demands. The scheme of the system operation is displayed in Figure 1.

![Figure 1](image)

We assume that the MAP flow of demands arrives for service in the system. This flow is defined by the underlying process \( \nu_t^{(c)}, t \geq 0 \), with the state space \( \{1, \ldots, W^{(c)}\} \) and matrices being \( D_0^{(c)} \) and \( D_1^{(c)} \), respectively, of size \( W^{(c)} \). The average arrival rate of demands \( \lambda_c \) is defined by the formula

\[
\lambda_c = \theta_c D_1^{(c)} e
\]

where the invariant probability vector \( \theta_c \) of the process \( \nu_t^{(c)} \) satisfies the system

\[
\theta_c (D_0^{(c)} + D_1^{(c)}) = 0, \quad \theta_c e = 1
\]

where \( 0 = (0, \ldots, 0) \), \( e = (1, \ldots, 1)^T \), and symbol \( T \) denotes the operation of transposition of the vector.

More information about the MAP, its characteristics, particular cases, and importance for an adequate description of flows in real-life systems can be found, e.g., in [43–46]. The key benefit of the MAP over the widely used stationary Poisson process model is the possibility to fit real flow with any value of the coefficients of variation and correlation of successive inter-arrival times. The stationary Poisson process does not give any freedom in fitting, except on account of the mean arrival rate. The coefficients of variation and correlation of inter-arrival times in the stationary Poisson process are equal to 1 and 0, correspondingly.

If service devices are unavailable at the arrival moment, the demand joins the buffer of an unlimited capacity and will be processed later on. Otherwise, the demand may try to start the receiving service.

The start of service of any demand requires the presence of a fixed number of e.u. E.u. arrive in the MAP flow defined by the underlying process \( \nu_t^{(e)} \), with the state space \( \{1, \ldots, W^{(e)}\} \) and matrices being \( D_0^{(e)} \) and \( D_1^{(e)} \), respectively. The average arrival rate of e.u. \( \lambda_e \) is defined by the formula

\[
\lambda_e = \theta_e D_1^{(e)} e
\]

where the row vector \( \theta_e \) is defined as the unique solution to the system

\[
\theta_e (D_0^{(e)} + D_1^{(e)}) = 0, \quad \theta_e e = 1.
\]
The e.u. are stored in the stock (battery) of a finite capacity $N$. If the stock is full upon the new e.u. arrival moment, this e.u. is permanently lost. Otherwise, it joins the stock.

Leakage of e.u. from the stock occurs with rate $\gamma_n$ when $n$ e.u. stay in the stock, $n = 1, N$. We assume that $\gamma_0 = 0$.

Service to demands can be provided via two service devices. The $r$th service device requires $n^{(r)}$ e.u. for a demand service. We assume that $n^{(2)} > n^{(1)}$, i.e., the second service device consumes more e.u. than the first one. The contrary case can be analyzed analogously.

Service time via the $r$th service device has the $PH$ distribution defined by the underlying $MC \eta^{(r)}_1$ with the state space being $\{1, \ldots, M^{(r)}\}$, $r = 1, 2$, of transient states. The state $M^{(r)} + 1$ is the absorbing one. It is said that such service time has an irreducible representation $(\beta^{(r)}, S^{(r)})$. Transition rates to the absorbing state are defined by the column vector $S_0^{(r)} = -S^{(r)} e$. The service rate is denoted by $\mu^{(r)} = (\beta^{(r)}(-S^{(r)})^{-1} e)^{-1}, r = 1, 2$. More useful information about the $PH$ distribution and its properties can be found, e.g., in [40].

If the required service beginning number of e.u. is unavailable, the service device suspends the work until the required number of e.u. will be accumulated in the stock. After that, the service device may immediately begin service. At the moment of the service beginning, the used service e.u. immediately disappear from the stock.

We consider the first service device as the main one. It cannot stay idle if the buffer is not empty and at least $n^{(1)}$ e.u. are available in the stock. The second service device is the reserved, assistant service device. The mechanism of its activation or deactivation is as follows. The thresholds $j$ and $k$ are assumed to be fixed, $1 \leq j \leq k$. An idle second service device is switched on if two conditions are simultaneously fulfilled: the number of demands in the queue is not less than $k$ and at least $n^{(2)}$ e.u. are available.

Interruption of service is not allowed. The second service device is switched off at the moment of service completion if (i) the queue length is below the value $j$ or (and) (ii) $n^{(2)}$ e.u. required for service of the next demand are not available.

The switching-on of the idle second service device occurs when the queue length is not less than $k$ and the number of available e.u. is not less than $n^{(2)}$. In other words, there are two scenarios of switching on (activating) the idle second service device: (i) there were not less than $n^{(2)}$ e.u., where the number of demands in the system was equal to $k - 1$ and a new demand arrives; (ii) there were not less than $k$ demands in the system, where there were $n^{(2)} - 1$ e.u. and new e.u. arrives. We will suppose that $n^{(1)} + n^{(2)} < N$. The contrary case of a small buffer capacity $N$ can be analyzed similarly to the analysis implemented here.

Demands residing in the buffer are, generally speaking, impatient. The patience time of a demand is exponentially distributed with rate $\varepsilon$. After this time finishes, the demand departs from the buffer, independently of other demands, and is considered lost.

We aim to analyze the behavior of this system and solve the optimization problem: to find thresholds $j$ and $k$ providing the minimum to the cost function. This function includes the charge for demands waiting in the buffer and the charges paid for both the use of the assistant service device during a unit of time and each activation of this device.

3. The Process Describing the Dynamics of the System and Its Generator

The dynamics of the described system are determined via the multidimensional $MC$ with continuous time

$$\xi_t = \{i_t, n_t, v_t^{(c)}, v_t^{(e)}, \eta_t\}, t \geq 0,$$

where at time $t$, $t \geq 0$,

- $i_t$ is the number of demands in the buffer, $i_t \geq 0$;
- $n_t$ is the number of e.u. in the stock, $n_t = 0, N$;
- $v_t^{(c)}$ is the state of the underlying process of the MAP of demands, $v_t^{(c)} = 1, W^{(c)}$;
- $v_t^{(e)}$ is the state of the underlying process of the MAP of e.u., $v_t^{(e)} = 1, W^{(e)}$;
vector process $\eta_i$ defines the states of the underlying processes of service in the devices. We will distinguish four macro-states of the process $\eta_i$: the first macro-state means that both service devices are idle; the second macro-state contains the states $(\eta_i^{(1)}, 0)$ when the first service device is busy while the second service device is idle; the third macro-state contains the states $(0, \eta_i^{(2)})$ when the second service device is busy while the first one is idle; and the fourth macro-state contains the states $(\eta_i^{(1)}, \eta_i^{(2)})$ when both service devices are busy. Here, $\eta_i^{(r)}$ denotes the current state of the underlying process of service in the $r$th service device, $\eta_i^{(r)} = \hat{1}, \tilde{M}^r$, $r = 1, 2$.

Here and below the notation $n = 0, N$ means that the parameter $n$ admits the values from the set $\{0, 1, \ldots, N\}$.

Note that the state space of the vector process $\eta_i$ defined by Formula (1) depends on the values of the components $\{i_t, n_t\}$ as follows.

- If $i_t = 0$ or $i_t > 0$ and $n_t = 0, n^{(1)} - 1$, then all macro-states of the process $\eta_i$ are feasible and the cardinality of its state space is equal to $M = (M^{(1)} + 1)(M^{(2)} + 1)$.
- If $i_t > 0$ and $n_t = n^{(1)}, N$, then the first and the third macro-states 0 of the process $\eta_i$ are not feasible because the first service device cannot stay idle when the queue is not empty and not less than $n^{(1)}$ e.u. are available. Correspondingly, the cardinality of the state space of the process $\eta_i$ is equal to $\hat{M} = (M^{(1)} + 1)$.

Finally, if $n_t = n^{(2)}$, $N$, then for $i_t \geq k$, the second macro-state $(\eta_i^{(1)}, 0)$ of the process $\eta_i$ is also not feasible and the cardinality of the state space of the process $\eta_i$ is equal to $\tilde{M} = M^{(1)}M^{(2)}$. In the following, when we say that the process $\eta_i$ currently has four macro-states, we mean all four possible macro-states described above. When we say that the process $\eta_i$ has two macro-states, we mean the second and fourth macro-states. When we say that the process $\eta_i$ has one macro-state, we mean the fourth macro-state.

To simplify the formulas, we will use the following notation:

- $W = W^{(c)}W^{(e)}$ is the elements’ number of the state space of the components $\{v_i^{(c)}, v_i^{(e)}\}$ of the MC $\zeta_i$;
- $D_0 = D_0^{(c)} \oplus D_0^{(e)}$, $\tilde{D}_1^{(e)} = I_{W_1} \otimes D_1^{(e)}$, $D_1^{(c)} = D_1^{(c)} \otimes I_{W_2}$;
- $\otimes$ and $\oplus$ mean the operations of the Kronecker product and the sum for the matrices, see, e.g., [47];
- $\text{diag}\{\ldots\}$ is the diagonal matrix with the diagonal entries given in the brackets;
- $\delta_{l,m}$ is equal to 1 if $l = m$ and equal to 0 if $l \neq m$;
- $N_- = Mn^{(1)} + \tilde{M}(N - n^{(1)} + 1)$;
- $N_+ = Mn^{(1)} + \tilde{M}(n^{(2)} - n^{(1)}) + \tilde{M}(N - n^{(2)} + 1)$;
- sub-generator $\bar{S}$ of the form

$$\bar{S} = \text{diag}\{0, S^{(1)}, S^{(2)}, S^{(1)} \oplus S^{(2)}\}$$

describing the rates of the vector process $\eta_i$ transition when all of its states are feasible and which do not cause service completions;

- matrix $\bar{S}_0^{(1)}$ of the form

$$\begin{pmatrix}
0 & 0_{M^{(1)}} & 0_{M^{(2)}} & 0_{\tilde{M}} \\
S_0^{(1)}_{M^{(2)} \times 1} & O_{M^{(1)} \times M^{(1)}} & O_{M^{(1)} \times M^{(2)}} & O_{M^{(1)} \times \tilde{M}} \\
O_{M^{(2)} \times 1} & O_{M^{(2)} \times M^{(1)}} & O_{M^{(2)} \times M^{(2)}} & O_{M^{(2)} \times \tilde{M}} \\
O_{\tilde{M} \times 1} & O_{\tilde{M} \times M^{(1)}} & O_{\tilde{M} \times M^{(2)}} & O_{\tilde{M} \times \tilde{M}}
\end{pmatrix},$$

describing the rates of the vector process $\eta_i$ transition when all of its states are feasible and which cause service completion in the first service device;
• matrix $S_0^{(2)}$ of the form
\[
\begin{pmatrix}
0 & \mathbf{0}_{M(1)} & \mathbf{0}_{M(2)} & \mathbf{0}_{M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(1) \times 1} & \mathbf{0}_{M(1) \times M(1)} & \mathbf{0}_{M(1) \times M(2)} & \mathbf{0}_{M(1) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(2)} & \mathbf{0}_{M(2) \times M(1)} & \mathbf{0}_{M(2) \times M(2)} & \mathbf{0}_{M(2) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M} & \mathbf{0}_{M \times M(1)} & \mathbf{0}_{M \times M(2)} & \mathbf{0}_{M \times M}
\end{pmatrix}
\]
describing the rates of the vector process $\eta_t$ transition when all of its states are feasible and which cause service completion in the second service device;

• matrix $B^{(1)}$ of the form
\[
\begin{pmatrix}
0 & \mathbf{0}_{M(1)} & \mathbf{0}_{M(2)} & \mathbf{0}_{M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(1) \times 1} & \mathbf{0}_{M(1) \times M(1)} & \mathbf{0}_{M(1) \times M(2)} & \mathbf{0}_{M(1) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(2)} & \mathbf{0}_{M(2) \times M(1)} & \mathbf{0}_{M(2) \times M(2)} & \mathbf{0}_{M(2) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M} & \mathbf{0}_{M \times M(1)} & \mathbf{0}_{M \times M(2)} & \mathbf{0}_{M \times M}
\end{pmatrix}
\]
describing the probabilities of the vector process $\eta_t$ transition when all of its states are feasible and service begins in the first service device;

• matrix $B^{(2)}$ of the form
\[
\begin{pmatrix}
0 & \mathbf{0}_{M(1)} & \mathbf{0}_{M(2)} & \mathbf{0}_{M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(1) \times 1} & \mathbf{0}_{M(1) \times M(1)} & \mathbf{0}_{M(1) \times M(2)} & \mathbf{0}_{M(1) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M(2)} & \mathbf{0}_{M(2) \times M(1)} & \mathbf{0}_{M(2) \times M(2)} & \mathbf{0}_{M(2) \times M}
\end{pmatrix}
\begin{pmatrix}
\mathbf{0}_{M} & \mathbf{0}_{M \times M(1)} & \mathbf{0}_{M \times M(2)} & \mathbf{0}_{M \times M}
\end{pmatrix}
\]
describing the probabilities of the vector process $\eta_t$ transition when all of its states are feasible and service begins in the second service device;

• matrices $A_{r,r'}$ defining the transitions of the vector process $\eta_t$ when it transits from the group of the states consisting of $r$ macro-states to the group of the states consisting of $r'$ macro-states:
\[
A_{1,2} = (O_{M \times M(1)} \ I_{M}), \quad A_{2,1} = (O_{M(1) \times M} \ I_{M}),
\]
\[
A_{4,1} = (O_{M(1) \times M} \ I_{M}), \quad A_{4,2} = (O_{1 \times M(1)} \ I_{M(1)}),
\]
\[
A_{2,4} = (O_{M(1) \times M} \ I_{M(1)}),
\]

• matrices $B_{r,r'}$ defining the transitions of the vector process $\eta_t$ when service by the second service device starts and the process $\eta_t$ transits from the group of the states consisting of $r$ macro-states to the group of the states consisting of $r'$ macro-states:
\[
B_{2,2} = (O_{M(1)} \ I_{M(1)}), \quad B_{2,1} = (O_{M(1) \times M(1)} \ I_{M(1)}),
\]
\[
B_{2,4} = (O_{M(1) \times M(1)} \ I_{M(1)}),
\]

• matrices $Z_{r,r'}$ defining the transitions of the vector process $\eta_t$ when the first service device completes service and starts the new one and the process $\eta_t$ transits from the
group of the states consisting of \( r \) macro-states to the group of the states consisting of \( r' \) macro-states:

\[
Z^{(1)}_{1,2} = \begin{pmatrix} O_{M \times M^{(1)}} & S_0^{(1)} \beta^{(1)} \otimes I_{M^{(2)}} \end{pmatrix},
\]

\[
Z^{(1)}_{2,2} = \begin{pmatrix} S_0^{(1)} \beta^{(1)} & O_{M^{(1)} \times M} \otimes S_0^{(1)} \beta^{(1)} \otimes I_{M^{(2)}} \end{pmatrix},
\]

\[
Z^{(1)}_{1,4} = \begin{pmatrix} O_{\hat{M} \times 1} & O_{\hat{M} \times M^{(1)}} & O_{\hat{M} \times M^{(2)}} & S_0^{(1)} \beta^{(1)} \otimes I_{M^{(2)}} \end{pmatrix},
\]

\[
Z^{(1)}_{2,4} = \begin{pmatrix} O_{M^{(1)} \times 1} & S_0^{(1)} \beta^{(1)} & O_{M^{(1)} \times M^{(2)}} & S_0^{(1)} \beta^{(1)} \otimes I_{M^{(2)}} \end{pmatrix};
\]

- matrices \( Z^{(2)}_{i,i'} \) defining the transitions of the vector process \( \eta_t \) when service by the second service device finishes service and starts the new one and the process \( \eta_t \) transits from the group of the states consisting of \( r \) macro-states to the group of the states consisting of \( r' \) macro-states:

\[
Z^{(2)}_{1,2} = \begin{pmatrix} O_{\hat{M} \times M^{(1)}} & I_{M^{(1)}} \otimes S_0^{(2)} \beta^{(2)} \end{pmatrix},
\]

\[
Z^{(2)}_{2,2} = \begin{pmatrix} O_{M^{(1)}} & O_{M^{(1)} \times \hat{M}} \otimes S_0^{(2)} \beta^{(2)} \end{pmatrix},
\]

\[
Z^{(2)}_{1,4} = \begin{pmatrix} O_{\hat{M} \times 1} & O_{\hat{M} \times M^{(1)}} & O_{\hat{M} \times M^{(2)}} & I_{M^{(1)}} \otimes S_0^{(2)} \beta^{(2)} \end{pmatrix},
\]

\[
Z^{(2)}_{2,4} = \begin{pmatrix} O_{M^{(1)} \times 1} & O_{M^{(1)} \times M^{(1)}} & O_{M^{(1)} \times M^{(2)}} & I_{M^{(1)}} \otimes S_0^{(2)} \beta^{(2)} \end{pmatrix};
\]

- matrix \( \hat{Z}^{(2)}_{2,2} \) defining the transitions of the vector process \( \eta_t \) when service by the second service device finishes service and the new service does not start, so the process \( \eta_t \) continues to stay in the group of the states consisting of two macro-states:

\[
\hat{Z}^{(2)} = \begin{pmatrix} O_{M^{(1)}} & O_{M^{(1)} \times \hat{M}} \\ I_{M^{(1)}} \otimes S_0^{(2)} \end{pmatrix}.
\]

Let also define the matrices \( F_{i,i'}^{n,n'} \), \( i, i' \geq 0 \), having blocks \( F_{i,i'}^{n,n'} = 0, \), defined by the following formulas:

- square matrix \( F_{0,0} \) of size \((N + 1)WM\) has:
  - diagonal blocks
    \[
    F_{0,0}^{(n,n)} = D_0 \otimes \tilde{S} + I_W \otimes (\tilde{S}_0^{(1)} + S_0^{(2)}) - \gamma_n I_{WM} + \hat{D}_1^{(e)} \otimes I_M \delta_{n,N}, \ n = 0, N; \]
  - sub-diagonal blocks \( F_{0,0}^{(n,n-1)} \) given by
    \[
    F_{0,0}^{(n,n-1)} = \gamma_n I_{WM}, \ n = 0, N; \]
  - up-diagonal blocks \( F_{0,0}^{(n,n+1)} \) given by
    \[
    F_{0,0}^{(n,n+1)} = \hat{D}_1^{(e)} \otimes I_M, \ n = 0, N - 1; \]
  - blocks \( F_{0,0}^{(n,n-n(1))}, \ n = n^{(1)}, N \) given by
    \[
    F_{0,0}^{(n,n-n(1))} = \tilde{D}_1^{(c)} \otimes B^{(1)}.
    \]


blocks \( F^{(n,n-n)}_{0,0} \), \( n = n^{(2)}, N \) given by

\[
F^{(n,n-n)}_{0,0} = D^{(c)}_1 \otimes B^{(2)} \delta_{k,1}. \tag{6}
\]

**Remark 1.** If \( n^{(1)} = 1 \), we presented above two different expressions for the matrix \( F^{(n,n-1)}_{0,0} \) (formulas for the matrix \( F^{(n-1)}_{0,0} \) and \( F^{(n,n-n)}_{0,0} \)). This happens because the number decrease in e.u. can occur either due to the leakage of e.u. or due to the service of a demand. In this and other similar cases, we merge two formulas into one formula via the summation of right-hand sides:

\[
F^{(n,n-1)}_{0,0} = D^{(c)}_1 \otimes B^{(1)} + \gamma_n I_{\hat{W}M}.
\]

- matrix \( F_{0,1} \) has only diagonal blocks.
  If \( k > 1 \), this matrix is defined by

\[
F^{(n,n)}_{0,1} = \begin{cases} 
D^{(c)}_1 \otimes I_{M}, & n = 0, n^{(1)} - 1, \\
D^{(c)}_1 \otimes A_{4,2}, & n = n^{(1)}, N.
\end{cases}
\]

If \( k = 1 \), the formula is changed to

\[
F^{(n,n)}_{0,1} = \begin{cases} 
D^{(c)}_1 \otimes I_{M}, & n = 0, n^{(1)} - 1, \\
D^{(c)}_1 \otimes A_{4,2}, & n = n^{(1)}, n^{(2)} - 1, \\
D^{(c)}_1 \otimes A_{4,1}, & n = n^{(2)}, N.
\end{cases}
\]

- Matrices \( F_{i,i}, i \geq 1 \) are square matrices of size \( WN^- \) if \( i = 1, k - 1 \) and size \( WN_+ \) if \( i > k \).

Matrices \( F_{i,i}, i \geq 1, i \neq k - 1 \) are the block-tridiagonal ones. Their diagonal blocks \( F^{(n,n)}_{i,j} \) are defined as follows.

If \( n = 0, n^{(1)} - 1 \), then for all \( i \geq 1 \)

\[
F^{(n,n)}_{i,i} = D_0 \oplus \tilde{S} + I_W \otimes (\tilde{S}^{(1)}_0 + S^{(2)}_0) - (\gamma_n + i \epsilon) I_{\hat{W}M}. \tag{9}
\]

If \( n = n^{(1)}, n^{(2)} - 1 \), then for all \( i \geq 1 \)

\[
F^{(n,n)}_{i,i} = D_0 \oplus \text{diag}\{S^{(1)}, S^{(1)} + S^{(2)}\} + I_W \otimes \hat{Z}^{(2)} - (\gamma_n + i \epsilon) I_{\hat{W}M}. \tag{10}
\]

If \( n = n^{(2)}, N \), then for \( 1 \leq i < j \), we have

\[
F^{(n,n)}_{i,j} = D_0 \oplus \text{diag}\{S^{(1)}, S^{(1)} + S^{(2)}\} + I_W \otimes \hat{Z}^{(2)} + D^{(e)}_1 \otimes I_{\hat{M}M} \delta_{n,N} - (\gamma_n + i \epsilon) I_{\hat{W}M}. \tag{11}
\]

for \( j \leq i < k \), we have

\[
F^{(n,n)}_{i,j} = D_0 \oplus \text{diag}\{S^{(1)}, S^{(1)} + S^{(2)}\} + D^{(e)}_1 \otimes I_{\hat{M}M} \delta_{n,N} - (\gamma_n + i \epsilon) I_{\hat{W}M}, \tag{12}
\]

and for \( i \geq k \), we have

\[
F^{(n,n)}_{i,j} = D_0 \oplus (S^{(1)} + S^{(2)}) + D^{(e)}_1 \otimes I_{\hat{M}M} \delta_{n,N} - (\gamma_n + i \epsilon) I_{\hat{W}M}. \tag{13}
\]
The sub-diagonal blocks $\mathcal{F}_{ij}^{(n,n-1)}$ for $1 \leq i < k$ are given by

$$\mathcal{F}_{ij}^{(n,n-1)} = \begin{cases} 
\gamma_i I_{WM}, & n = 1, n^{(1)} - 1, \\
\gamma_i I_W \otimes A_{2,4}, & n = n^{(1)}, \\
\gamma_i I_{WM}, & n = n^{(1)} + 1, N.
\end{cases}$$  \hspace{1cm} (14)

For $i \geq k$, these blocks have the form

$$\mathcal{F}_{ij}^{(n,n-1)} = \begin{cases} 
\gamma_i I_{WM}, & n = 1, n^{(1)} - 1, \\
\gamma_i I_W \otimes A_{2,4}, & n = n^{(1)}, \\
\gamma_i I_{WM}, & n = n^{(1)} + 1, n^{(2)} - 1, \\
\gamma_i I_W \otimes A_{1,2}, & n = n^{(2)}, \\
\gamma_i I_{WM}, & n = n^{(2)} + 1, N.
\end{cases}$$  \hspace{1cm} (15)

The up-diagonal blocks $\mathcal{F}_{ij}^{(n,n+1)}$ for $i = 1, k-1$ have the following form:

$$\mathcal{F}_{ij}^{(n,n+1)} = \begin{cases} 
D_i^{(e)} \otimes I_{M}, & n = 0, n^{(1)} - 2, \\
D_i^{(e)} \otimes A_{4,2}, & n = n^{(1)} - 1, \\
D_i^{(e)} \otimes I_{M}, & n = n^{(1)}, N - 1.
\end{cases}$$  \hspace{1cm} (16)

The blocks $\mathcal{F}_{ij}^{(n,n+1)}$ for $i \geq k$ are given by

$$\mathcal{F}_{ij}^{(n,n+1)} = \begin{cases} 
\hat{D}_i^{(e)} \otimes A_{4,2}, & n = n^{(1)} - 1, \\
\hat{D}_i^{(e)} \otimes I_{M}, & n = n^{(1)}, n^{(2)} - 2, \\
\hat{D}_i^{(e)} \otimes A_{2,1}, & n = n^{(2)} - 1, \\
\hat{D}_i^{(e)} \otimes I_{M}, & n = n^{(2)}, N - 1.
\end{cases}$$  \hspace{1cm} (17)

- Matrix $\mathcal{F}_{k-1,k-1}$’s non-zero blocks $\mathcal{F}_{k-1,k-1}^{(n,n)}$, $\mathcal{F}_{k-1,k-1}^{(n,n-1)}$, $\mathcal{F}_{k-1,k-1}^{(n,n+1)}$ are already defined by the formulas above (when $i = k - 1$) and it has the complementary non-zero blocks

$$\mathcal{F}_{k-1,k-1}^{(n,n-n^{(2)})} = \begin{cases} 
\hat{D}_i^{(e)} \otimes B_{2,4}, & n = n^{(1)}, n^{(2)} + n^{(1)} - 1, \\
\hat{D}_i^{(e)} \otimes B_{2,2}, & n = n^{(2)} + n^{(1)}, N.
\end{cases}$$  \hspace{1cm} (18)

- Matrices $\mathcal{F}_{i,i-1}$, $i \geq 2$ have the entries caused by the demands’ departure from the queue due to impatience and also non-zero blocks

$$\mathcal{F}_{i,i-1}^{(n^{(1)}-1,0)}, \mathcal{F}_{i,i-1}^{(n^{(2)}-1,0)}, \mathcal{F}_{i,i-1}^{(n,n-n^{(1)})}, \mathcal{F}_{i,i-1}^{(n,n-n^{(2)})},$$

which have different sizes and forms in the following three cases: $i = \frac{2}{k}, k$, and $i > k$. The size is equal to $WN_- \times WN_-$ for $i = \frac{2}{k}$, $WN_+ \times WN_-$ for $i = k$, and $WN_+ \times WN_+$ for $i > k$.

In the case of $i = \frac{2}{k}-1$, the matrix $\mathcal{F}_{i,i-1}$ is the sum of both the matrix $i\epsilon I_{WN_-}$ and the matrix $\mathcal{H}_-$ having the non-zero blocks

$$\mathcal{H}_{-}^{(n^{(1)}-1,0)}$$

defined below:

$$\mathcal{H}_{-}^{(n^{(1)}-1,0)} = D_1^{(e)} \otimes B^{(1)},$$  \hspace{1cm} (19)

and

$$\mathcal{H}_{-}^{(n,n-n^{(1)})} = \begin{cases} 
I_W \otimes Z_{2,4}^{(1)}, & n = n^{(1)}, 2n^{(1)} - 1, \\
I_W \otimes Z_{2,2}^{(1)}, & n = 2n^{(1)}, N.
\end{cases}$$  \hspace{1cm} (20)
For \( j \leq i \leq k - 1 \), there exist the blocks \( \mathcal{H}_{-}^{(n,n-n(2))} \), \( n = \overline{(n(2),N)} \):
\[
\mathcal{H}_{-}^{(n,n-n(2))} = \begin{cases} 
I_W \otimes Z_{2,4}^{(2)}, & n = \overline{(n(2),n(2)+n(1)-1)}, \\
I_W \otimes Z_{2,2}^{(2)}, & n = \overline{(n(2)+n(1),N)}. 
\end{cases}
\] (21)

For \( i = \overline{2,j-1} \), blocks \( \mathcal{H}_{-}^{(n,n-n(2))} \) for \( n = \overline{(n(2),N)} \) are equal to zero.

In the case of \( i \leq k \), the matrix \( F_{i,-1} \) is the sum of both the matrix \( \epsilon_l M_{WN} \) and the matrix \( \mathcal{H}_{+} \) having the non-zero blocks \( \mathcal{H}_{+}^{(n(1)-1,0)}, \mathcal{H}_{+}^{(n(2)-1,0)}, \mathcal{H}_{+}^{(n,n-n(1))}, \mathcal{H}_{+}^{(n,n-n(2))} \), defined below:
\[
\mathcal{H}_{+}^{(n(1)-1,0)} = \hat{D}_1^{(e)} \otimes B^{(1)}; \\
\mathcal{H}_{+}^{(n(2)-1,0)} = \hat{D}_1^{(e)} \otimes B_{2,4}.
\] (22)

Formulas for \( \mathcal{H}_{+}^{(n,n-n(1))} \) have different forms depending on the relation between the numbers \( n(2) \) and \( 2n(1) - 1 \).

In the case of \( n(2) \leq 2n(1) - 1 \), the formula is:
\[
\mathcal{H}_{+}^{(n,n-n(1))} = \begin{cases} 
I_W \otimes Z_{2,4}^{(1)}, & n = \overline{(n(1),n(2)-1)}, \\
I_W \otimes Z_{1,4}^{(1)}, & n = \overline{(n(2),2n(1)-1)}, \\
I_W \otimes Z_{1,2}^{(1)}, & n = \overline{(2n(1),n(1)+n(2)-1)}, \\
I_W \otimes S_0^{(1)} \beta^{(1)} \otimes I_{M(2)}, & n = \overline{(n(1)+n(2),N)}. 
\end{cases}
\] (24)

In the case of \( n(2) > 2n(1) - 1 \), we have the formula:
\[
\mathcal{H}_{+}^{(n,n-n(1))} = \begin{cases} 
I_W \otimes Z_{2,4}^{(1)}, & n = \overline{(n(1),2n(1)-1)}, \\
I_W \otimes Z_{1,2}^{(1)}, & n = \overline{(2n(1),n(2)-1)}, \\
I_W \otimes Z_{1,2}^{(1)}, & n = \overline{(n(2),n(1)+n(2)-1)}, \\
I_W \otimes S_0^{(1)} \beta^{(1)} \otimes I_{M(2)}, & n = \overline{(n(1)+n(2),N)}. 
\end{cases}
\] (25)

Formulas for \( \mathcal{H}_{+}^{(n,n-n(2))} \) have the following form:
\[
\mathcal{H}_{+}^{(n,n-n(2))} = \begin{cases} 
I_W \otimes Z_{1,4}^{(2)}, & n = \overline{(n(2),n(2)+n(1)-1)}, \\
I_W \otimes Z_{1,2}^{(2)}, & n = \overline{(n(2)+n(1),2n(2)-1)}, \\
I_W \otimes I_{M(2)} \otimes S_0^{(2)} \beta^{(2)}, & n = \overline{(2n(2),N)}. 
\end{cases}
\] (26)

As it was already stated above, the matrix \( F_{kk-1} \) is not square. Its size for \( k \geq 2 \) is \( WN_+ \times WN_- \). It is defined by the formula:
\[
F_{kk-1} = \epsilon k R + \mathcal{H}_0 
\] (27)

where the matrix \( R \) has the blocks \( R^{(n,n)} \) given by
\[
R^{(n,n)} = \begin{cases} 
I_W \otimes I_M, & n = \overline{(0,n(1)-1)}, \\
I_W \otimes I_{\bar{M}}, & n = \overline{(n(1),n(2)-1)}, \\
I_W \otimes (O_{\bar{M} \times M(1)} I_{\bar{M}}), & n = \overline{(n(2),N)}. 
\end{cases}
\]
and the matrix $\mathcal{H}_0$ is defined as follows:

$$\mathcal{H}_0^{(n(1)-1,0)} = \hat{D}_1^{(e)} \otimes B^{(1)},$$

(28)

$$\mathcal{H}_0^{(n(2)-1,0)} = \hat{D}_1^{(e)} \otimes B_{2,4}.$$  

(29)

Formulas for $\mathcal{H}_0^{(n,n-n^{(1)})}$ have different forms depending on the relation between the numbers $n^{(2)}$ and $2n^{(1)} - 1$.

In the case of $n^{(2)} \leq 2n^{(1)} - 1$, the formula is:

$$\mathcal{H}_0^{(n,n-n^{(1)})} = \begin{cases}
I_W \otimes Z_{2,4}^{(1)}, & n = \overline{n^{(1)}, n^{(2)} - 1}, \\
I_W \otimes Z_{1,4}^{(1)}, & n = \overline{n^{(2)}, 2n^{(1)} - 1}, \\
I_W \otimes Z_{1,2}^{(1)}, & n = 2n^{(1)}, N.
\end{cases}$$

(30)

In the case of $n^{(2)} > 2n^{(1)} - 1$, we have the formula:

$$\mathcal{H}_0^{(n,n-n^{(1)})} = \begin{cases}
I_W \otimes Z_{2,4}^{(2)}, & n = \overline{n^{(1)}, 2n^{(1)} - 1}, \\
I_W \otimes Z_{2,2}^{(1)}, & n = \overline{2n^{(1)}, n^{(2)} - 1}, \\
I_W \otimes Z_{1,2}^{(1)}, & n = n^{(2)}, N.
\end{cases}$$

(31)

Formulas for $\mathcal{H}_0^{(n,n-n^{(2)})}$ have the following form:

$$\mathcal{H}_0^{(n,n-n^{(2)})} = \begin{cases}
I_W \otimes Z_{1,4}^{(2)}, & n = \overline{n^{(2)}, n^{(2)} + n^{(1)} - 1}, \\
I_W \otimes Z_{1,2}^{(2)}, & n = n^{(2)} + n^{(1)}, N.
\end{cases}$$

(32)

- Matrix $\mathcal{F}_{1,0}$ has different forms for various values of the threshold $k$: $k = 1$ and $k > 1$. Its form is not obtained from the above formula for $\mathcal{F}_{k,k-1}$ via formal setting $k = 1$ because level 0 of the chain has another cardinality than the other levels.

In the case of $k = 1$, the size of this matrix is $\mathcal{W}N_+ \times \mathcal{W}(N + 1)M$ and the matrix is equal to the sum of the matrix $\varepsilon \mathcal{R}_1$ where the matrix $\mathcal{R}_1$ has the blocks $\mathcal{R}_1^{(n,n)}$ defined by the formula

$$\mathcal{R}_1^{(n,n)} = \begin{cases}
I_W \otimes I_M, & n = \overline{0,n^{(1)}-1}, \\
I_W \otimes (O_{M^{(1)} \times (M-M^{(1)})} I_{M^{(1)})}, & n = \overline{n^{(1)}, n^{(2)}-1}, \\
I_W \otimes (O_{M^{(1)} \times (M-M^{(1)})} I_{M^{(1)})}, & n = \overline{n^{(2)}, N},
\end{cases}$$

and the matrix has the following non-zero blocks:

Block $\mathcal{F}_{1,0}^{(n,n-n^{(1)})}$ is defined by

$$\mathcal{F}_{1,0}^{(n,n-n^{(1)})} = \begin{cases}
I_W \otimes Z_{2,4}^{(1)}, & n = \overline{n^{(1)}, n^{(2)} - 1}, \\
I_W \otimes Z_{1,4}^{(1)}, & n = n^{(2)}, N.
\end{cases}$$

(33)

Blocks $\mathcal{F}_{1,0}^{(n,n-n^{(2)})}$ are defined by the formula

$$\mathcal{F}_{1,0}^{(n,n-n^{(2)})} = I_W \otimes Z_{1,4}^{(2)} \delta_{1,1}, n = \overline{n^{(2)}, N},$$

(34)

$$\mathcal{F}_{1,0}^{(n^{(1)}-1,0)} = \hat{D}_1^{(e)} \otimes B^{(1)},$$

(35)
The cardinality of the mini-levels \( \text{MC} \) order of its components and call the group of the states of the \( \text{MC} \) having the value \( i \) of the first component as level \( i \) of the chain. The group of the states having the value \( (i, n) \) of the first two components is called the mini-level \( (i, n) \).

The cardinality of level 0 is equal to \( (N + 1)WM \). The cardinality of the mini-levels \( (i, n), i \geq 1, n = 0, n(1) - 1 \) is WM. The cardinality of the mini-levels \( (i, n), i = 1, k - 1, n = n(1), N \) is WM. The cardinality of the mini-levels \( (i, n), i \geq k, n = n(1), n(2) - 1 \) is WM. The cardinality of the mini-levels \( (i, n), i \geq k, n = n(2), N \) is WM.

Let \( F \) be the generator of \( \text{MC} \) \( \zeta_i \), \( t \geq 0 \).
Theorem 1. The infinitesimal generator $\mathcal{F}$ has the block-tridiagonal structure:

$$
\mathcal{F} = \begin{pmatrix}
F_{0,0} & F_{0,1} & 0 & 0 & \cdots \\
F_{1,0} & F_{1,1} & F_{1,2} & 0 & \cdots \\
0 & F_{2,1} & F_{2,2} & F_{2,3} & \cdots \\
& & & & \ddots \\
\end{pmatrix}
$$

where the square blocks, $F_{i,i'}$, $i, i' \geq 0$, $|i - i'| \leq 1$, have the following meaning.

The matrix $F_{i,i'}$ defines (except the diagonal elements of the matrix $F_{i,i}$) transition rates from level $i$ to level $i'$. The diagonal elements of the matrix $F_{i,i}$ are negative. The modules of these elements specify the departure rate of the MC from the respective state of MC belonging to level $i$. In turn, the matrix $F_{i,i'}$ consists of the blocks $(F_{i,i'})^{(n,n')}$ that define the rates of the chain $\xi_t$ transition from the mini-level $(i, n)$ to the mini-level $(i', n')$. An explicit form of the blocks $(F_{i,i'})^{(n,n')}$ is given by Formulas (2)–(41).

Proof of Theorem 1 is given in Appendix A.

Having explicit formulas for the blocks of the generator of the MC $\zeta_t$, we have an opportunity to analyze its stationary behavior.

4. The Case of Absolutely Patient Demands—The Ergodicity Condition of the System and Its Stationary Distribution

Due to the control by the number of active service devices, the dynamics of the MC $\zeta_t$ is level-dependent. However, if the demands are patient (i.e., $\epsilon = 0$), this dependence disappears for levels higher than $k$. Therefore, the MC $\zeta_t$ can be treated as level-independent QBD with many boundary states. It follows from [40] that the ergodicity criterion for this MC is formulated as follows.

Let $F^0 = F_{k,k}$, $F^+ = F_{k+1,k}$, $F^- = F_{k+1,k}$.

Theorem 2. The criterion for the ergodicity of MC $\zeta_t$ is the fulfillment of the inequality

$$
y F^- e > y F^+ e
$$

where the vector $y$ satisfies the system

$$
y (F^- + F^0 + F^+) = 0, \ y e = 1.
$$

The vector $y$ has a finite size $WN_+$ and the solution of the system (43) is easily implementable on the computer. After the computation of this vector, the validity of inequality can also easily be obtained.

Let us now suppose that the parameters of the system are such that inequality (42) holds good.

Then the limits

$$
\varphi\{i, n, v^{(e)}, \nu^{(e)}, \eta\} = \lim_{t \to \infty} P\{i_t = i, n_t = n, v^{(e)} = v^{(e)}, \nu^{(e)} = \nu^{(e)}, \eta_t = \eta\},
$$

which are called stationary probabilities exist and the minimal non-negative solution of the quadratic matrix equation

$$
\mathcal{T}^2 F^- + TF^0 + F^+ = 0
$$

for the matrix $\mathcal{T}$ exists as well.

Let $\varphi_i = (\varphi(i, 0), \varphi(i, 1), \ldots, \varphi(i, N))$, $i \geq 0$, be the row vector of stationary probabilities of the states of MC that belong to the level $i$. 


Denote
\[
\tilde{F} = \begin{pmatrix}
F_{0,0} & F_{0,1} & O & \ldots & O & O \\
F_{1,0} & F_{1,1} & F_{1,2} & \ldots & O & O \\
O & F_{2,1} & F_{2,2} & \ldots & O & O \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
O & O & O & \ldots & F_{k-1,k-1} & F_{k-1,k} \\
O & O & O & \ldots & F_{k,k-1} & F_{0} + TF - F \\
\end{pmatrix}
\] (45)

\textbf{Theorem 3.} The probability vectors $\varphi_i$, $i \geq 0$, are computed as follows.

The vectors $\varphi_i$, $i > k$, are computed as

\[
\varphi_i = \varphi_k T^{i-k}, \ i > k,
\]

where the matrix $T$ is the solution of Equation (44) having a spectral radius less than 1, while the vector $(\varphi_0, \ldots, \varphi_k)$ is found as the solution to the finite system of equations

\[
(\varphi_0, \ldots, \varphi_k) \tilde{F} = 0,
\]

\[
(\varphi_0, \ldots, \varphi_k) \begin{pmatrix}
I \\
I \\
\vdots \\
I \\
(I - T)^{-1}
\end{pmatrix} e = 1
\]

and matrix $\tilde{F}$ is given by Formula (45).

Proof of Theorem 3 immediately follows from [40].

\section{5. The Case of Impatient Demands—The Ergodicity Condition of the System and Its Stationary Distribution}

\textbf{Theorem 4.} If the demands are impatient, i.e., $\epsilon \neq 0$, then the MC $\zeta_t$ is ergodic for any choice of the system parameters.

\textbf{Proof.} If $\epsilon \neq 0$, then the dependence of the blocks of generator $F$ on the level does not disappear as in the previous section. In this situation, it is feasible to prove that the MC $\zeta_t$ belongs to the class of asymptotically quasi-Toeplitz Markov chains (AQTMC), which was defined in [48].

Indeed, the definition of AQTMC given in [48] requires the existence of limiting matrices

\[
Y_0 = \lim_{i \to \infty} T_i^{-1} F_{i-1} Y_1 = \lim_{i \to \infty} T_i^{-1} F_{i} + I, \ Y_2 = \lim_{i \to \infty} T_i^{-1} F_{i+1}
\]

where $T_i = -I \circ F_{i}$, $i \geq 0$ and where $\circ$ is the Hadamard product of matrices symbol, see, e.g., [49].

Using the above-obtained explicit expressions for the blocks of generator $F$, it is easy to demonstrate that if $\epsilon \neq 0$, then for the MC $\zeta_t$, the matrices $Y_0, Y_1,$ and $Y_2$ exist and are equal to $I, O,$ and $O$, respectively. Thus, the MC $\zeta_t$ belongs to the class of AQTMC.

It is proven in [48] that the condition sufficient for the ergodicity of the MC $\zeta_t$ is the fulfillment of inequality

\[
wY_0 e > wY_2 e
\]

where the row-vector $w$ satisfies the system:

\[
w(Y_0 + Y_1 + Y_2) = w, \ we = 1.
\]
Taking into account that \( Y_0 = I, Y_1 = O, \) and \( Y_2 = O, \) inequality (46) turns to the trivial inequality \( 1 > 0. \) Therefore, in the case of impatient demands, the MC \( \xi_t \) is ergodic for any choice of the system parameters. Therefore, Theorem 4 is proven.

The statement of this theorem is intuitively clear because, due to impatience, the departure rate of demands from the system unboundedly increases when the length of a queue increases. Therefore, the queue length never becomes infinite.

Let us introduce the stationary probabilities of the MC \( \xi_t \) and row vectors \( \varphi_i = (\varphi(i, 0), \varphi(i, 1), \ldots, \varphi(i, N)), \ i \geq 0, \) in the same way as in the previous section.

These vectors are defined as the unique solution of the infinite system of equilibrium equations

\[
(\varphi_0, \varphi_1, \varphi_2, \ldots) \mathcal{F} = 0, \ (\varphi_0, \varphi_1, \varphi_2, \ldots) \mathbf{e} = 1.
\]

In contrast to the case of the absolutely patient demands, in which generator \( \mathcal{F} \) for large values has the block Toeplitz-like structure, here the solution of this system does not have a matrix-geometric form. In general case, the solution of such systems is not possible. Some kinds of direct or soft truncation methods are applied to obtain approximate solutions. The most popular in the kind of soft truncation in the literature is described in [50]. However, the truncation methods may not lead to success because a good quality of approximation can be achieved only for large values of the truncation levels and the existing computer memory is not enough to solve the corresponding finite truncated system.

As it was shown above, the Markov chain \( \xi_t \) belongs to the class of AQTMCS, of which the system of equilibrium equations can be solved via the corresponding algorithm proposed in [48]. An effective and numerically stable algorithm for solving this system, which exploits the block tri-diagonal form of generator \( \mathcal{F}, \) was elaborated in [51]. We use this algorithm below for the numerical study of the system.

6. Performance Measures

With vectors \( \varphi_i, \ i \geq 0, \) computed, the values of several system performance metrics may be calculated.

The mean number \( L_{\text{buffer}} \) of demands in the buffer is calculated as

\[
L_{\text{buffer}} = \sum_{i=1}^{\infty} i \varphi_i \mathbf{e}.
\]

The mean number \( N_{\text{energy}} \) of e.u. in the stock is calculated as

\[
N_{\text{energy}} = \sum_{i=0}^{\infty} \sum_{n=1}^{N} n \varphi(i, n) \mathbf{e}.
\]

Denote

\[
\mathbf{e}^{(1)} = \begin{pmatrix} 0 \\ \mathbf{e}_{M(1)} \\ \mathbf{e}_{M(2)} \end{pmatrix}, \quad \mathbf{e}^{(2)} = \begin{pmatrix} 0 \\ \mathbf{e}^T_{M(1)} \\ \mathbf{e}^T_{M(2)} \\ \mathbf{e}_{M} \end{pmatrix}, \quad \mathbf{e}^{(*)} = \begin{pmatrix} 0 \\ \mathbf{e}^T_{M(1)} \\ \mathbf{e}^T_{M(2)} \\ \mathbf{e}_{M} \end{pmatrix},
\]

\[
\mathbf{e}^{(0)} = \begin{pmatrix} 1 \\ 0^T_{M(1)} \\ 0^T_{M(2)} \\ 0^T_{\bar{M}} \end{pmatrix}, \quad \mathbf{s} = \begin{pmatrix} 0 \\ S^{(1)} \\ S^{(2)} \\ (S^{(1)} \oplus S^{(2)}) \mathbf{e}_{M} \end{pmatrix}.
\]
The probability \( P_{1\text{,busy}} \) that the first service device is occupied at any time is calculated as
\[
P_{1\text{,busy}} = \varphi_0(e_{N+1} \otimes e^{(1)}_W) + \sum_{n=0}^{n^{(1)}-1} \sum_{i=1}^{k-1} \varphi(i,n)(e_W \otimes e^{(1)}_M) + \\
\sum_{i=1}^{k-1} \sum_{n=m^{(1)}}^{N} \varphi(i,n)(e_W \otimes e^{(1)}_M) + \\
\sum_{i=k}^{\infty} \left( \sum_{n=m^{(1)}}^{n^{(1)}-1} \varphi(i,n)(e_W \otimes e^{(1)}_M) + \sum_{n=m^{(2)}}^{N} \varphi(i,n)(e_W \otimes e^{(1)}_M) \right).
\]

The probability \( P_{2\text{,busy}} \) that the second service device is occupied at any time is calculated as
\[
P_{2\text{,busy}} = \varphi_0(e_{N+1} \otimes e^{(2)}_W) + \sum_{n=0}^{n^{(1)}-1} \sum_{i=1}^{k-1} \varphi(i,n)(e_W \otimes e^{(2)}_M) + \\
\sum_{i=1}^{k-1} \sum_{n=m^{(1)}}^{N} \varphi(i,n)(e_W \otimes e^{(2)}_M) + \\
\sum_{i=k}^{\infty} \left( \sum_{n=m^{(1)}}^{n^{(1)}-1} \varphi(i,n)(e_W \otimes e^{(2)}_M) + \sum_{n=m^{(2)}}^{N} \varphi(i,n)(e_W \otimes e^{(2)}_M) \right).
\]

The probability \( P_{\text{busy}} \) that both service devices are occupied at any time is given by
\[
P_{\text{busy}} = \varphi_0(e_{N+1} \otimes e^{(s)}_W) + \sum_{n=0}^{n^{(1)}-1} \sum_{i=1}^{k-1} \varphi(i,n)(e_W \otimes e^{(s)}_M) + \\
\sum_{i=1}^{k-1} \sum_{n=m^{(1)}}^{N} \varphi(i,n)(e_W \otimes e^{(s)}_M) + \\
\sum_{i=k}^{\infty} \left( \sum_{n=m^{(1)}}^{n^{(1)}-1} \varphi(i,n)(e_W \otimes e^{(s)}_M) + \sum_{n=m^{(2)}}^{N} \varphi(i,n)(e_W \otimes e^{(s)}_M) \right).
\]

The probability \( P_{\text{idle}} \) that both service devices are not occupied at any time is calculated as
\[
P_{\text{idle}} = \varphi_0(e_{N+1} \otimes e^{(0)}_W) + \sum_{n=0}^{n^{(1)}-1} \sum_{i=1}^{k-1} \varphi(i,n)(e_W \otimes e^{(0)}_M) + \\
\sum_{i=1}^{k-1} \sum_{n=m^{(1)}}^{N} \varphi(i,n)(e_W \otimes e^{(0)}_M) + \\
\sum_{i=k}^{\infty} \left( \sum_{n=m^{(1)}}^{n^{(1)}-1} \varphi(i,n)(e_W \otimes e^{(0)}_M) + \sum_{n=m^{(2)}}^{N} \varphi(i,n)(e_W \otimes e^{(0)}_M) \right).
\]

The probability \( P_{\text{empty}} \) that the system is empty at any time is calculated as
\[
P_{\text{empty}} = \varphi_0(e_{N+1} \otimes e^{(0)}_W).
\]

The mean number of demands in service at any time \( N_{\text{service}} \) is calculated as
\[
N_{\text{service}} = \varphi_0(e_{N+1} \otimes (e^{(1)} + e^{(2)})) + \sum_{n=0}^{n^{(1)}-1} \sum_{i=1}^{k-1} \varphi(i,n)(e_W \otimes (e^{(1)} + e^{(2)})) + \\
\sum_{i=1}^{k-1} \sum_{n=m^{(1)}}^{N} \varphi(i,n)(e_W \otimes (e^{(1)} + e^{(2)})) + \\
\sum_{i=k}^{\infty} \left( \sum_{n=m^{(1)}}^{n^{(1)}-1} \varphi(i,n)(e_W \otimes (e^{(1)} + e^{(2)})) + \sum_{n=m^{(2)}}^{N} \varphi(i,n)(e_W \otimes (e^{(1)} + e^{(2)})) \right).
\]
Remark 2. Since the generator of the considered MC is quite cumbersome, it is necessary to manage some kind of control by using the accuracy of the theoretical derivations and computer implementation of the construction of the generator of the MC and the computation of its steady-state distribution. We present here a few of the existing possibilities used for such a control.

\[
\begin{align*}
\sum_{i=1}^{k-1} \sum_{n \in \mathbb{N}(i)} \varphi(i, n) & \left( e_W \otimes \left( \frac{e_M^{(1)}}{2e_M} \right) \right) + \\
\sum_{i=k}^{\infty} \left( \sum_{n \in \mathbb{N}(i)} \varphi(i, n) \left( e_W \otimes \left( \frac{e_M^{(1)}}{2e_M} \right) \right) + 2 \sum_{n \in \mathbb{N}(2)} \varphi(i, n) \left( e_W \otimes e_M \right) \right).
\end{align*}
\]

The mean number of demands in the system at any time \(L_{\text{system}}\) is calculated as

\[
L_{\text{system}} = L_{\text{buffer}} + N_{\text{service}}.
\]

The probability \(P_{\text{energy-loss}}\) of any e.u. loss due to the battery overflow is calculated as

\[
P_{\text{energy-loss}} = \lambda^{-1}_e \left[ \varphi(0, N)(\hat{D}_1^{(e)} e_W \otimes e_M) + \sum_{i=1}^{k-1} \sum_{n \in \mathbb{N}(i)} \varphi(i, n)(\hat{D}_1^{(e)} e_W \otimes e_M) + \\
\sum_{i=k}^{\infty} \varphi(i, N)(\hat{D}_1^{(e)} e_W \otimes e_M) \right].
\]

The share \(P_{\text{energy-leak}}\) of arriving energy, which is leaked from the battery, is calculated as

\[
P_{\text{energy-leak}} = \lambda^{-1}_e \sum_{i=0}^{\infty} \sum_{n=1}^{N} \gamma_n \varphi(i, n) e.
\]

It is worth noting here that the size of the vector \(e\) in this and similar formulas is not constant. It depends on the size of the vector \(\varphi(i, n)\), which is defined by the cardinality of the corresponding mini-level \((i, n)\).

The share of the accumulated e.u., which is leaked from the battery \(P_{\text{energy-leak}}\), is calculated as

\[
P_{\text{energy-leak}} = \frac{P_{\text{energy-leak}}}{1 - P_{\text{energy-loss}}}.
\]

The probability of an arbitrary demand loss due to impatience \(P_{\text{imp-loss}}\) is calculated as

\[
P_{\text{imp-loss}} = \lambda^{-1}_e \sum_{i=1}^{\infty} i \varphi(i, n) e = \frac{\varepsilon}{\lambda_e} L_{\text{buffer}}.
\]

The departure rate \(\lambda_{\text{out}}^{(\text{out})}\) of demands that received service is calculated as

\[
\lambda_{\text{out}}^{(\text{out})} = \varphi_0 (e_{N+1} \otimes e_W \otimes \hat{s}) + \sum_{n=0}^{n-1} \sum_{i=1}^{k-1} \sum_{n \in \mathbb{N}(i)} \varphi(i, n) (e_W \otimes \left( S_0^{(1)} \otimes e_{M(2)} + e_{M(1)} \otimes S_0^{(2)} \right)) + \\
\sum_{i=k}^{\infty} \left( \sum_{n \in \mathbb{N}(i)} \varphi(i, n) \left( e_W \otimes \left( S_0^{(1)} \otimes e_{M(2)} + e_{M(1)} \otimes S_0^{(2)} \right) \right) + \\
\sum_{n \in \mathbb{N}(2)} \varphi(i, n) (e_W \otimes (S_0^{(1)} \otimes e_{M(2)} + e_{M(1)} \otimes S_0^{(2)})) \right).
\]

Remark 2. Since the generator of the considered MC is quite cumbersome, it is necessary to manage some kind of control by using the accuracy of the theoretical derivations and computer implementation of the construction of the generator of the MC and the computation of its steady-state distribution. We present here a few of the existing possibilities used for such a control.
It is clear that the following relation must be fulfilled:

\[ \lambda_c = \lambda_c^{(out)} + \epsilon L_{\text{buffer}}. \]

The departure rate of demands that received service by the first service device \( \lambda_1^{(1, out)} \) is calculated as

\[
\lambda_1^{(1, out)} = \varphi_0(e_{N+1} \otimes e_W \otimes s^{(1)}) + \sum_{n=0}^{N-1} \sum_{i=1}^{\infty} \varphi(i, n)(e_W \otimes s^{(1)}) + \sum_{i=1}^{k-1} \sum_{n=0}^{N-1} \varphi(i, n)\left( e_W \otimes \left( s_0^{(1)} \otimes e_M^{(2)} \right) \right) + \sum_{i=k}^{\infty} \left( \sum_{n=0}^{N-1} \varphi(i, n)\left(e_W \otimes \left( s_0^{(1)} \otimes e_M^{(2)} \right) \right) + \sum_{n=n+2}^{N} \varphi(i, n)(e_W \otimes (s_0^{(1)} \otimes e_M^{(2)})) \right).
\]

The departure rate \( \lambda_2^{(2, out)} \) of demands that received service by the second service device is calculated as

\[
\lambda_2^{(2, out)} = \varphi_0(e_{N+1} \otimes e_W \otimes s^{(2)}) + \sum_{n=0}^{N-1} \sum_{i=1}^{\infty} \varphi(i, n)(e_W \otimes s^{(2)}) + \sum_{i=1}^{k-1} \sum_{n=0}^{N-1} \varphi(i, n)\left( e_W \otimes \left( 0_M^{(1)} \otimes e_M^{(2)} \right) \right) + \sum_{i=k}^{\infty} \left( \sum_{n=0}^{N-1} \varphi(i, n)\left(e_W \otimes \left( 0_M^{(1)} \otimes e_M^{(2)} \right) \right) + \sum_{n=n+2}^{N} \varphi(i, n)(e_W \otimes (0_M^{(1)} \otimes s_0^{(2)})) \right).
\]

The share \( Q_1 \) of demands that received service by the first service device is calculated as

\[ Q_1 = \frac{\lambda_1^{(1, out)}}{\lambda_c^{(out)}}. \]

The share \( Q_2 \) of demands that received service by the second service device is calculated as

\[ Q_2 = \frac{\lambda_2^{(2, out)}}{\lambda_c^{(out)}}. \]

The rate of switching off the second service device \( Q_{off} \) is calculated as

\[
Q_{off} = \varphi_0(e_{N+1} \otimes e_W \otimes S_0^{(2)} e_M) + \sum_{i=0}^{N-1} \left( \sum_{n=0}^{N-1} \varphi(i, n)(e_W \otimes S_0^{(2)} e_M) + \sum_{i=1}^{k-1} \sum_{n=0}^{N-1} \varphi(i, n)\left( e_W \otimes \left( 0_M^{(1)} \otimes S_0^{(2)} \right) \right) + \sum_{i=k}^{\infty} \left( \sum_{n=0}^{N-1} \varphi(i, n)\left(e_W \otimes \left( 0_M^{(1)} \otimes S_0^{(2)} \right) \right) + \sum_{n=n+2}^{N} \varphi(i, n)(e_W \otimes (0_M^{(1)} \otimes S_0^{(2)})) \right) \right) + \sum_{n=n+2}^{N} \varphi(i, n)(e_W \otimes (0_M^{(1)} \otimes S_0^{(2)})) \right).
\]

The first summands here define the rate of switching off the second service device because the number of demands in the buffer at the service completion epoch is below the threshold \( j \). The last summand is the rate of switching off the second service device because the number of demands in the buffer at the service completion epoch is not below the threshold \( j \), but there is not enough e.u. to start a new service.
The rate $Q_{on}$ of switching on the second service device is calculated as

$$Q_{on} = \sum_{n=0}^{N} \varphi(k-1,n) F_{k-1}^{(n,n-n^{(2)})} e + \sum_{i=k}^{\infty} \varphi(i,n^{(2)}-1) H_{+}^{(n^{(2)}-1,0)} e.$$ 

**Remark 3.** According to the formula of the total probability, the mean number of e.u. spent for service of one demand is equal to $n^{(1)}Q_{1} + n^{(2)}Q_{2}$. It is clear that all which are not lost or leaked e.u. in the stationary operating system is spent for demands service.

Therefore, the following relation should be fulfilled and can be used for the accuracy of computation control:

$$\lambda_{c}(1 - P_{\text{energy-loss}} - P_{\text{energy-leak}}) = \lambda_{c}(1 - P_{\text{imp-loss}})(n^{(1)}Q_{1} + n^{(2)}Q_{2}).$$

Also, it is obvious that in the stationary operating system, the relation

$$Q_{on} = Q_{off}$$

must be fulfilled.

### 7. Numerical Examples and Optimization Problem

Let us provide some findings from numerical experiments to demonstrate the influence of various system parameters on its performance metrics.

#### 7.1. Dependence of Performance Measures on Demands’ Arrival Rate $\lambda_{c}$ for Different Values of e.u.

**Arrival Rate $\lambda_{c}$**

Let us fix the following set of system parameters.

The MAP of demands is defined by the matrices

$$D_{0}^{(c)} = \begin{pmatrix} -1.37024 & 0.0180038 \\ 0.0488727 & -0.0927569 \end{pmatrix}, \quad D_{1}^{(c)} = \begin{pmatrix} 1.34323 & 0.0900188 \\ 0.0244326 & 0.0194516 \end{pmatrix}.$$ 

This $MAP$ has the rate $\lambda_{c} = 1$ and the coefficient of the correlation of successive inter-arrival times is equal to 0.0864406. The coefficient of the variation is 6.33639.

The MAP of e.u. is defined by the matrices

$$D_{0}^{(e)} = \begin{pmatrix} -4.0702 & 0.0135021 \\ 0.0366469 & -0.168292 \end{pmatrix}, \quad D_{1}^{(e)} = \begin{pmatrix} 4.0297 & 0.0270042 \\ 0.0733051 & 0.0583402 \end{pmatrix}.$$ 

This $MAP$ has the rate $\lambda_{e} = 3$ and the coefficient of the correlation of successive inter-arrival times is equal to 0.153007. The coefficient of the variation is 9.85627. To vary in the experiments of the arrival rates $\lambda_{c}$ and $\lambda_{e}$, we multiply the matrices $D_{k}^{(c)}, D_{k}^{(e)}, k = 0, 1$ by the corresponding constants.

$PH$ distribution of service time by the first service device has the irreducible representation $(\beta^{(1)}, S^{(1)})$ where

$$\beta^{(1)} = (0.5, 0.5) \quad S^{(1)} = \begin{pmatrix} -5 & 0 \\ 0 & -1.8 \end{pmatrix}.$$ 

The mean service rate is $\mu^{(1)} = 2.64706$. The squared coefficient of variation is equal to 1.44291.
The PH distribution of service time by the second service device has the irreducible representation \((\beta^{(2)}, S^{(2)})\) where

\[
\beta^{(2)} = (1, 0.) \quad S^{(2)} = \begin{pmatrix} -2.5 & 2.5 \\ 0 & -2.5 \end{pmatrix}.
\]

The mean service rate is \(\mu^{(2)} = 1.25\). The squared coefficient of variation is equal to 0.5.

The number of e.u. \(n^{(1)}\) required for a demand service by the first service device is equal to 1. The number of e.u. \(n^{(2)}\) required for a demand service by the first service device is equal to 3. The rate \(\epsilon\) of demands reneging due to impatience is equal to 0.01. The energy leakage rate \(\gamma_n\) does not depend on the number of the e.u. in the stock and is equal to 0.05.

Let control parameters \(j\) and \(k\) be fixed as 1 and 3. This means that the second service device is activated (conditional that the required amount of e.u. is available) when the queue length reaches a value of 3. This service device is deactivated only when it finishes service when the queue is empty.

Figures 2–8 show the reliance of certain system performance indicators on the arrival rate \(\lambda_c\) for different values of the e.u. arrival rate \(\lambda_e\).

**Figure 2.** Dependence of the output rate \(\lambda_c^{(\text{out})}\) of serviced demands on the arrival rate \(\lambda_c\) for different values of e.u. arrival rate \(\lambda_e\).

**Figure 3.** Dependence of the output rate \(\lambda_c^{(1,\text{out})}\) of demands from the first service device and rate \(\lambda_c^{(2,\text{out})}\) from the second service device on arrival rate \(\lambda_c\) for different values of e.u. arrival rate \(\lambda_e\).
Figure 4. Dependence of the mean number of demands in the buffer $L_{\text{buffer}}$ on arrival rate $\lambda_c$ for different values of e.u. arrival rate $\lambda_e$.

Figure 5. Dependence of the mean number of e.u. in the stock $N_{\text{energy}}$ and the mean number of demands in service $N_{\text{service}}$ on arrival rate $\lambda_c$ for different values of e.u. arrival rate $\lambda_e$.

Figure 6. Dependence of the probability $P_1,\text{busy}$ that the first service device and the probability $P_2,\text{busy}$ that the second service device is busy on arrival rate $\lambda_c$ for different values of e.u. arrival rate $\lambda_e$.

Figure 7. Dependence of the probability $P_{\text{energy-loss}}$ of e.u. loss on arrival rate $\lambda_c$ for different values of e.u. arrival rate $\lambda_e$. 
It is evident from Figures 2 and 3 that the output rate of serviced demands increases along with the increase in demands’ arrival rate $\lambda_c$. However, for a small rate $\lambda_e = 1$ of e.u. arrival, this output becomes practically constant for $\lambda_c \geq 1.5$. It is explained by the lack of e.u. for the activation of the second service device. This is clearly seen in Figure 3 that the second service device is practically not used. This implies the sharp increase in the queue length observed in Figure 4. This increase, however, is smoothed by the mass exodus of demands from the system due to their impatience.

Figure 5, on the left, illustrates the behavior of the mean number of e.u. in the stock. This number decreases with the increase in demands’ arrival rate (and higher consumption of energy). For a small value of the e.u. rate, this number is negligible, which implies the very rare use of the second service device and the stabilization of the output rate of serviced demands which we already commented on above. Figure 5, on the right, presents the mean number of demands in service, which also confirms this stabilization. Also, it illustrates the higher involvement of the second service device into demands’ service when the e.u. flow has a higher rate. This involvement explains the increase in the rate of successfully serviced demands with the growth of $\lambda_c$ that we observed in Figure 2.

The essentially higher involvement of the second service device into demands’ service with the increase in $\lambda_c$ when the e.u. flow has a higher rate is demonstrated in Figure 6. Figure 7 shows the decrease in the loss probability of e.u. due to the stock overflow when $\lambda_c$ increases and more e.u. is spent for service. For small $\lambda_e$ (equal to 1), this loss probability becomes practically equal to zero for $\lambda_c > 1.8$ because all e.u. is quickly used for service. Figure 8, on the left, illustrates the higher involvement of the second service device to provide service when there is enough e.u. and the rate $\lambda_c$ of demands’ arrival increases.

Figure 8, on the right, shows the behavior of the rate of the second service device deactivation. It is interesting to note here that this rate is higher for the middle value (3) of $\lambda_c$. For $\lambda_e = 1$, the rate of the second service device deactivation is small because it is rarely activated (due to the lack of energy). For $\lambda_e = 5$, the rate of the second service device deactivation is relatively small because the queue length is not small and there is enough e.u. to use the second service device for a relatively long time.

Note that the form of the blocks of the generator given in Theorem 1 and the expressions for the system performance measures given in Section 6 are quite complicated. Thus, it is very important to control the accuracy of the analytical derivation of the obtained expressions of these blocks and the computer implementation of the computations of these blocks as well as the stationary probability vectors and the various performance measures of the system. During our experiments, we used many tests for the control, including the ones mentioned in Remarks 2 and 3 above, and all of them were satisfactory with an accuracy of order $10^{-11}$.

7.2. Illustration of the Impact of Correlation in Arrival Processes of Demands and e.u.

In this experiment, we define the MAPs of demands and e.u. by the following matrices.
The MAP of demands is defined by the matrices

\[ D^{(c)}_0 = \begin{pmatrix} -3.51599 & 0.0915192 \\ 0.0273807 & -0.13727 \end{pmatrix}, \quad D^{(c)}_1 = \begin{pmatrix} 3.40871 & 0.0157533 \\ 0.0120025 & 0.0978955 \end{pmatrix}. \]

This MAP has the rate \( \lambda^{(c)} = 1 \) and the coefficient of the correlation of successive inter-arrival times is equal to 0.312757. The coefficient of the variation is 10.0512.

The MAP of e.u. is defined by the matrices

\[ D^{(e)}_0 = \begin{pmatrix} -10.3867 & 0.11477 \\ 0.0225039 & -0.352749 \end{pmatrix}, \quad D^{(e)}_1 = \begin{pmatrix} 10.2258 & 0.046133 \\ 0.0365688 & 0.293676 \end{pmatrix}. \]

This MAP has the rate \( \lambda^{(e)} = 3 \) and the coefficient of the correlation of successive inter-arrival times is equal to 0.374795. The coefficient of the variation is 11.6831.

To make clear the importance of correlation in the arrival processes’ account, let us present a comparison of the shapes of dependence of three performance measures on the demand arrival rate \( \lambda^{(c)} \) for the presented correlated MAPs and for the stationary Poisson arrival processes (having zero correlation) with the rates \( \lambda^{(c)} = 1 \) and \( \lambda^{(e)} = 3 \), correspondingly.

It is obvious from Figures 9 and 10 that the existence of correlation in arrival processes has a significant effect on the system performance indicators. As a result, if the flows in a real-world system described by the model are correlated or have a significant variance in inter-arrival time, using the stationary Poisson process leads to an overly optimistic assessment of the system’s performance characteristics. This clearly motivates the necessity to analyze the systems with the MAP.

**Figure 9.** Dependence of the output rate of serviced demands \( \lambda^{(out)}_c \) on the arrival rate \( \lambda_c \) for different correlations in demand arrival processes.

**Figure 10.** Dependence of the mean number of e.u. in the stock \( N_{\text{energy}} \) and the probability of an e.u. loss \( P_{\text{energy-loss}} \) for different correlations in demand arrival processes on the mean arrival rate \( \lambda_c \).
7.3. Illustration of the Effect of the First Service Device’s Coefficient of Variation in Service Time

An account of the coefficient of variation in service time by the first service device is achieved via the assumption that this time has a $PH$ distribution. It should be noted that if service time distribution would be assumed exponential (which is a very particular case of $PH$), the analysis implemented above would not become essentially simpler. This is because the main difficulties in the analysis consist of an account of the number of e.u. available at a given moment and the moment just after service begins and also on account of the states of two service devices and the change in cardinality of the levels of $M$ depending on these states. Anyway, it is interesting to look at the influence of the variance of service time distribution. We illustrate such an influence by considering different variances of the service time distribution at the first service device.

We assume the same system’s parameters as in the first experiment and the following three variants of $PH$ distribution of service time at the first service device having the same mean service rate.

In the first variant, service time has an exponential distribution with the rate $\mu_1$. It is defined by the irreducible representation $(\beta^{(1)}, S^{(1)})$ where $\beta^{(1)} = 1$ and $S^{(1)} = -\mu_1$. The squared coefficient of variation in this distribution is equal to 1.

In the second variant, service time has Erlangian distribution with the mean value $\mu_1^{-1}$ defined by the vector $\beta^{(1)} = (1, 0)$ and sub-generator

$$S^{(1)} = \begin{pmatrix} -2\mu_1 & 2\mu_1 \\ 0 & -2\mu_1 \end{pmatrix}.$$ 

The squared coefficient of variation in this distribution is equal to 0.5.

In the third variant, service time has the hyper-exponential distribution with the mean value $\mu_1^{-1}$ defined by the vector $\beta^{(1)} = (0.02, 0.98)$ and sub-generator

$$S^{(1)} = \begin{pmatrix} -0.04\mu_1 & 0 \\ 0 & -1.96\mu_1 \end{pmatrix}.$$ 

The squared coefficient of variation in this distribution is equal to 24.5102.

In Figures 11 and 12, we show the dependencies of the output rate of the serviced demands, the mean number of the demands in the buffer, and the probability of an e.u. loss on the rate $\mu_1$ for service times at the first service device with different coefficients of variation.

![Figure 11. Dependence of the output rate of serviced demands $\lambda^{(out)}$ on the rate $\mu_1$ for service times with different coefficients of variation.](image-url)
Figure 12. Dependence of the mean number of demands in the buffer $L_{\text{buffer}}$ and the probability of an e.u. loss $P_{\text{energy-loss}}$ on the rate $\mu_1$ for service times with different coefficients of variation.

Figures 11 and 12 show that the variance in the service time at the first service device has some impact, but, under the fixed set of the system parameters, this impact is not very essential, but not negligible. For the mean number of the demands in the buffer, the difference is about 14 percent. A higher coefficient of variation worsens the quality of the system operation.

7.4. Solution of Optimization Problem

We may conceive and solve numerous optimization issues after we have solved the challenge of computing the stationary distribution of the system states and the main performance metrics of the system for an arbitrarily defined set $(j, k)$ of the thresholds, e.g., let us consider the problem of selection of control parameters $j, k$ that minimize the cost criterion (the economic function representing the system charges during a unit of time) of the form

$$E(j, k) = b\lambda c^{-1} L_{\text{buffer}} + c_2 P_{2,\text{busy}} + c_{\text{off}} Q_{\text{off}},$$

where $b$ is the charge paid per unit of time for an arbitrary demand waiting in the buffer, $c_2$ is the payment of the system due to the use (lease) of an assistant service device during a unit of time, and $c_{\text{off}}$ is the charge paid for each switching-off of the assistant service device. Evidently, the system obtains a profit from the service of demands and is charged for not enough quick access of demands to the service, for the use of an assistant service device, and for managing its switching on (or off). To maximize the revenue of the system, the cost criterion $E(j, k)$ has to be minimized via the proper choice of the thresholds $j$ and $k$.

Let us fix the following parameters of the system. Arrival flows of demands and e.u. are defined by the same matrices $D_k^{(c)}$ and $D_k^{(e)}$, $k = 0, 1$ as in the first experiment scaled to provide mean arrival rates $\lambda_c = 5$ and $\lambda_e = 7$. The rate of service by the second service device is assumed to be five times higher ($\mu_2 = 6.25$) and the impatience rate $\epsilon$ is assumed to be ten times higher than in the first experiment. The remaining system settings are the same as in the first experiment.

The costs are chosen as follows: $b = 1$, $c_2 = 1$, and $c_{\text{off}} = 15$.

Figure 13, on the left, depicts the cost criterion $E(j, k)$ dependency on the variable thresholds $j$ and $k$ in the area $1 \leq j \leq k, 1 \leq k \leq 30$. Figure 13, on the right, shows the same dependence when the arrival flows are replaced with the stationary Poisson flows having the same arrival rate.
Some comments on these Figures are as follows. In the case of the MAP flows, the value $E(1,1)$ corresponding to the permanent use of the assistant service device when enough e.u. is available is equal to 21.9769. The value $E(30,30)$ corresponding to the switching on of the assistant service device when the queue reaches the value 30 and switching it off when the number of demands in the buffer drops below 30 is equal to 10.4861. The optimal value of the cost criterion is 9.04651. It is achieved when $j = 1$ and $k = 30$. This corresponds to the switching on of the assistant service device when the queue reaches the value of 30 and switching it off only when the number of demands in the buffer drops below 1. The absolute profit gained via control defined as $\Delta = \min\{E(1,1), E(30,30)\} - E(1,30)$ is equal to 1.43959. The relative profit gained via control defined as $\frac{\Delta}{E(1,30)}$ is equal to 15.91%.

In the case of the stationary Poisson flows $E(1,1) = 23.3223$, $E(30,30) = 7.03713$, $E(1,30) = 5.72215$, and $\Delta = 1.31498$, the relative profit is 22.98%, i.e., the optimal set of the thresholds is the same, but the relative profit is higher in the case of the stationary Poisson flows. Lower profit in the case of the correlated flows is explained by the fact that periods of the presence of a long queue occur due to the bursty arrivals.

Many other different conditional and unconditional optimization problems with various criteria of quality and restrictions on values of some performance measures can be solved with the use of the presented above analysis, e.g., the cost function may be assumed not user-centric and include the penalty of the system for waiting for the demands as suggested above, but it may be provider-centric and includes the penalty for the loss of demands leading to the decrease in both the throughput of the system and the profit of the service provider.

It is worth noting that we intentionally restricted ourselves from the early beginning to the consideration of a concrete control policy of the hysteresis type defined above. This allowed us to reduce, via consideration of a Markov chain with the generator having quite a complicated form, the problem of the optimization of the cost criterion to the problem of minimization of a function of two integer variables. If a more complicated policy would be considered, the solution of the optimization problem, e.g., via the use of Markov decision processes, looks to be extremely difficult, if feasible.

8. Conclusions

In this paper, we analyzed a queuing system with two distinct service devices providing different service rates and consuming different numbers of e.u. for service of one demand. The main service device cannot stay idle if the buffer of the system is not full and the number of available e.u. is enough for service. The second, assistant service device can start service only when the queue length reaches the high threshold and the number of available e.u. is enough for service. The assistant service device can continue service after a given service completion if the queue length exceeds the low threshold and the number of available e.u. is enough for service. The parameters of the system are quite general: arrivals of demands and e.u. occur according to the MAPs, and the service time distribution is of
the PH type. Demands and e.u. can depart the system after some random waiting time, having the exponential distribution.

Under any fixed pair of the low and high thresholds of the strategy of control via activation and deactivation of the assistant service device, the behavior of the system is described via a multidimensional Markov chain. In the case of absolutely patient demands, this chain belongs to the class of level-independent QBD processes with many boundary levels and its analysis is more or less standard. When the demands are impatient, this chain belongs to the class of Asymptotically Quasi-Toeplitz MC and its analysis was implemented based on the respective results from [48,51]. The primary system performance measures are computed using non-trivial formulas. Their dependence on some parameters of the system is highlighted via numerical results. An example of solving the problem of optimal selection of the thresholds is presented.

The model under consideration can be utilized for analysis and optimization of real-world systems with service devices that require an online supply from some resource, e.g., it is suitable for modeling the operation of the node of the sensor network with e.u. harvesting.

The presented analysis can be extended to the cases of the correlated marked Markov arrival process of demands and energy, separate stocks for e.u. storing, finite buffer for demands and retrying of demands, disastrous disappearance of demands or (and) energy, service devices breakdowns, etc. An extension of the model to the controlled priority systems, see, e.g., [32], is planned.

Author Contributions: Conceptualization, C.D., A.D. and R.M.; methodology, A.D. and R.M.; software, C.D., M.P.D. and R.M.; validation, M.P.D., A.D. and R.M.; formal analysis, A.D. and R.M.; investigation, C.D., M.P.D., A.D. and R.M.; writing—original draft preparation, A.D. and R.M.; writing—review and editing, A.D. and R.M.; supervision, A.D. and R.M.; project administration, R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

Proof of Theorem 1 is carried out via the thorough examination of all potential transitions of the MC $\zeta_t$ during the infinitesimally tiny span of time. This analysis is quite complicated, first of all, due to the different state spaces of the MC for different values of the components $i_t$ and $n_t$ and the possible changes in the cardinalities of these state spaces when (a) the number of demands in the buffer $i$ departs from the state 0 and returns to this state, crosses up and down the levels $j$ and $k$; and (b) the number of e.u. $n_t$ crosses up and down the levels $n^{(1)}$ and $n^{(2)}$. Due to the possibilities of the different sizes of the jumps of the number of e.u. down (by 1, $n^{(1)}$, and $n^{(2)}$), it is necessary to keep track of the value of the number of e.u. $n_t$ not only at the moment of a jump but also just after it. This makes some formulas pretty “branching”, requiring consideration of many possible variants of relations between the state of the chain $\zeta_t$, the thresholds $j$, and $k$, and the numbers $n^{(1)}$ and $n^{(2)}$.

The three-block diagonal form of the generator $F$ is obviously explained by the fact that the number of demands arriving at the system and departing from the system is equal to one. No batch arrivals or services, like in [53], are assumed in this model.

The operation of the Kronecker product of matrices is extremely useful for representing the transition rates of several distinct Markov chains (in our case, the underlying processes of MAP flows of demands and e.u. and the vector process $\eta_t$ describing the states of the service devices). Transitions of the process $\eta_t$ are defined by the self-explanatory matrices introduced before the formulation of Theorem 1. Therefore, we present here only a brief proof of the part of the formulas.
The blocks $F_{0,0}^{(n,n)}$ have the diagonal entries, which are negative, and the non-diagonal non-negative entries defining the rates of the transitions of the MC between its states without the change in the queue length is equal to zero. The diagonal entry modulus specifies the rate of MC departure from the related state. Form (2) of $F_{0,0}^{(n,n)}$ is explained easily.

The summand
\[ D_0 \otimes \hat{S} = D_0 \otimes I_M + I_W \otimes \hat{S} \]
defines the transition rates of (i) the underlying processes of the arrival of demands and e.u. and (ii) the service underlying process $\eta_i$ without the change in the queue length. The summand $I_W \otimes \left( \hat{S}_0^{(1)} + \hat{S}_0^{(2)} \right)$ defines the rates of transitions of the underlying processes of both the arrival of demands and e.u. and of the process $\eta_i$ when the first or the second service devices complete the current service (new service does not begin because the queue is empty). The term $\gamma_{n_i} I_{WM}$ defines the departure rate $0$ from the respective state due to a leakage of an e.u. The summand $\hat{D}_1^{(e)} \otimes I_M \delta_{n,N}$ corresponds to the arrival and immediate loss of a new e.u. arriving when the stock for e.u. storing is full.

Form (3) of the block $F_{0,0}^{(n,n-1)}$ is valid when $n^{(1)} \neq 1$ (see Remark 1), which is explained simply by the fact that such a transition occurs only if an e.u. leaks from the storage.

Form (4) of the block $F_{0,0}^{(n,n+1)}$ is obvious because the component $n_1$ increases its value from $n$ to $n + 1$ only when a new e.u. arrives.

The block $F_{0,0}^{(n,n-u^{(1)})}$ in the form (5) describes the scenario when the queue is empty, the first service device is idle, and there is enough e.u. to start service by this service device ($n \geq n^{(3)}$), and a demand arrives. This demand immediately starts service (without joining the queue), which is reflected by the Kronecker multiplier $B^{(1)}$, and $n^{(1)}$ e.u. disappear from the stock.

The block $F_{0,0}^{(n,n-u^{(2)})}$ in the form (6) corresponds to the scenario when the queue is empty, the first service device is busy, the second service device is idle, there is enough e.u. to start service by the second service device ($n \geq n^{(2)}$), and a demand arrives. This demand immediately starts service (without joining the queue), which is reflected by the Kronecker multiplier $B^{(2)}$, and $n^{(2)}$ e.u. disappear from the stock.

The block $F_{0,1}^{(n,n-u^{(1)})}$ relates to the scenario in which a fresh demand arrives, is not immediately accepted for service, and joins the queues. Correspondingly, nothing can happen at the same moment with the number of accumulated e.u. Therefore, $F_{0,1}^{(n,n')}$ is equal to a zero matrix if $n \neq n'$. If the number of available e.u. is less than $n^{(1)}$, nothing can happen with the process $\eta_i$ along with the demand arrival. Therefore $F_{0,1}^{(n,n)} = \hat{D}_1^{(c)} \otimes I_M$ when $n = 0, n^{(1)} - 1$. If $n = n^{(1)}, N$, then the arrival of a new demand leads to the change in the cardinality of the state space of the process $\eta_i$. The states when both service devices are idle and the first service device is idle become impossible. The reduction in the state space is reflected via Kronecker multiplication by the matrix $A_{2,4}$. As a result of this consideration, we obtain Formula (7).

This formula was obtained in the assumption that $k > 1$. If $k = 1$, then when $n = n^{(2)}, N$, the arrival of the demand leads to the activation of the idle service device if any. Thus, the arriving demand moves to the queue only when both service devices are busy. The corresponding reduction in the state space of the process $\eta_i$ is performed via Kronecker multiplication by the matrix $A_{1,4}$.

As a result of this consideration, we obtain Formula (8).

Formulas (9)–(13) for the diagonal blocks $F_{i,j}^{(n,n)}$, $i \geq 1$, of the matrices $F_{i,j}$ are derived similar to the derivation of Formula (2) for $F_{0,0}^{(n,n)}$. However, several situations have to be treated separately.

In the case of $n = 0, n^{(1)} - 1$, the cardinality of mini-levels $(i,n)$ is the same as for mini-levels $(0,n)$. Therefore, Formula (9) repeats Formula (2) up to two modifications: the
block \( IeI_{WM} \) appears due to impatience of demands and the block \( \tilde{\mathcal{D}}^{(e)}_1 \otimes I_M \delta_{n,N} \) naturally disappears because all values on \( n \) in the considered range \( n = 0, n^{(1)} - 1 \) are not equal to \( N \).

In the case \( n = n^{(1)}, n^{(2)} - 1 \), the state space of the vector process \( \eta_i \) is reduced due to the impossibility of having an idle first service device in the presence of the queue and not less than \( n^{(1)} \) e.u. The term \( \tilde{Z}^{(2)} \) appears in (10) instead of the term \( I_W \otimes \left( \tilde{S}^{(1)}_0 + \tilde{S}^{(2)}_0 \right) \) because (a) service completion by the first service device in case \( n \geq n^{(1)} \) implies the immediate start of the next service (and decreasing the value of \( i \) to \( i - 1 \)); and (b) service completion by the second service device, in this case, is not accompanied by the immediate start of the next service due to the lack of the required e.u. (\( n < n^{(2)} \)).

Formula (11) is similar to Formula (10). Only here is the possibility of the e.u. loss in case a full stock is accounted for and a new service by the second service device cannot start, not due to the lack of e.u., but because in the considered case \( 1 \leq i < j \), the queue length is below the threshold value \( j \).

In the case when \( j \leq i < k \), the second service device can start service, but this will again lead to the decrease in the value of \( i \) to \( i - 1 \). Thus, we obtain Formula (12).

In the case when \( i \geq k \) and \( n = n^{(2)}, \tilde{N} \), both service devices are always busy. This implies the decrease in the cardinality of the state space of the process \( \eta_i \) to \( \tilde{M} \). As a result, we obtain Formula (13).

Formulas (14) and (15) describe the transitions of the considered \( MC \) within the level \( i \) caused by the e.u. leakage. Both formulas reflect that the other components of the \( MC \) \( \zeta_i \) have no transitions during the leakage moment. But, the cardinality of the state space of the process \( \eta_i \) increases from \( \tilde{M} \) to \( M \) when the number of e.u. drops below the value \( n^{(1)} \) and this cardinality increases from \( \tilde{M} \) to \( \tilde{M} \) when the number of e.u. drops below the value \( n^{(2)} \) when \( i \geq k \). The change in the cardinality from \( M \) to \( \tilde{M} \) is reflected by the matrix \( A_{24} \). The change in the cardinality from \( M \) to \( \tilde{M} \) is reflected by the matrix \( (O_{\tilde{M},M^{(1)}} I_{\delta_3}) \). This explains Formulas (14) and (15).

Formulas (16) and (17) describe the transitions of the \( MC \) \( \zeta_i \) within the level \( i \) caused by an e.u. arrival. All other components of the \( MC \) \( \zeta_i \) have no transitions during the e.u. arrival moment. But, the cardinality of the state space of the process \( \eta_i \) decreases from \( M \) to \( \tilde{M} \) when the number of e.u. reaches the value \( n^{(1)} \) and this cardinality decreases from \( M \) to \( \tilde{M} \) when the number of e.u. reaches the value \( n^{(2)} \) when \( i \geq k \). The change in the cardinality from \( M \) to \( \tilde{M} \) is reflected by the matrix \( A_{42} \).

The change in the cardinality from \( M \) to \( \tilde{M} \) is reflected by the matrix \( A_{21} \).

The form of the complementary non-zero block \( \mathcal{F}^{(n,n-n^{(2)})}_{k-1,k-1} \) in the matrix \( \mathcal{F}_{k-1,k-1} \) can be commented as follows. This block corresponds to the following scenario. The number of demands in the queue is equal to \( k - 1 \), the second service device is idle, and the number of available e.u. is enough to provide service by this service device. At this moment, a new demand arrives. According to the considered hysteresis strategy of control, this arrival triggers the switching on of the second service device and the service beginning. Simultaneously, \( n^{(2)} \) e.u. disappear from the stock. Matrix \( B_{24} \) defines the probabilities of the process \( \eta_i \) transition in the case when \( n = n^{(2)}, n^{(2)} + n^{(1)} - 1 \) and, consequently, the number of available e.u. drops below the level \( n^{(1)} \) and the cardinality of the state space increases from \( \tilde{M} \) to \( M \).

Matrix \( B_{22} \) corresponds to the case when \( n = n^{(2)} + n^{(1)}, \tilde{N} \), i.e., the number of available e.u. becomes not less than \( n^{(1)} \) and cardinality \( \tilde{M} \) does not change. Summarizing, we obtain Formula (18).

Transitions of the \( MC \zeta_i \) from level \( i, i \geq 2 \), to level \( i - 1 \) correspond to a demand departure from a queue. Such a departure can happen in the following five cases: (a) a demand abandons the queue due to impatience; (b) the first service device was idle, the number of available e.u. was \( n^{(1)} - 1 \), new e.u. arrives, and the first service device starts work (the matrix \( \mathcal{F}_{ij-1}^{(n^{(1)}-1,0)} \) provides the related rates); (c) the first service device finishes service and starts the new one because the required e.u. is available (the matrix
\( \mathcal{F}_{i,j-1}^{(n,n-n^{(1)})} \) provides the related rates; (d) the first service device was busy while the second one was idle, the number of available e.u. was \( n^{(2)} - 1 \), new e.u. arrives, the queue length is not less than \( k \), and the second service device starts work (the matrix \( \mathcal{F}_{i,j-1}^{(n^{(2)}-1,0)} \) provides the related rates); and (e) the second service device finishes service and starts the new one because the first service device is busy, the required e.u. is available, and the queue length is not less than \( j \) (the matrix \( \mathcal{F}_{i,j-1}^{(n,n-n^{(2)})} \) provides the related rates). The listed blocks

\[
\mathcal{F}_{i,j-1}^{(n^{(1)}-1,0)}, \mathcal{F}_{i,j-1}^{(n^{(2)}-1,0)}, \mathcal{F}_{i,j-1}^{(n,n-n^{(1)})}, \mathcal{F}_{i,j-1}^{(n,n-n^{(2)})},
\]

are further called in the case when \( i \leq k - 1 \) as

\[
\mathcal{H}_{-}^{(n^{(1)}-1,0)}, \mathcal{H}_{-}^{(n^{(2)}-1,0)}, \mathcal{H}_{-}^{(n,n-n^{(1)})}, \mathcal{H}_{-}^{(n,n-n^{(2)})},
\]

respectively, aiming to stress their explicit independence of \( i \). The form of these blocks is explained below.

Form (19) of the matrix \( \mathcal{H}_{-}^{(n^{(1)}-1,0)} \) is due to the fact that the e.u. arrival rate is given by the matrix \( \tilde{D}_i^{(e)} \), while the matrix \( \tilde{B}^{(1)} \) defines the transition of the vector process \( \eta_i \) at the moment of service begins with the first service device.

Matrix \( \mathcal{H}_{-}^{(n,n-n^{(1)})} \) has form (20). Here, the matrix \( Z_{2,4}^{(1)} \) contains the rates of transitions of the vector process \( \eta_i \) at the epochs of service completion and beginning by the first service device when the number of available e.u. is not enough to guarantee that the remaining (after service beginning) number of e.u. is sufficient to provide service by the first service device to the next demand, i.e., \( n = n^{(1)}, 2n^{(1)} - 1 \).

The matrix \( Z_{2,2}^{(1)} \) contains the rates of transitions of the vector process \( \eta_i \) at the moment of service completion and beginning by the first service device when the number of available e.u. allows us to guarantee that the remaining number of e.u. is enough to provide service by the first service device to the next demand, i.e., \( n = 2n^{(1)}, N \).

Analogously, the matrices \( \mathcal{H}_{+}^{(n,n-n^{(2)})}, \mathcal{H}_{+}^{(n,n-n^{(3)})} \), \( n = n^{(2)}, N \), define the rates of transitions of the vector process \( \eta_i \) at the moments of service completion and beginning by the second service device. As a result, we obtain Formula (21).

In the case \( i > k \), the blocks

\[
\mathcal{F}_{i,j-1}^{(n^{(1)}-1,0)}, \mathcal{F}_{i,j-1}^{(n^{(2)}-1,0)}, \mathcal{F}_{i,j-1}^{(n,n-n^{(1)})}, \mathcal{F}_{i,j-1}^{(n,n-n^{(2)})},
\]

are further called as

\[
\mathcal{H}_{+}^{(n^{(1)}-1,0)}, \mathcal{H}_{+}^{(n^{(2)}-1,0)}, \mathcal{H}_{+}^{(n,n-n^{(1)})}, \mathcal{H}_{+}^{(n,n-n^{(2)})},
\]

respectively. The form of these blocks is explained below.

Formula (22) is the direct analog of Formula (19). Formula (23) describes the scenario when a new e.u. arrival makes it possible to start service by the second service device (while the first one is busy and the queue length \( i \) is larger than \( k \)).

Recall that for \( i > k \), the state space of the underlying process of service consists of four macro-states if the number of available e.u. is such that \( n = 0, n^{(1)} - 1 \); two macro-states if \( n = n^{(1)}, n^{(2)} - 1 \); and one macro-state if \( n = n^{(2)}, N \). Formulas (24) and (25) describe the transition rates of the MC under study when the first server finishes service and begins the new one. During such a transition, the number of available e.u. decreases by \( n^{(1)} \). Then, the new number of macro-states of the underlying process of service depends also on the relation between the value \( n - n^{(1)} \) and the thresholds \( n^{(1)} \) and \( n^{(2)} \), leading to different scenarios of changing the number of macro-states. In the scenario \( n^{(2)} \leq 2n^{(1)} - 1 \), we obtain Formula (24), otherwise Formula (25) holds.

The explanation of Formulas (26)–(41) is similar to the one given for Formulas (2)–(25).
References

1. Dudin, A. Optimal multithreshold control for a $BMAP/G/1$ queue with $N$ service modes. *Queueing Syst.* 1998, 30, 273–287. [CrossRef]


4. Tijms, H. On a switch-over policy for controlling the workload in a queueing system with two constant service rates and fixed switch-over costs. *Z. Oper. Res.* 1977, 21, 19–32. [CrossRef]


22. Le Ny, L.; Tuffin, B. A Simple Analysis of Heterogeneous Multi-Server Threshold Queues with Hysteresis; IRISA: Sonoma, CA, USA, 2000. [CrossRef]


42. Asmussen, S. Applied Probability and Queues; Springer: New York, NY, USA, 2003; Volume 2. [CrossRef]


47. Graham, A. Kronecker Products and Matrix Calculus with Applications; Ellis Horwood: Chichester, UK, 1981. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.