mathematics

IMPACT FACTOR 2.592

CITESCORE 2.9

Article

Optimization of Open Queuing Networks with Batch Services

Elena Stankevich, Igor Tananko and Michele Pagano

Special Issue
Advances in Queueing Theory

Edited by
Prof. Dr. Anatoly Nazarov and Prof. Dr. Alexander Dudin

MDPI

https://doi.org/10.3390/math10163027

mathematics

MDPI

Article

Optimization of Open Queuing Networks with Batch Services

Elena Stankevich [1,*,†], Igor Tananko [1,*,†] and Michele Pagano [2,*,†]

[1] Department of System Analysis and Automatic Control, Faculty of Computer Science and Information Technologies, Saratov State University, 83 Astrakhanskaya St., 410012 Saratov, Russia
[2] Department of Information Engineering, University of Pisa, Via G. Caruso 16, 56122 Pisa, Italy
* Correspondence: sysan@info.sgu.ru (E.S.); tanankoie@info.sgu.ru (I.T.); michele.pagano@unipi.it (M.P.)
† These authors contributed equally to this work.

Abstract: In this paper, open queuing networks with Poisson arrivals and single-server infinite buffer queues are considered. Unlike traditional queuing models, customers are served (with exponential service time) in batches, so that the nodes are non-work-conserving. The main contribution of this work is the design of an efficient algorithm to find the batch sizes which minimize the average response time of the network. As preliminary steps at the basis of the proposed algorithm, an analytical expression of the average sojourn time in each node is derived, and it is shown that this function, depending on the batch size, has a single minimum. The goodness of the proposed algorithm and analytical formula were verified through a discrete-event simulation for an open network with a non-tree structure.

Keywords: open queuing networks; batch service; optimization

MSC: 60K20

Citation: Stankevich, E.; Tananko, I.; Pagano, M. Optimization of Open Queuing Networks with Batch Services. *Mathematics* **2022**, *10*, 3027. https://doi.org/10.3390/math10163027

Academic Editors: Anatoly Nazarov and Alexander Dudin

Received: 29 July 2022
Accepted: 19 August 2022
Published: 22 August 2022

1. Introduction

The recent growing interest in queuing networks with batch services is motivated by their use as mathematical models of packet switching networks [1], wireless sensor networks [2], manufacturing systems [3–6], Web servers [7], data-processing systems [8] and transportation systems [9].

The analysis of queues with batch services was originally introduced in [10,11] and extended to queuing networks in [12]. A review of the main results on queues with batch services can be found in [13,14].

Such a feature as batch service significantly complicates the model analysis, since it violates the one step assumption, which is at the basis of product-form queuing networks. Indeed, in general it is not possible to obtain a product-form of the stationary distribution [6,15] for a queuing network with batch services, and this consideration strongly limits their practical applicability. To overcome this problem, approximate analysis methods, based on mean value analysis (MVA) [1] or the decomposition method [6,16], have been proposed in the literature. In [15], an infinitesimal operator was constructed to obtain a stationary distribution. In [17–22], a product form of the stationary distribution for queuing networks was obtained under certain assumptions. In more detail, Chao et al. [18] supposed that at service completion, the entire batch coalesces into a single unit, and it either leaves the system or goes to another node according to given routing probabilities; the product-form solution of such a model can provide relevant bounds for the behavior of assembly processes, especially when the system is operating under heavy traffic. In [19], conditions were obtained under which Markovian queues (i.e., with Poisson arrivals of batches of tasks and exponential task-service times), with both batch arrivals and departures, have a geometric queue length probability distribution at equilibrium. Furthermore, in [20], response time density was obtained for a tandem pair of such Markovian queues. Miyazawa and Taylor [21] considered queuing

networks with batch arrivals and departures, in which additional batches arrive to the nodes when they are empty. The introduction of these extra arrivals made it possible to obtain a geometric stationary distribution, which is a (stochastic) upper bound for the original network. Such systems were further analyzed by Chao [22], who proved that the stationary distribution satisfies a set of non-standard partial balance equations and that the extra arrivals are the necessary and sufficient conditions for having a product form solution, as well as for the partial balance equation to hold.

In addition to the above-discussed probabilistic analysis, both theoretical [23,24] and practical [3–5,8,25] parameter optimization of queuing networks with batch services represent open issues of significant relevance. In more detail, one of the classes of optimization problems involves determining the optimal size of the batch, since it has a significant impact on network performance. Rabta and Reiner [3] proposed a general-purpose genetic algorithm and an approximate decomposition procedure for determining optimal batch sizes in a multi-product manufacturing system with the goal of minimizing the total cycle time. The problem of determining the optimal buffer size in a material handling system with constraints on throughput and cycle time was addressed in [4] by using queue decomposition and an iterative method. In more detail, in a material handling system a vehicle moves back and forth between consecutive workstations that have input/front and output/rear buffers. The system was analyzed as an open queuing network, in which input buffers corresponded to single-server queues with batch arrivals, workstations were modeled as traditional single-server queues and the output buffers corresponded to queues with batch services.

In [5], a queuing network with batch services was used to model semiconductor fabrication facilities. A procedure for distributing plant resources was proposed, which ensures the minimum cost of equipment for a given set of technical characteristics (volume and cycle time targets). Kar and Harrison [8] investigated data processing systems: the optimization problem of finding the batch size by maximizing the throughput was solved using mean-field techniques. Finally, [25] dealt with $M/G^{[a,b]}/1/N$ queues with bulking threshold a and maximum service capacity b; by means of renewal theory, busy period analysis and decomposition techniques. It was shown how increasing the bulking threshold affects performance indexes, such as the mean waiting time and the time-averaged number of loss customers. Then, a necessary and sufficient condition for the optimal bulking threshold that minimizes the expected waiting time was established, and an algorithm which guarantees to find the optimal threshold in polynomial time was proposed.

This paper extends the results published in [15,26] for queuing networks consisting of $M/M^b/1$ systems, where b denotes the batch size. In more detail, in [26], based on the fact that the stationary distributions of the considered $M/M^b/1$ queue and an $M/M/1$ queue with (adequately chosen) state-dependent service rates are identical, a product-form for the stationary distribution of a general queuing network with batch services has been obtained. Taking advantage of our previous result, in this work the analytical expression of the average response time of the queuing network is obtained, and it is shown that it has one minimum. This is the basis for the main original contribution of the paper, represented by an algorithm for finding the optimal batch sizes that minimize the mean response time.

It should be noted that similar results were recently derived in [27] for infinite servers queues: Using the factorial moment generating function, the authors obtained a product-form for the stationary distribution of the $M/M^b/\infty$ system; showed that the stationary distributions of the $M/M^1/\infty$ (i.e., $b = 1$) and $M/M/\infty$ queues are identical; and derived some performance indexes. In general, the assumption of infinite servers simplifies the tractability of queuing systems, and to the best of our knowledge a similar analysis has not been extended to the single server case.

The rest of the paper is organized as follows. Section 2 describes the main features of open queuing networks with batch services, and Section 3 introduces the equivalent Jackson network, originally proposed by the authors in [26]. Then, in Section 4, the expressions of the average sojourn time in each node and of the response time of the network are derived.

Section 5 presents the main original contribution of this paper, namely, an optimization algorithm for finding the vector of the batch sizes for which the average response time is minimal. Finally, Section 6 numerically illustrates the original contributions of the paper.

2. Statement of the Problem

The paper deals with large-scale networks with batch services and individual routing of the customers. In more detail, a continuous-time open queuing network N consisting of L nodes, S_i, $i = 1, \ldots, L$, and an external traffic source S_0 is considered. The arrival process is Poissonian with rate λ_0, and customer transitions between nodes S_i, $i = 1, \ldots, L$, are described by the routing matrix $\Theta = (\theta_{ij})$, $i, j = 0, \ldots, L$, where θ_{ij} defines the transition probability from node S_i to node S_j.

As in classic open queuing networks, each node S_i, $i = 1, \ldots, L$, consists of an infinite capacity single-server queue, and arriving customers are put in the waiting queue if the server is busy. It is assumed that in node S_i the customers are served in batches of size b_i. In more detail, the server stays idle until b_i customers arrive at the node; if more customers are present in the waiting queue when the server is idle, then b_i customers are selected in any order, while the others remain in the queue. The service times are exponentially distributed with parameter μ_i, $i = 1, \ldots, L$. At the end of the service, the destination node S_j of each customer, independently of the others, is determined accordingly to the routing probability θ_{ij}, $i = 1, \ldots, L$, $j = 0, 1, \ldots, L$.

The following analysis is based on the assumption that the number of possible destinations from each node S_i is significantly larger than the batch size. Therefore, the simultaneous arrival of two or more customers in a node has an infinitesimal probability, and the input flow in each node of N can be approximated as a Poisson stream of customers with intensity depending on λ_0 and $\Theta = (\theta_{ij})$.

The state of the network is described by a vector $s = (s_1, \ldots, s_L)$, where s_i indicates the number of customers at node S_i. Hence, the state space of the queuing network N can be denoted as $X = \{s : s_i \geq 0\}$.

The first step is to determine the stationary distribution $\pi(s) = (\pi_1(s_1), \ldots, \pi_L(s_L))$, $s \in X$, for the queuing network N, where the stationary distributions $\pi_i(s_i)$ in the nodes S_i, $s_i = 0, 1, \ldots$, $i = 1, \ldots, L$, can be calculated considering each node in isolation, as originally proved in [26] and briefly summarized in the next section.

3. Single Node Analysis and Equivalent Jackson Network

Let us consider the generic node S_i, $i = 1, \ldots, L$, in isolation, with input rate λ_i given by

$$\lambda_i = \frac{\omega_i}{\omega_0}\lambda_0, \ i = 1, \ldots, L, \tag{1}$$

where the vector of visitation rates $\omega = (\omega_0, \omega_1, \ldots, \omega_L)$ is the solution of the equation $\omega\Theta = \omega$ with the normalization condition $\sum_{i=0}^{L} \omega_i = 1$.

It is easy to derive the equilibrium equations for node S_i:

$$\begin{cases} \lambda_i \pi_i(n) = \mu_i \pi_i(b_i), \ n = 0, \\ \lambda_i \pi_i(n) = \lambda_i \pi_i(n-1) + \mu_i \pi_i(b_i + n), \ 1 \leq n \leq b_i - 1, \\ (\lambda_i + \mu_i)\pi_i(n) = \lambda_i \pi_i(n-1) + \mu_i \pi_i(b_i + n), \ n \geq b_i. \end{cases} \tag{2}$$

Let us introduce a birth–death process ξ_i, equivalent in steady-state probabilities to the Markov process associated with the node S_i. The process ξ_i is defined on the set of states $\{0, 1, \ldots\}$, with birth rates $\lambda_i = \lambda_i(n)$, which do not depend on the state n, $n \in \{0, 1, \ldots\}$, and death rates $\widetilde{\mu}_i(n)$, $n \in \{1, 2, \ldots\}$. The state space and the parameters λ_i of the process ξ_i correspond to the state space $\{0, 1, \ldots\}$ and the input rate of node S_i defined by Equation (1).

To determine the rates $\widetilde{\mu}_i(n)$, $n = 1, 2, \ldots$, note that the steady-state probabilities of the birth–death process ξ_i are given by

$$\pi_i(k) = \pi_i(0) \prod_{n=1}^{k} \frac{\lambda_i}{\widetilde{\mu}_i(n)},\ k = 1, 2, \ldots, \tag{3}$$

where

$$\pi_i(0) = \left(1 + \sum_{k=1}^{\infty} \prod_{n=1}^{k} \frac{\lambda_i}{\widetilde{\mu}_i(n)}\right)^{-1},\ i = 1, \ldots, L.$$

By replacing (3) in (2), the following expressions are obtained for $\widetilde{\mu}_i(n)$, $n = 1, 2, \ldots$:

$$\begin{cases} \widetilde{\mu}_i(n) = \lambda_i - \mu_i \dfrac{\lambda_i^{b_i}}{\widetilde{\mu}_i(n+1) \cdot \ldots \cdot \widetilde{\mu}_i(b_i + n)},\ 1 \le n \le b_i - 1, \\ \widetilde{\mu}_i(n) = \lambda_i + \mu_i - \mu_i \dfrac{\lambda_i^{b_i}}{\widetilde{\mu}_i(n+1) \cdot \ldots \cdot \widetilde{\mu}_i(b_i + n)},\ n \ge b_i. \end{cases} \tag{4}$$

Let $M_i = \lim\limits_{n \to \infty} \widetilde{\mu}_i(n)$; if the limit exists, then:

$$\mu_i \lambda_i^{b_i} = (\lambda_i + \mu_i - M_i) M_i^{b_i}$$

or

$$M_i^{b_i+1} - (\lambda_i + \mu_i) M_i^{b_i} + \lambda_i^{b_i} \mu_i = 0\,. \tag{5}$$

The existence of the equivalent birth–death process ξ_i depends upon the existence of a positive solution of the previous equation, which satisfies the stability condition for each node S_i.

As proved in [26], taking into account the stability condition, equation (5) has a unique root M_i, belonging to the open interval

$$\left(\frac{b_i(\lambda_i + \mu_i)}{b_i + 1},\ \frac{(\lambda_i + \mu_i)^{b_i+1} - \lambda_i^{b_i} \mu_i}{(\lambda_i + \mu_i)^{b_i}}\right), \tag{6}$$

which in the general case can be determined numerically. From (4) it follows that

$$\widetilde{\mu}_i(b_i) = \widetilde{\mu}_i(b_i + 1) = \widetilde{\mu}_i(b_i + 2) = \cdots = M_i,$$

and the service rates $\widetilde{\mu}_i(b_i - 1)$, $\widetilde{\mu}_i(b_i - 2), \ldots, \widetilde{\mu}_i(1)$ can be easily derived recursively.

Since the equivalent process ξ_i can be built for any node S_i, it is possible to define a Jackson network $\widetilde{N}$ with nodes $\widetilde{S}_i$, equivalent in stationary distribution to the original queuing network N with batch services. According to the previous results for the single node, $\widetilde{N}$ has state-dependent service rates $\widetilde{\mu}_i(n)$, where n is the number of customers in the node $\widetilde{S}_i$, $n = 1, 2, \ldots$, $i = 1, \ldots, L$, and its arrival rates and routing probabilities are the same as in the original queuing network N.

Moreover, N and its equivalent network $\widetilde{N}$ are stable if and only if the utilization coefficient in the node S_i, $i = 1, \ldots, L$,

$$\rho_i = \frac{\lambda_i}{b_i \mu_i} < 1.$$

Under such conditions, the stationary distribution for $\widetilde{N}$ is given by

$$\pi(s) = \prod_{i=1}^{L} \pi_i(s_i),\ s \in X, \tag{7}$$

where

$$\pi_i(s_i) = \pi_i(0) \prod_{n=1}^{s_i} \frac{\lambda_i}{\widetilde{\mu}_i(n)}.$$

4. Stationary Response Time

The stationary distribution $\pi(s)$ for $\widetilde{N}$ allows us to calculate the average values of various performance parameters.

To characterized the queuing network as a whole, the most relevant index is the average time spent by a customer in the queuing network, known as response time:

$$\bar{\tau} = \frac{1}{\lambda_0} \sum_{i=1}^{L} \lambda_i \bar{u}_i,$$

which is given by the weighted sum of the average sojourn times $\bar{u}_i$ in all the nodes S_i.

An elegant closed-form expression for the average sojourn time is provided by the following Theorem, which is valid for a generic node S_i, $i = 1, \ldots, L$. For sake of simplicity, in the proof of the theorem, the subscript i identifying the node is omitted.

Theorem 1. *The average sojourn time in a queuing system with batch services of fixed size b is given by*

$$\bar{u} = \frac{b-1}{2\lambda} + \frac{1}{M-\lambda}, \tag{8}$$

where λ is the arrival rate and M is the unique solution of (5), satisfying the stability condition.

Proof of Theorem 1. By substituting (3) into the first equation of (2), we get

$$\frac{\lambda}{\mu} = \frac{\lambda^b}{\widetilde{\mu}(1)\widetilde{\mu}(2)\ldots\widetilde{\mu}(b)},$$

and by taking into account the latter equality:

$$\pi(n+b) = \pi(0)\frac{\lambda}{\mu}\left(\frac{\lambda}{M}\right)^n, \; n \geq 0, \; b \geq 1. \tag{9}$$

Having defined $x = \frac{\lambda}{M}$, the system (2) can be rewritten as follows:

$$\begin{cases} \pi(b) = \pi(0)\frac{\lambda}{\mu}, \\ \pi(n) = \pi(0)\sum\limits_{i=0}^{n} x^i, \; 1 \leq n \leq b-1, \\ \pi(n) = \pi(0)x^{n-b}\frac{\lambda}{\mu}, \; n \geq b. \end{cases} \tag{10}$$

When taking into account Little's law and the definition of the average number of customers in the system, the average sojourn time is given by

$$\bar{u} = \frac{1}{\lambda} \sum_{n=0}^{\infty} n\pi(n),$$

and by substituting the expressions of the state probabilities given by (10), the previous expression becomes

$$\bar{u} = \frac{1}{\lambda}\Big(\pi(0)(1+x) + 2\pi(0)(1+x+x^2) + \cdots + (b-1)\pi(0)(1+x+\cdots+x^{b-1}) +$$

$$+\sum_{n=b}^{\infty} n\pi(0)x^{n-b}\frac{\lambda}{\mu}\Bigg)$$

or rearrange the terms according to the powers of x:

$$\bar{u}=\frac{\pi(0)}{\lambda}\left((1+x)\sum_{i=1}^{b-1} i+x^2\sum_{i=2}^{b-1} i+\cdots+x^{b-1}(b-1)+\sum_{n=b}^{\infty} nx^{n-b}\frac{\lambda}{\mu}\right).$$

The next step of the proof consists in finding the value of $\pi(0)$. To this end, the system (2) can be rewritten in the form

$$\begin{cases}\lambda\pi(0)=\mu\pi(b),\\ \lambda\pi(1)=\mu\pi(b)+\mu\pi(b+1),\\ \ldots\\ \lambda\pi(b-1)=\mu(\pi(b)+\pi(b+1)+\cdots+\pi(2b-1)),\\ \lambda\pi(b)=\mu(\pi(b+1)+\pi(b+2)+\cdots+\pi(2b+1)),\\ \ldots\end{cases}$$

By summing up all the equations of this system, we obtain

$$\lambda\sum_{n=0}^{\infty}\pi(n)=b\mu\sum_{n=b}^{\infty}\pi(n)$$

or

$$\sum_{n=b}^{\infty}\pi(n)=\frac{\lambda}{b\mu}. \tag{11}$$

By substituting (9) into (11) and applying the formula for the sum of an infinite geometric progression, the probability of an empty system becomes

$$\pi(0)=\frac{1-x}{b}.$$

The last part of the proof consists in manipulating the expression of $\bar{u}$, which can be significantly simplified thanks to the specific structure of the state probabilities of the system and the properties of M. In more detail:

$$\bar{u}=\frac{1-x}{b\lambda}\left((1+x)\sum_{i=1}^{b-1} i+x^2\sum_{i=2}^{b-1} i+\cdots+x^{b-1}(b-1)+\sum_{n=b}^{\infty} nx^{n-b}\frac{\lambda}{\mu}\right)$$

or, by rearranging the terms:

$$\bar{u}=\frac{1}{b\lambda}\left((1-x^2)\sum_{i=1}^{b-1} i+(1-x)x^2\sum_{i=2}^{b-1} i+\cdots+(1-x)x^{b-1}(b-1)\right)+$$

$$+\frac{1-x}{b\lambda}\sum_{n=b}^{\infty} nx^{n-b}\frac{\lambda}{\mu}=\frac{1}{b\lambda}\Bigg(\sum_{i=1}^{b-1} i-\Bigg[\left(\sum_{i=1}^{b-1} i-\sum_{i=2}^{b-1} i\right)x^2+\left(\sum_{i=2}^{b-1} i-\sum_{i=3}^{b-1} i\right)x^3+\ldots$$

$$+((b-2)+(b-1)-(b-1))x^{b-1}+(b-1)x^b\Bigg]\Bigg)+\frac{1-x}{b\lambda}\sum_{n=b}^{\infty} nx^{n-b}\frac{\lambda}{\mu}$$

and after simple algebraic manipulations:

$$\bar{u}=\frac{1}{b\lambda}\left(\sum_{i=1}^{b-1} i-\sum_{i=1}^{b-1}(i-1)x^i\right)+\frac{1-x}{b\lambda}\sum_{n=b}^{\infty} nx^{n-b}\frac{\lambda}{\mu}.$$

Applying the formulas for the sums of arithmetic and arithmetic-geometric progressions, we obtain

$$\bar{u} = \frac{b-1}{2\lambda} - \frac{x^2(1-x^{b-1})}{\lambda b(1-x)^2} + \frac{(b-1)x^{b+1}}{\lambda b(1-x)} + \frac{x}{b\mu(1-x)} + \frac{1}{\mu}$$

and, after the inverse substitution of $x = \frac{\lambda}{M}$:

$$\bar{u} = \frac{b-1}{2\lambda} - \frac{\frac{\lambda^2}{M^2}(1-\frac{\lambda^{b-1}}{M^{b-1}})}{\lambda b(1-\frac{\lambda}{M})^2} + \frac{(b-1)\frac{\lambda^{b+1}}{M^{b+1}}}{\lambda b(1-\frac{\lambda}{M})} + \frac{\frac{\lambda}{M}}{b\mu(1-\frac{\lambda}{M})} + \frac{1}{\mu} =$$

$$= \frac{b-1}{2\lambda} - \frac{\lambda(M^{b-1}-\lambda^{b-1})}{bM^{b-1}(M-\lambda)^2} + \frac{(b-1)\lambda^b}{bM^b(M-\lambda)} + \frac{\lambda}{b\mu(M-\lambda)} + \frac{1}{\mu}.$$

The sum of the second, third and fourth terms in the latter equation becomes

$$\frac{-\lambda M(M^{b-1}-\lambda^{b-1}) + (M-\lambda)(b-1)\lambda^b}{bM^b(M-\lambda)^2} + \frac{\lambda}{b\mu(M-\lambda)} =$$

$$= \frac{-\lambda M^b + \lambda^b M + Mb\lambda^b - M\lambda^b - \lambda^{b+1}b + \lambda^{b+1}}{bM^b(M-\lambda)^2} + \frac{\lambda}{b\mu(M-\lambda)} =$$

$$= \frac{-\lambda M^b + Mb\lambda^b - \lambda^{b+1}b + \lambda^{b+1}}{bM^b(M-\lambda)^2} + \frac{\lambda}{b\mu(M-\lambda)} = \frac{b\lambda^b(M-\lambda) - \lambda(M^b-\lambda^b)}{bM^b(M-\lambda)^2} +$$

$$+ \frac{\lambda}{b\mu(M-\lambda)} = \frac{b\lambda^b\mu(M-\lambda) - \lambda\mu(M^b-\lambda^b) + M^b(M-\lambda)\lambda}{bM^b(M-\lambda)^2\mu}$$

and, taking into account the well-known formula:

$$a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \cdots + ab^{n-2} + b^{n-1}),$$

and the average sojourn time can be rewritten as:

$$\bar{u} = \frac{b-1}{2\lambda} + \frac{b\lambda^b\mu - \lambda\mu(M^{b-1} + M^{b-2}\lambda + \cdots + M\lambda^{b-2} + \lambda^{b-1}) + M^b\lambda}{bM^b\mu(M-\lambda)} + \frac{1}{\mu}.$$

Finally, from equation (5) it is easy to find that

$$M^{b+1} - (\lambda+\mu)M^b + \lambda^b\mu = (M-\lambda)(M^b - \mu M^{b-1} - \mu M^{b-2}\lambda - \cdots - \mu M\lambda^{b-2} - \mu\lambda^{b-1}) = 0.$$

Therefore, since $M > \lambda$,

$$M^b - \mu(M^{b-1} + M^{b-2}\lambda + \cdots + M\lambda^{b-2} + \lambda^{b-1}) = 0$$

and then:

$$\bar{u} = \frac{b-1}{2\lambda} + \frac{\lambda^b}{M^b(M-\lambda)} + \frac{1}{\mu} =$$

$$= \frac{b-1}{2\lambda} + \frac{\lambda^b\mu + M^{b+1} - M^b\lambda}{M^b(M-\lambda)\mu} = \frac{b-1}{2\lambda} + \frac{M^b\mu}{M^b(M-\lambda)\mu} = \frac{b-1}{2\lambda} + \frac{1}{M-\lambda}.$$

Thus, we got expression (8). □

Note that for $b = 1$, the root of the equation (5) is $M = \mu$, and hence $\bar{u} = \frac{1}{\mu-\lambda}$ in accordance with the well-known result for the $M/M/1$ queue.

5. Average Response Time Optimization

The average response time is a relevant performance index that depends, as highlighted in the previous section, on the routing matrix and the node sojourn times. The latter depend on the network parameters not only explicitly, but also through the values of the roots M_i of (5).

Assuming that the network structure and the incoming traffic rate are fixed parameters, in this section we propose an efficient algorithm to find the optimal batch size vector $b = (b_1, \ldots, b_L)$, for which the average response time of the queuing network N is minimized.

Our optimization algorithm is based on the following two considerations. Firstly, as shown above, the evolution of each node is independent of the rest of the network in a probabilistic sense: the arrival rate to any node of N does not change when the service discipline (namely, the batch size) changes in other nodes, and the input rates λ_i depends only on the routing matrix. Therefore, the problem of finding the minimum average response time $\bar{\tau}$ of N is decomposed into solving L minimization problems for the functions $\bar{u}_i(b_i)$ in the nodes S_i, $i = 1, \ldots, L$. Secondly, the function $\bar{u}_i(b_i)$ has only one minimum for values of b_i satisfying the stability condition $\lambda_i < b_i\mu_i$.

As in the previous section, for sake of clarity, in the following we omit the subscript i, since the same considerations are valid for all the nodes. To prove that $\bar{u}(b)$ has one minimum, we treat b as a continuous parameter and calculate the second derivative of $\bar{u}(b)$ with respect to b:

$$\bar{u}'' = \frac{(M-\lambda)M'' - 2(M')^2}{(\lambda - M)^3}.$$

From (6) it is easy to see that M converges to $\lambda + \mu$ as $b \to \infty$. Moreover, the sequence of the roots M_b is monotone increasing for any $b > \frac{\lambda}{\mu}$. To prove the latter statement, consider the function

$$f_b(M) = M^{b+1} - (\lambda + \mu)M^b + \lambda^b\mu;$$

i.e., M_b is the largest root of $f_b(M) = 0$. Hence:

$$M_b^{b+1} = (\lambda + \mu)M_b^b - \lambda^b\mu$$

and

$$f_{b+1}(M_b) = M_b^{b+2} - (\lambda + \mu)M_b^{b+1} + \lambda^{b+1}\mu \;=\; \lambda^b\mu(\lambda - M_b) \;< 0,$$

since $M_b > \lambda$. This means that for the batch size $b+1$, the largest solution of $f_{b+1}(M) = 0$ must satisfy the inequality $M_{b+1} > M_b$, since $f_{b+1}(M)$ is continuous and $f_{b+1}(\lambda + \mu) > 0$. Hence, the function $y = M(b)$, again considering b as a continuous variable, has an horizontal asymptote as $b \to \infty$ and is monotone increasing; as a consequence of the existence of a finite limit, its increase rate decreases for sufficiently high values of b, and its second derivative M'' must be negative. Moreover, as shown in the Figures 1–3 below for various ratios of λ and μ, the function $y = M(b)$ is concave (convex upwards) for all values of b satisfying the stability condition. Furthermore, since $M > \lambda$, $\bar{u_i}'' > 0$ and the average sojourn time is a convex function (and has one minimum).

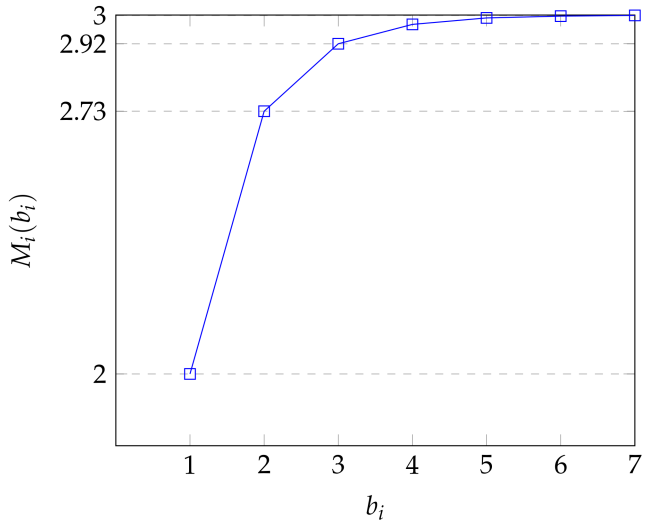

Figure 1. Limit service rate in the node S_i for $\lambda_i = 1.0$, $\mu_i = 2.0$.

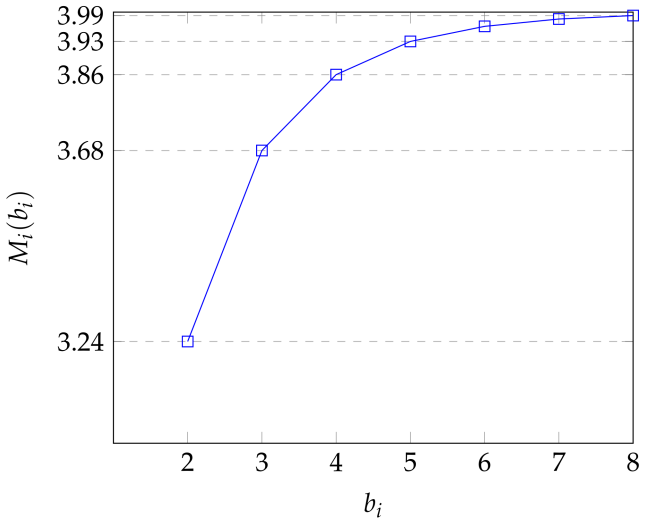

Figure 2. Limit service rate in the node S_i for $\lambda_i = 2.0$, $\mu_i = 2.0$.

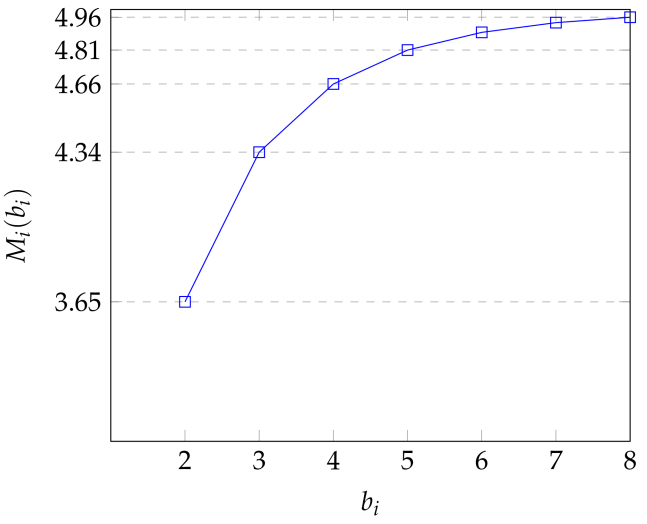

Figure 3. Limit service rate in the node S_i for $\lambda_i = 3.0$, $\mu_i = 2.0$.

Hence, we propose the following optimization algorithm to determine the optimal batch size vector:

1. Find the vector ω solution of equation $\omega\Theta = \omega$ under the normalization condition $\sum_{i=0}^{L} \omega_i = 1$.
2. Compute the arrival rates vector $\lambda = (\lambda_1, \ldots, \lambda_L)$, as
$$\lambda_i = \frac{\omega_i}{\omega_0}\lambda_0, i = 1, \ldots, L.$$
3. Initialize $i = 0$.
4. $i = i + 1$. If $i > L$, then go to step 14.
5. Initialize $k = 0$ that represents the iteration number.
6. Find the initial value b_i satisfying the condition $\lambda_i < b_i\mu_i$.
7. Find the root M_i of the equation (5).
8. Calculate $\widetilde{\mu}_i^k = (\widetilde{\mu}_i^k(1), \ldots, \widetilde{\mu}_i^k(b_i))$, where $\widetilde{\mu}_i^k(n) = M_i$, if $n \geq b_i$ and the rates $\widetilde{\mu}_i^k(n)$ are calculated according to (4), if $n < b_i$.
9. Calculate the stationary distribution according to (7).
10. Calculate the average sojourn time at the node S_i according to (8).
11. $k = k + 1$ and increase b_i by 1.
12. Repeat steps 7–10.
13. If $\bar{u}_i^k < \bar{u}_i^{k-1}$, then go to step 11. If $\bar{u}_i^k \geq \bar{u}_i^{k-1}$, then decrease b_i by 1 and go to step 4.
14. Return the vector b. The algorithm is complete.

6. Numerical Examples

To highlight the correctness of the results in the previous sections, we provide some numerical examples for a single node and a large queuing network, satisfying the assumptions under which the product-form solution was derived.

At first we focus on the sojourn time in a generic node S_i, $i \in \{1, \ldots, L\}$, with input rate λ_i and service rate μ_i for batches of b_i customers. Figures 4 and 5 show the behavior of the sojourn time $\bar{u}_i(b_i)$ for $\mu_i = 2.0$ and $\mu_i = 3.0$, respectively. In more detail, $\bar{u}_i(b_i)$ is plotted for different values of λ_i (namely, $\lambda_i = 1.0, \ldots, 5.0$) as a function of the "allowed" values of the batch size, i.e., the values of b_i satisfying the condition $\rho_i < 1$ for the existence of a stationary regime.

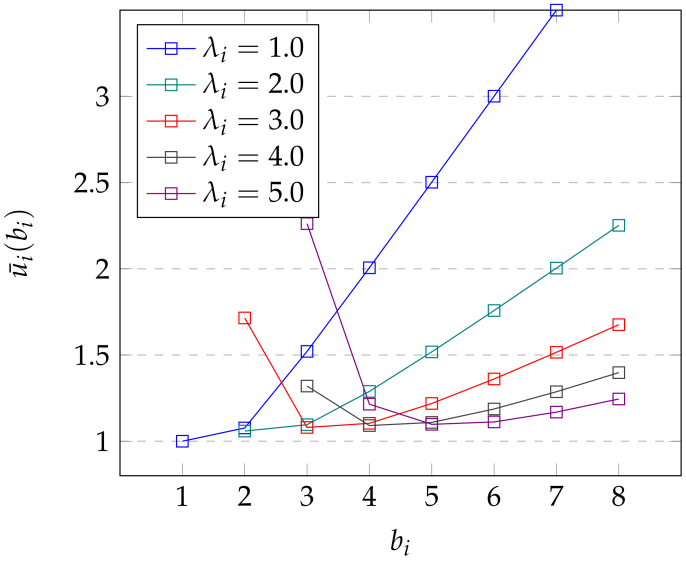

Figure 4. Average sojourn times in the node S_i for $\mu_i = 2.0$.

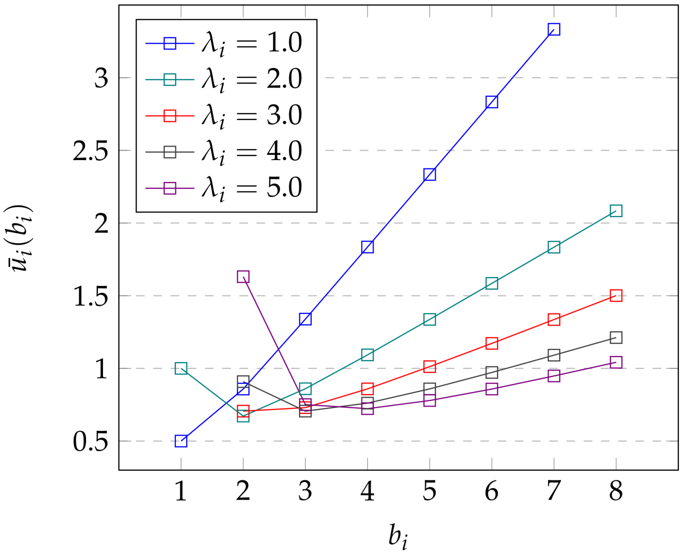

Figure 5. Average sojourn times in the node S_i for $\mu_i = 3.0$.

Both figures show that the function $\bar{u}_i(b_i)$ has one minimum for a specific value of ρ_i depending on the system's parameters. For instance, if $\lambda_i = 1$, then the optimal batch size is one for both values of μ_i, whereas for $\lambda_i = 2$ the optimal b_i is already equal to two; and for higher input rates it depends on μ_i (compare the graphs for $\lambda_i = 5$ in Figures 4 and 5).

Next, consider the queuing network N, consisting of $L = 11$ nodes, with input rate $\lambda_0 = 2$, service rate vector $\mu = (1.1,\ 0.8,\ 1.1,\ 0.3,\ 0.4,\ 0.2,\ 0.4,\ 0.3,\ 0.2,\ 0.4,\ 0.1)$ and routing matrix

$$\Theta = \begin{pmatrix} 0.0 & 0.3 & 0.3 & 0.4 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 \\ 0.7 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.8 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}.$$

From the source S_0, the nodes S_1, S_2 and S_3 may be reached. After being served in these systems, the customers may return to the source S_0 (i.e., leave the network) with a probability of 0.5, and with small probabilities equal to 0.1, may move to nodes S_4–S_8. After service is completed in S_4–S_8, the customers may go to S_0 with a probability of 0.7 or to one of the nodes S_9–S_{11} with probabilities equal to 0.1. Finally, customers from S_9–S_{11} leave the network with sufficiently high probabilities or move to other nodes with the same probability of 0.1.

Links and route probabilities were chosen in such a way that the developed analysis method can be applied. Indeed, small and comparable (in our case, equal) transition probabilities permit one to approximate the input flow, coming to any queue from a large number of nodes, close to a Poisson flow. Moreover, backwards paths (from S_9–S_{11}) have been introduced to verify the applicability of the proposed methodology also to queuing networks with a non-tree structure.

Using the optimization algorithm proposed in this paper, the optimal batch size vector is $b = (2,\ 2,\ 2,\ 2,\ 1,\ 2,\ 2,\ 2,\ 1,\ 1,\ 3)$. This choice leads to an average response time equal

to 7.61, and a mean number of customers in the network equal to 15.22 (in good agreement with the simulation estimates of 7.64 and 15.28, respectively). Here and in the following, all the parameters, derived by discrete-event simulation, were calculated in stationary conditions with a confidence interval of 0.01 and a confidence level higher than 0.95.

For the sake of completeness, Tables 1 and 2, respectively, show the average number of customers and the average sojourn time in each node of the network. Note that all the analytical values and the corresponding estimates by discrete-event simulation are very close, confirming the applicability of the analytical approach proposed in this paper not only for the whole queue, but also at the node level. In more detail, the largest difference for the average number of customers, observed at node S_6, is less than 5%, whereas for the average sojourn time the maximum was attained at nodes S_9 and S_{10}, and it was equal to 2%.

Table 1. Average numbers of customers in the nodes.

Node	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
Approximation	1.174	1.554	1.493	1.447	1.065	2.118	1.139	1.381	1.111	0.357	2.289
Simulation	1.175	1.553	1.494	1.451	1.076	2.221	1.145	1.388	1.131	0.365	2.283

Table 2. Average sojourn times in the nodes.

Node	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
Approximation	1.890	2.501	1.818	6.672	5.163	10.750	5.254	6.694	10.556	3.393	21.647
Simulation	1.890	2.501	1.820	6.691	5.219	10.769	5.280	6.720	10.735	3.468	21.684

Moreover, Table 2 highlights that the average sojourn times in nodes S_6, S_9 and S_{11} significantly exceed the values in the other nodes of the network. Since in our example the transition probabilities are comparable, this is due to the service rates, which in these nodes are lower than in the rest of the network.

7. Conclusions

In this paper we proposed an efficient method for the analysis of large-scale open queuing networks with batch services and an original algorithm to determine the optimal vector of the service batch sizes, which minimizes the average response time of the network. The proposed algorithm can be used to determine the optimal capacity of vehicles for various purposes, such as the transportation of passengers or goods. The algorithm is also applicable for optimizing systems for the accumulation and subsequent processing of customers requests, which may include information in electronic or paper forms, financial resources, materials, parts and production waste. In more detail, due to the conditions under which the product-form was obtained, the proposed algorithm is applicable to large-scale networks with individual routing of the customers, assuming that the number of possible destinations is significantly larger than the batch size.

Finally, it is worth mentioning that the memory requirement of the algorithm is $\mathcal{O}(1)$, and its computational complexity is $\mathcal{O}(n)$, where n is the number of the values of the $\bar{u}$ function that must be calculated (in general, n depends on the queuing network parameters).

Author Contributions: Conceptualization, E.S., I.T. and M.P.; methodology, I.T.; software, I.T.; validation, E.S., I.T. and M.P.; formal analysis, E.S., I.T. and M.P.; investigation, E.S., I.T. and M.P.; writing—original draft preparation, E.S., I.T. and M.P.; writing—review and editing, E.S., I.T. and M.P.; visualization, E.S., I.T. and M.P.; supervision, I.T. and M.P.; project administration, E.S.; funding acquisition, E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of science and education of the Russian Federation in the framework of the basic part of the scientific research state task, project FSRR-2020-0006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Evequoz, C.; Tropper, C. Approximate analysis of bulk closed queueing networks. *Int. J. Prod. Res.* **1995**, *33*, 179–204.
2. Mitici, M.; Goseling, J.; Ommeren, J.-K.; Graaf, M.; Boucherie, R.J. On a tandem queue with batch service and its applications in wireless sensor networks. *Queueing Syst.* **2017**, *87*, 81–93. [CrossRef]
3. Rabta, B.; Reiner, G. Batch sizes optimisation by means of queueing network decomposition and genetic algorithm. *Int. J. Prod. Res.* **2012**, *50*, 2720–2731. [CrossRef]
4. Yu, A.-L.; Zhang, H.-Y.; Chen, Q.-X.; Mao, N.; Xi, S.-H. Buffer allocation in a flow shop with capacitated batch transports. *J. Oper. Res. Soc.* **2021**, *73*, 888–904. [CrossRef]
5. Hopp, W.J.; Spearman, M.L.; Chayet, S.; Donohue, K.L.; Gel, E.S. Using an optimized queueing network model to support wafer fab design. *IIE Trans.* **2002**, *34*, 119–130. [CrossRef]
6. Hanschke, T; Zisgen, H. Queueing networks with batch service. *Eur. J. Ind. Eng.* **2011**, *5*, 313–326. [CrossRef]
7. Antunes, N.; Nunes, C.; Pacheco A. A queueing network with Markov modulated input. In Proceedings of the Second Euro-Japanese Workshop on Stochastic Risk Modelling for Finance, Chamonix, France, 18–20 September 2000; pp. 15–24.
8. Kar, S.; Rehrmann, R.; Mukhopadhyay, A.; Alt, B.; Ciucu, F.; Koeppl, H.; Binnig C.; Rizk A. On the throughput optimization in large-scale batch-processing systems. *Perform. Eval.* **2020**, *144*, 102142. [CrossRef]
9. Grippa, P.; Schilcher, U.; Bettstetter, C. On access control in cabin-based transport systems. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 2149–2156. [CrossRef]
10. Beiley, N.T.J. On Queueing Processes with Bulk Service. *J. R. Stat. Soc.* **1954**, *16*, 80–87. [CrossRef]
11. Downton, F. Waiting time in bulk service queues. *J. R. Stat. Soc.* **1955**, *17*, 256–261. [CrossRef]
12. Henderson, W.; Pearce, C.E.M.; Taylor, P.G.; Dijk, N.M. Closed Queueing Networks with Batch Services . *Queueing Syst.* **1990**, *6*, 59–70. [CrossRef]
13. Sasikala, S.: Indhira, K. Bulk Service Queueing Models—A Survey. *Int. J. Pure Appl. Math.* **2016**, *106*, 43–56.
14. Chaudhry, M.L.;Templeton, J.G.C. *A First Course in Bulk Queues*; John Wiley & Sons Inc.: New York, NY, USA, 1983.
15. Stankevich, E.P.; Tananko, I.E.; Dolgov, V.I. Analysis of closed queueing networks with batch service. *Izv. Saratov Univ. (N. S.) Ser. Math. Mech. Inform.* **2020**, *20*, 527–533. [CrossRef]
16. Klunder, W. Decomposition of Open Queueing Networks with Batch Service. In *Operations Research Proceedings 2016*; Springer: Cham, Switzerland, 2018; pp. 575–581.
17. Daduna, H. *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*; Springer: New York, NY, USA, 2001.
18. Chao, X.; Pinedo, M.; Shaw, D. Networks of Queues with Batch Services and Customer Coalescence. *J. Appl. Probab.* **1996**, *33*, 858–869. [CrossRef]
19. Harrison, P.G. Product-form queueing networks with batches. In *Computer Performance Engineering, Proceedings of the 15th European Workshop, EPEW 2018, Paris, France, 29–30 October 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; pp. 250–264.
20. Harrison, P.G.; Bor, J. Response time distribution in a tandem pair of queues with batch processing. *J. ACM* **2021**, *68*, 1–41. [CrossRef]
21. Miyazawa, M.; Taylor, P.G. A Geometric Product-Form Distribution for a Queueing Network with Non-Standard Batch Arrivals And Batch Transfers. *Adv. Appl. Prob.* **1997**, *29*, 523–544. [CrossRef]
22. Chao, X. Partial Balances in Batch Arrival Batch Service and Assemble-Transfer Queueing Networks. *Adv. Appl. Prob.* **1997**, *34*, 745–752. [CrossRef]
23. Neely, M.J. *Stochastic Network Optimization with Application to Communication and Queueing Systems*; Morgan & Claypool: San Rafael, CA, USA, 2010.
24. Xia, C.H.; Michailidis, G.; Bambos, N.; Glynn, P.W. Optimal control of parallel queues with batch service. *Probab. Eng. Inform. Soc.* **2002**, *16*, 289–307. [CrossRef]
25. Zeng, Y.; Xia, C.H. Optimal bulking threshold of batch service queues. *J. Appl. Prob.* **2017**, *54*, 409–423. [CrossRef]
26. Stankevich, E.; Tananko, I.; Pagano, M. Analysis of Open Queueing Networks with Batch Services. In *Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Proceedings of the 20th International Conference, ITMM 2021, Named after A.F. Terpugov, Tomsk, Russia, 1–5 December 2021*; Communications in Computer and Information Science; Springer: Cham, Switzerland, 2022; pp. 40–51.
27. Nakamura, A.; Phung-Duc, T. Stationary Analysis of Infinite Server Queue with Batch Service. In *Performance Engineering and Stochastic Modeling, Proceedings of the 17th European Workshop, EPEW 2021, and 26th International Conference, ASMTA 2021, Virtual Event, 9–10 December and 13–14 December 2021*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; pp. 411–424.