mathematics

MDPI

Article

Improved Bounds on the Rate of Convergence to Stationarity of the $M/M/N$ Queue as Both the Number of Servers and the Arrival Rate Grow Large

Alexander Zeifman [1,2,*], Yacov Satin [1] and Rostislav Razumchik [2]

[1] Department of Applied Mathematics, Vologda State University, 160000 Vologda, Russia; yacovi@mail.ru
[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., 119133 Moscow, Russia; rrazumchik@frccsc.ru
* Correspondence: a_zeifman@mail.ru

Abstract

Consideration is given to the stable $M/M/N/\infty$ queueing system as both the arrival rate and the number of servers tend to infinity, including the Halflin–Whitt regime. One is interested in the asymptotic behavior of the rate of convergence to stationarity of the total number of customers in the system. Bounds for the rate of convergence are provided, which improve upon those available in the literature. The derivations, simpler than those available in the literature, are based on the well-known logarithmic norm method. It is also shown that the obtained results can be adopted to $M/M/N/N+R$ queues as well.

Keywords: continuous-time Markov chains; birth–death process; $M/M/N$ queue; ergodicity bounds; logarithmic norm

MSC: 60J27; 60J28; 60K25

1. Introduction

In this paper, we revisit the problem of bounding the rate of convergence (to stationarity) of the birth–death processes, describing the total number of customers in the classic $M/M/N/\infty$ queueing system, as the number of servers tends to infinity. Thus, we allow the arrival rate λ to be a function of N, i.e., $\lambda = \lambda(N)$, whereas the service rate μ remains a constant and the (ergodicity) condition $\lambda(N) < N\mu$ is assumed to hold.

The problem of bounding the rate of convergence to stationarity of ergodic birth–death processes is well-known. In particular, for $M/M/N/\infty$ queues (and for closely related $M/M/N/N+R$ queues), many results have been obtained during the past decades; see, for example, [1–12]. The analysis (in particular, in the Halflin–Whitt regime as $N \to \infty$) of the asymptotic behavior of the rate of convergence has come to be very fruitful (see [2,6–8,13]).

One of the relatively general techniques to tackle the considered problem is known as the logarithmic norm method. As shown in [10,14–16], it allows one to obtain both the bounds and, sometimes, exact values of the rate of convergence. In this paper, we use this method to improve some of the bounds for the rate of convergence (of $M/M/N/\infty$ as $N \to \infty$) reported in the literature.

In Section 2, we review the logarithmic norm method and state (without proofs) the necessary representations and bounds for the rate of convergence. In Section 3, we switch to the asymptotic regime and obtain the main theorem (see Theorem 1), concerning the upper and lower bounds for the rate of convergence. Section 4 is devoted to the Halflin–Whitt

Check for updates

Academic Editors: Antonio Di Crescenzo and Virginia Giorno

Received: 3 October 2026
Revised: 25 October 2026
Accepted: 26 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

regime. Here, in Theorem 2, one can find the new bounds for the rate of convergence. In Section 5, we show an example, demonstrating that the obtained results can be adopted to other queues as well.

2. Preliminaries

The common approach for obtaining the upper bounds of convergence (in the l_1-norm) to the limiting regime (and perturbation bounds) of homogeneous as well as inhomogeneous BDPs can be found in [17]. For the purpose of this paper, however, we need only its special case, particularly suited for homogeneous BDPs. To make the exposition clearer, we concentrate on the particular BDP $X(t)$, which describes the total number of customers in the standard $M/M/N$ queue with the arrival rate λ and service rate μ. Clearly, the state space of $X(t)$ is $\mathcal{X} = \{0, 1, ...\}$ and its birth and death rates are λ and $\mu_k = \min(k, N)\mu$, $k = 1, 2, ...$ Let us denote the state probability by $p_i(t) = \Pr\{X(t) = i\}$, and the vector of state probabilities by $p(t) = (p_0(t), p_1(t), ...)^T$. It is well-known that the probabilities $p(t)$ satisfy the forward Kolmogorov system

$$\frac{d}{dt}\mathbf{p}(t) = A(t)\mathbf{p}(t),\ t \geq 0, \tag{1}$$

where the initial distribution $\mathbf{p}(0)$ is assumed to be given and A is the transposed infinitesimal matrix of $X(t)$, i.e.,

$$A = \begin{pmatrix} -\lambda_0 & \mu_1 & 0 & \cdots & 0 & \cdots & \cdots \\ \lambda_0 & -(\lambda_1+\mu_1) & \mu_2 & \cdots & 0 & \cdots & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \\ 0 & \cdots & \cdots & \lambda_{r-1} & -(\lambda_r+\mu_r) & \mu_{r+1} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Let $\|\cdot\|$ be the usual l_1-norm (Manhattan distance); i.e., for the vector $p(t)$ and the matrix A, it holds that $\|p(t)\| = \sum_i |p_i(t)|$ and $\|A\| = \sup_j \sum_i |(A)_{ij}|$. Denote by Ω the set of all l_1 vectors with non-negative coordinates and a unit norm. One can consider (1) as a linear differential equation with a bounded operator in the space l_1, and apply to its analysis the theory developed in [18]. Specifically, consider the Cauchy matrix $U(t, s) = e^{A(t-s)}$. If the vector $p(t)$ is a solution to the Equation (1), then the set Ω is invariant: if $p(\tau) \in \Omega$, then $p(t) \in \Omega$ for all $t \geq \tau$. Note that the transition probability matrix $P(s, t)$ for the Markov chain $X(t)$ is related to the Cauchy matrix by the equality $P(s, t) = U^T(t, s)$.

Let us consider the system of ordinary differential equations

$$\frac{d}{dt}x(t) = Bx(t), \tag{2}$$

in the space of sequences l_1 with a bounded operator B, which is given by the matrix $B = (b_{ij})$. Let $V(t, s) = e^{B(t-s)}$ be its Cauchy operator. The main tool for the analysis is the so-called logarithmic norm $\gamma(B)$ of the operator function B, which is introduced in one of the following equivalent ways:

— $\gamma(B) = \lim_{h\to+0} \frac{\|I+hB(t)\|-1}{h}$;
— $\gamma(B) = \lim_{h\to+0} \frac{\|V(t+h,t)\|-1}{h}, t \geq 0.$

It can be verified that the explicit formula for the logarithmic norm in the space l_1 is

$$\gamma(B) = \sup_{j\in\mathcal{X}} \left(b_{jj} + \sum_{i\neq j} |b_{ij}| \right). \tag{3}$$

The usefulness of the logarithmic norm in the analysis of Markov chains stems from the fact that

$$\|V(t, s)\| \leq e^{(t-s)\gamma(B)}. \tag{4}$$

Let the matrix B in (2) be essentially non-negative; i.e., all its off-diagonal elements are non-negative. For such a matrix B, if $\mathbf{x}(s) \geq 0$, then $\mathbf{x}(t) \geq 0$ for any $t \geq s$. Denote $\sup_{j\in\mathcal{X}} \sum_i b_{ij} = -h^*$ and by $= \inf_{j\in\mathcal{X}} \sum_i b_{ij} = -h_*$. Then, $\gamma(B) = -h^*$ and

$$\|x(t)\| \leq e^{-h^* t}\|x(0)\| \tag{5}$$

for any $t \geq 0$ and any (not necessarily non-negative) initial condition $x(0)$, and

$$\|\mathbf{x}(t)\| \geq e^{-h_* t}\|\mathbf{x}(0)\|. \tag{6}$$

for any non-negative initial condition $x(0)$.

Whenever all the off-diagonal entries of the matrix B are non-negative, it is often possible to obtain an exact upper bound for the rate of convergence as follows. The concept of a logarithmic norm for a finite matrix was first introduced independently by Lozinsky and Dahlquist (in the late 1950s), and for the general case, see [18,19]. The opportunity to apply the logarithmic norm to the study of Markov chains was indicated in the paper [20]. This approach was developed and implemented in [10,14–16].

The approach is put into practice in several steps. Firstly one removes from system (1) the state (0). In order to do this, one replaces $p_0(t)$ with $1 - \sum_{i=1}^{\infty} p_i(t)$. Then from (1) one obtains the following equation:

$$\frac{d}{dt}\mathbf{z}(t) = B\mathbf{z}(t) + \mathbf{f}, \tag{7}$$

where $\mathbf{f} = (\lambda_0, 0, \ldots)^T$,

$$B = \begin{pmatrix} -(\lambda_0+\lambda_1+\mu_1) & \mu_2-\lambda_0 & -\lambda_0 & \cdots & -\lambda_0 & \cdots & \cdots \\ \lambda_1 & -(\lambda_2+\mu_2) & \mu_3 & \cdots & 0 & \cdots & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \\ 0 & \cdots & \cdots & \lambda_{r-1} & -(\lambda_r+\mu_r) & \mu_{r+1} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{8}$$

Consider a sequence of positive numbers $d_1, d_2, d_3, \ldots$, and an upper triangular matrix D of the form

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ & \ddots & \ddots & \ddots \end{pmatrix}. \tag{9}$$

Let us introduce the weighted space $l_{1D} = \{z = (p_1, p_2, \cdots)^T : \|z\|_{1D} = \|Dz\| < \infty\}$. Then $\|B\|_{1D} = \|DBD^{-1}\|$. For brevity put $B^{**} = DBD^{-1}$. The explicit form of this matrix is

$$B^{**} = \begin{pmatrix} -(\lambda_0+\mu_1) & \frac{d_1}{d_2}\mu_1 & 0 & \cdots & 0 & \cdots & \cdots \\ \frac{d_2}{d_1}\lambda_1 & -(\lambda_1+\mu_2) & \frac{d_2}{d_3}\mu_2 & \cdots & 0 & \cdots & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \\ 0 & \cdots & \cdots & \frac{d_r}{d_{r-1}}\lambda_{r-1} & -(\lambda_{r-1}+\mu_r) & \frac{d_r}{d_{r+1}}\mu_r & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{10}$$

Since the vectors z^* and z^{**} do not contain the probability p_0, we define $\|p^* - p^{**}\|_{1D}$ by $\|p^* - p^{**}\|_{1D} = \|z^* - z^{**}\|_{1D}$. Note that this matrix is essentially non-negative. Denote by $-\alpha_k$ the sum of all elements in the kth column of B^{**}, i.e.,

$$\alpha_k = \lambda_k + \mu_{k+1} - \frac{d_{k+2}}{d_{k+1}}\lambda_{k+1} - \frac{d_k}{d_{k+1}}\mu_k,\ k \geq 0,\ \mu_0 = 0.$$

Then $h^* = \sup_j \sum_i (-b_{ij}) = \inf_j \alpha_j$, $h_* = \inf_j \sum_i (-b_{ij}) = \sup_j \alpha_j$. From the inequalities (5) and (6), we obtain:

$$\|p(t) - \pi\|_{1D} \le e^{-h^* t} \|p(0) - \pi\|_{1D}, \tag{11}$$

for any $t \ge 0$ and any initial condition $p(0)$, and

$$\|p(t) - \pi\|_{1D} \ge e^{-h_* t} \|p(0) - \pi\|_{1D}. \tag{12}$$

for any initial condition $p(0)$ such that $p(0) - \pi$ is non-negative.

Note that the decay parameter (or the spectral gap) β_0 was determined in [10] for a finite Markov chain by the condition

$$\beta_0 := \sup\{a > 0 : \|\mathbf{p}(t) - \pi\| = O(e^{-at}) \quad \text{as} \quad t \to \infty \quad \text{for any} \quad \mathbf{p}(0)\}. \tag{13}$$

The general definition of the decay parameter (see, for example, [9]) is introduced as

$$\beta_0^* := \sup\{a > 0 : p_{ij}(t) - \pi_j = O(e^{-at}) \quad \text{as} \quad t \to \infty \quad \text{for any} \quad i, j\}, \tag{14}$$

It should also be noted that considering the quantity $p_{ij}(t)$ entails $X(0) = i$, from which it follows that $\mathbf{p}(0) \in l_{1D}$. Hence $\beta_0^* = \beta_0$.

In case of the finite state space $\mathcal{X}$, when all norms are equivalent, for any sequence $\{d_i, i \ge 1\}$, we have that

$$\beta_0 \in [h^*, h_*]. \tag{15}$$

As shown in [10], there exists a unique sequence $\{d_i, i \ge 1\}$, for which $h_* = h^* = \beta_0$. In the case of the countable state space $\mathcal{X}$, such a sequence also exists and is unique (see [16]).

In [14], it was shown that the bound

$$\beta_0 \ge h^*, \tag{16}$$

is valid for any sequence $\{d_i, i \ge 1\}$, such that $\inf_{i \ge 1} d_i = d > 0$.

3. General Bounds

In what follows, we assume that the stability condition $\lambda(N) < N\mu$ holds, but the value of $\lambda(N)$ is chosen such that $\lambda(N) > \mu$. The sequence $\{d_n, n \ge 0\}$, introduced in Section 2, is defined by $d_0 = d_1 = 1$, $d_{k+1} = \delta_k d_k$. Put

$$\delta_k = \begin{cases} \sqrt{\frac{k\mu}{\lambda}}, & 1 \le k \le N, \\ \sqrt{\frac{N\mu}{\lambda}}, & k > N, \end{cases}$$

and let us study the properties of α_k, considering it as a function of k. By plugging δ_k into the expression for α_k, we get:

$$\alpha_k = \begin{cases} \lambda + (k+1)\mu - \sqrt{(k+1)\lambda\mu} - \sqrt{k\lambda\mu}, & 0 \le k < N, \\ \left(\sqrt{\lambda} - \sqrt{N\mu}\right)^2, & k \ge N. \end{cases} \tag{17}$$

Hence $h^* = \inf\limits_{k \in \{0,1,\dots\}} \alpha_k = \min(\alpha_0, \alpha_1, \dots, \alpha_N)$. Consider the continuous real-valued function $\phi(x) = \lambda + (x+1)\mu - \sqrt{(x+1)\lambda\mu} - \sqrt{x\lambda\mu}$, $x \in [0, N-1]$. Clearly, $\phi(x) = \alpha_x$ whenever x is an integer. Its derivative for $x \in (0, N-1)$ is equal to

$$\phi'(x) = \mu - 0.5\sqrt{\frac{\lambda\mu}{x+1}} - 0.5\sqrt{\frac{\lambda\mu}{x}},$$

and consequently, $\phi'(x)$ is a strictly increasing function. Consider the two points $x_1 = \frac{\lambda}{\mu} - 1$ and $x_2 = \frac{\lambda}{\mu}$. Due to the assumption $\mu < \lambda < N\mu$, we have that $0 < x_1 < x_2 < N$, and since

$$\phi'(x_1) = \mu - 0.5\sqrt{\frac{\lambda\mu}{\frac{\lambda}{\mu} - 1 + 1}} - 0.5\sqrt{\frac{\lambda\mu}{\frac{\lambda}{\mu} - 1}} = 0.5\mu - 0.5\mu\sqrt{\frac{\lambda}{\lambda - \mu}} < 0,$$

$$\phi'(x_2) = \mu - 0.5\sqrt{\frac{\lambda\mu}{\frac{\lambda}{\mu} + 1}} - 0.5\sqrt{\frac{\lambda\mu}{\frac{\lambda}{\mu}}} = 0.5\mu - 0.5\mu\sqrt{\frac{\lambda}{\lambda + \mu}} > 0,$$

the function $\phi(x)$ attains its minimum at a point $x \in (\lambda/\mu - 1, \lambda/\mu)$. Consequently the value of $h^* = \min(\alpha_0, \alpha_1, \dots, \alpha_N)$ is equal either to $\alpha_{\lfloor \lambda/\mu - 1 \rfloor}$ or $\alpha_{\lfloor \lambda/\mu \rfloor}$, or α_N.

In addition to the assumptions made at the beginning of Section 3, assume that $\lambda(N)$ is such that $\lambda(N) \to \infty$ as $N \to \infty$ (but μ is a constant). It is well-known (see [21]) that the rate of convergence β_0 satisfies the inequality

$$\beta_0 \le \left(\sqrt{\lambda} - \sqrt{N\mu}\right)^2 = \alpha_N = \beta^{**}. \tag{18}$$

As was shown above, the function α_k attains its minimum at a point $k \in \{\lfloor \lambda/\mu - 1 \rfloor, \lfloor \lambda/\mu \rfloor, N\}$. Let us derive the bound for the minimal value. For *any* $x^* = x_1 + \eta = \lambda/\mu - 1 + \eta$, where $0 \le \eta \le 1$, we have

$$\phi(x^*) = \lambda + \left(\frac{\lambda}{\mu} + \eta\right)\mu - \sqrt{\left(\frac{\lambda}{\mu} + \eta\right)\lambda\mu} - \sqrt{\left(\frac{\lambda}{\mu} - (1 - \eta)\right)\lambda\mu} =$$

$$= 2\lambda + \eta\mu - \sqrt{\lambda^2 + \eta\lambda\mu} - \sqrt{\lambda^2 - (1 - \eta)\lambda\mu} =$$

$$= \frac{\eta\lambda\mu + \eta^2\mu^2}{\lambda + \eta\mu + \sqrt{\lambda^2 + \eta\lambda\mu}} + \frac{(1 - \eta)\lambda\mu}{\lambda + \sqrt{\lambda^2 - (1 - \eta)\lambda\mu}},$$

and since we have assumed that $\lambda = \lambda(N) \to \infty$ as $N \to \infty$, it holds that

$$\left|\phi(x^*) - \left(\frac{\eta}{2} + \frac{1 - \eta}{2}\right)\right|\mu \le \frac{\mu}{\lambda},$$

and

$$\lim_{N\to\infty} \phi(x^*) = \left(\frac{\eta}{2} + \frac{1 - \eta}{2}\right)\mu = \frac{\mu}{2}.$$

Hence, since the minimal value of α_k when $k \in \{0, 1, \dots, N - 1\}$ can be considered as $\phi(x^*)$, we have $\min_{k\in\{0,1,\dots,N-1\}} \alpha_k = \frac{\mu}{2} + o(1)$. Therefore, the following theorem holds.

Theorem 1. *Assume $\lambda = \lambda(N) \to \infty$ as $N \to \infty$, and μ is a constant. Then the rate of convergence for $M/M/N$ as $N \to \infty$ satisfies the double inequality*

$$\min\left(\frac{\mu}{2}, \left(\sqrt{\lambda} - \sqrt{N\mu}\right)^2\right) + o(1) \le \beta_0 \le \left(\sqrt{\lambda} - \sqrt{N\mu}\right)^2. \tag{19}$$

4. Halfin–Whitt Regime

Consider the case, when $\lambda(N) = N - N^{1-a}$, and $a \in (0, 1)$, $\mu = 1$. Usually, one says that the Halfin–Whitt regime holds if $a = 0.5$; the case $a < 0.5$ corresponds to the sub-Halfin–Whitt regime, and the case $a > 0.5$ corresponds to the super-Halfin–Whitt regime (see, for example, [2,6,7,22]). From Theorem 1, one immediately has the following corollary.

Corollary 1. *The rate of convergence for $M/M/N$ in Halfin–Whitt regimes satisfies as $N \to \infty$ the double inequality*

$$\min\left(\frac{1}{2}, \left(\sqrt{\lambda} - \sqrt{N}\right)^2\right) + o(1) \leq \beta_0 \leq \left(\sqrt{\lambda} - \sqrt{N}\right)^2. \tag{20}$$

These inequalities can be made sharper. After plugging $\lambda(N) = N - N^{1-a}$ in the rightmost expression, we get:

$$\left(\sqrt{\lambda} - \sqrt{N}\right)^2 = \left(\sqrt{N - N^{1-a}} - \sqrt{N}\right)^2 = \left(\frac{N^{1-a}}{\sqrt{N - N^{1-a}} + \sqrt{N}}\right)^2 = \beta^{**}(N).$$

Consider the three exhaustive cases for the value of a:

— when $a = 0.5$, we have

$$\lim_{N\to\infty} \beta^{**}(N) = \left(\frac{N^{0.5}}{\sqrt{N - N^{1-a}} + \sqrt{N}}\right)^2 = \frac{1}{4};$$

— when $a > 0.5$, it holds that

$$\lim_{N\to\infty} \beta^{**}(N) = \left(\frac{N^{1-a}}{\sqrt{N - N^{1-a}} + \sqrt{N}}\right)^2 = 0;$$

— when $a < 0.5$, we have

$$\lim_{N\to\infty} \beta^{**}(N) = \left(\frac{N^{1-a}}{\sqrt{N - N^{1-a}} + \sqrt{N}}\right)^2 = \infty.$$

In the case where $a > 0.5$, for sufficiently large values of N, it holds that $\alpha_k \geq \frac{1}{2} + o(1)$ for any $k \leq N - 1$.

Therefore, from Theorem 1, one obtains another statement about the asymptotics of the decay parameter in the Halflin–Whitt regime.

Theorem 2. *The decay parameter β_0 of the $M/M/N$ queue in the Halflin–Whitt regimes satisfies as $N \to \infty$ the following inequalities:*

(i) $\beta_0 \geq 0.5 + o(1)$, when $a < 0.5$;

(ii) $\beta_0 = 0.25 + o(1)$, when $a = 0.5$;

(iii) $\beta_0 = \left(\frac{N^{1-a}}{\sqrt{N-N^{1-a}}+\sqrt{N}}\right)^2 = o(1)$, when $a > 0.5$.

Theorem 2 provides a simpler proof for the case $a = 0.5$ (see [2]) and better estimates for the sub-Halflin–Whitt and the super-Halflin–Whitt regimes (see [7] for the known so far bounds for the convergence rates in different metrics for $M/M/N$ queues in the Halflin–Whitt regime). The graphs below (Figures 1–5) show the values of h^* (the lower bound for β_0) as a function of N for various cases of a. These results are consistent with the asymptotic result of Theorem 2.

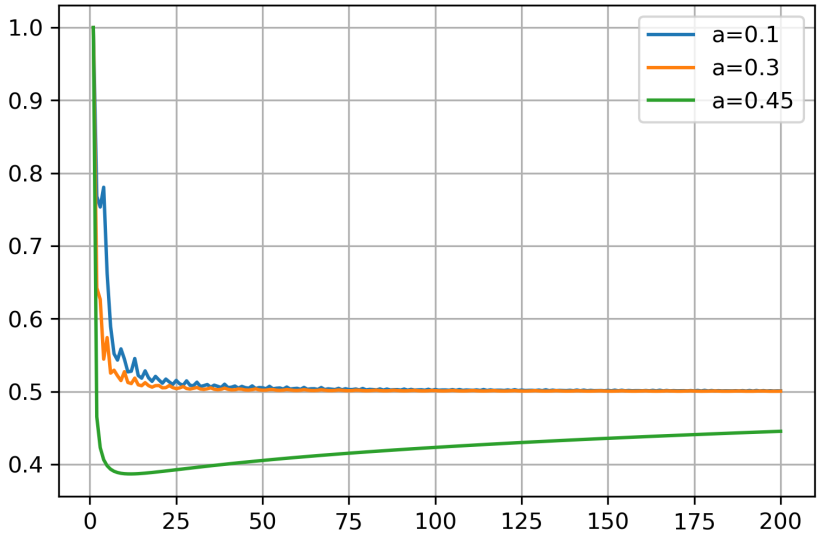

Figure 1. Behavior of h^* for various $N \in \{1, \ldots, 200\}$ and a within the sub-Halfin–Whitt regime.

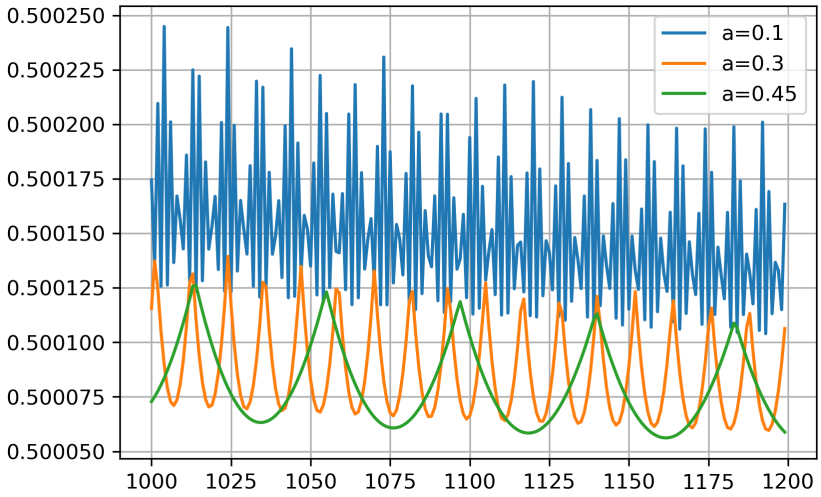

Figure 2. Behavior of h^* for various $N \in \{1000, \ldots, 1200\}$ and a within the sub-Halfin–Whitt regime.

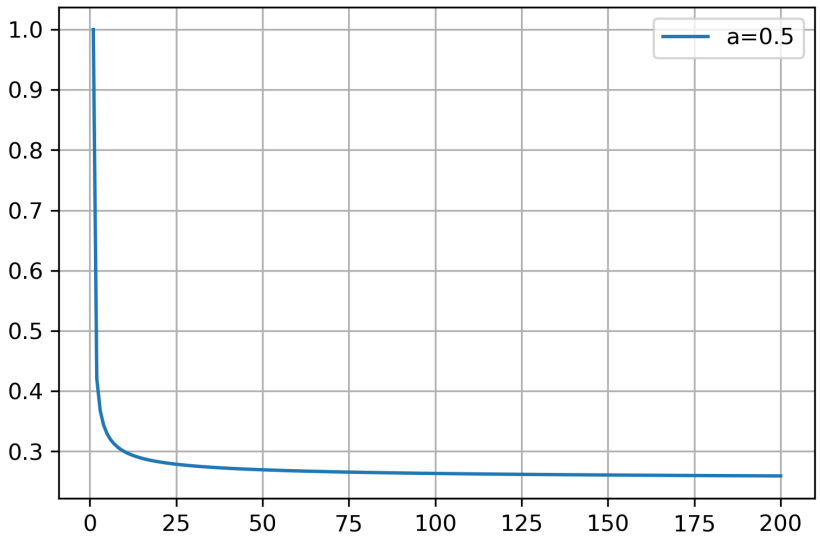

Figure 3. Behavior of h^* for various $N \in \{1, \ldots, 200\}$ in the Halfin–Whitt regime.

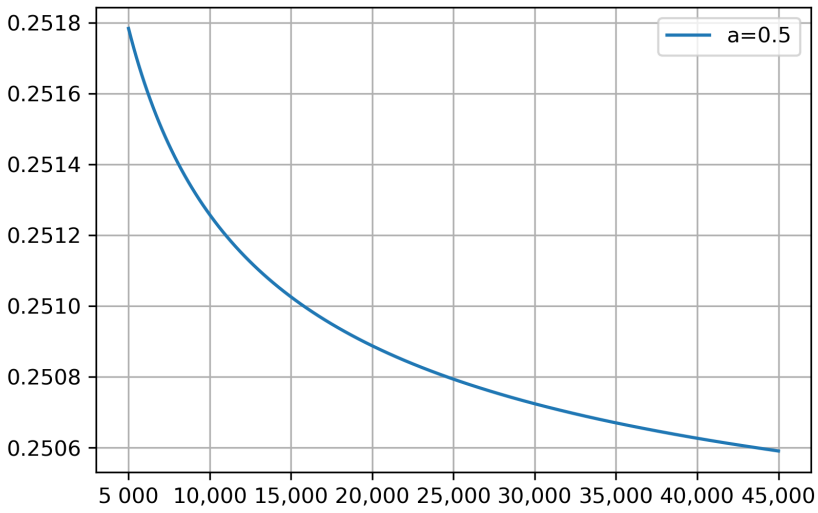

Figure 4. Behavior of h^* for various $N \in \{5000, \ldots, 45{,}000\}$ in the Halfin–Whitt regime.

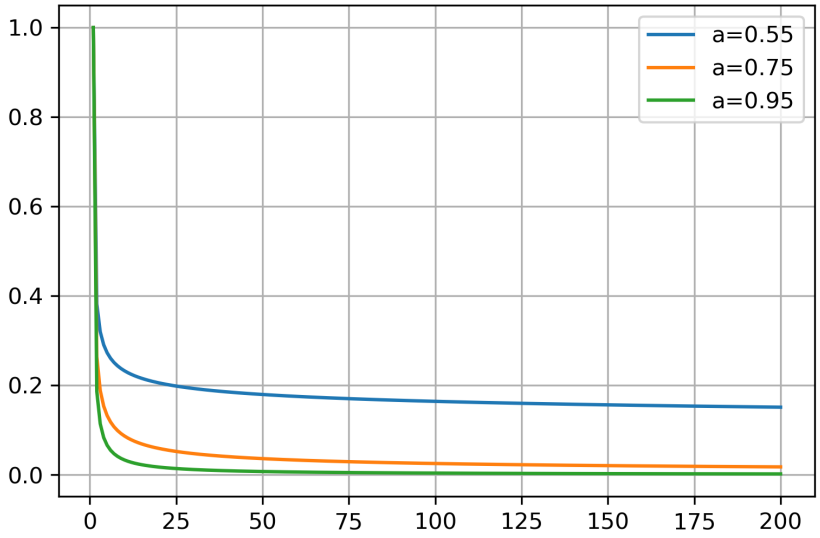

Figure 5. Behavior of h^* for various $N \in \{1, \ldots, 200\}$ and a within the super-Halfin–Whitt regime.

5. Conclusions

The bounds for the decay rate obtained in Theorem 2 can be adopted to other queues as well. Consider, for example, the queue $M/M/N/N+R$ (extensively studied in [10,11,23–26]). As shown in [17], the matrix B^{**} defined by (10) has the form

$$B^{**}=\begin{pmatrix} -(\lambda_0+\mu_1) & \frac{d_1}{d_2}\mu_1 & 0 & \cdots & 0 \\ \frac{d_2}{d_1}\lambda_1 & -(\lambda_1+\mu_2) & \frac{d_2}{d_3}\mu_2 & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & \frac{d_r}{d_{r-1}}\lambda_{r-1} & -(\lambda_{r-1}+\mu_r) \end{pmatrix}.$$

Therefore, all the values of α_k, defined by (17), remain unchanged, except for the last one α_{N+R-1}, which is equal to a larger value $\alpha_{N+R-1} = \lambda + N\mu - \sqrt{N\lambda\mu}$, if $R > 0$, and $\alpha_{N-1} = \lambda + N\mu - \sqrt{(N-1)\lambda\mu}$, if $R = 0$. Thus, Theorem 2 gives the bounds for the decay parameter (as $N \to \infty$ in the Halfin–Whitt regimes). It is important to note that these bounds do not depend on R. Moreover, since the state space of the Markov chain is finite, we have

$$\|\mathbf{p}(t) - \pi\|_1 \leq 8(d_1 + \cdots + d_{N+R})e^{-\beta_0 t},\ t > 0.$$

Author Contributions: Conceptualization, supervision, A.Z.; methodology, Y.S.; software, validation, visualization, A.Z.; investigation, Writing—original draft, R.R., Y.S. and A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Ministry of Science and Higher Education of the Russian Federation under project No. 075-15-2024-544.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Banerjee, S.; Mukherjee, D. Join-the-shortest queue diffusion limit in halfin–whitt regime. *Ann. Appl. Probab.* **2020**, *30*, 80–144. [CrossRef]
2. Gamarnik, D.; Goldberg, D.A. On the rate of convergence to stationarity of the M/M/N queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* **2013**, *23*, 1879–1912. [CrossRef]
3. Kijima, M. Evaluation of the decay parameter for some specialized birth–death processes. *J. Appl. Probab.* **1992**, *29*, 781–791. [CrossRef]
4. Kijima, M. *Markov Processes for Stochastic Modeling*; CRC Press: Boca Raton, FL, USA, 1997; Volume 6.
5. Liu, X.; Ying, L. Steady-state analysis of load-balancing algorithms in the sub-Halfin–Whitt regime. *J. Appl. Probab.* **2020**, *57*, 578–596. [CrossRef]
6. Nguyen, H.H.; Varma, S.M.; Maguluri, S.T. Finite-Time Behavior of Erlang-C Model: Mixing Time, Mean Queue Length and Tail Bounds. *ACM SIGMETRICS Perform. Eval. Rev.* **2025**, *53*, 4–6. [CrossRef]
7. Nguyen, H.; Varma, S.; Theja Maguluri, S. Finite-Time behaviors of M/M/n Queue: Mixing Time, Mean Queue Length and Tail Bounds. *arXiv* **2025**, arXiv:2504.02207.
8. Nguyen, H.H. Mixing time of Queuing Systems: From Erlang-C to Join-the-Shortest-Queue. *ACM SIGMETRICS Perform. Eval. Rev.* **2026**, *53*, 6–8. [CrossRef]
9. Van Doorn, E. Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Adv. Appl. Probab.* **1985**, *17*, 514–530. [CrossRef]
10. Van Doorn, E.A.; Zeifman, A.I.; Panfilova, T.L. Bounds and asymptotics for the rate of convergence of birth-death processes. *Theory Probab. Its Appl.* **2010**, *54*, 97–113. [CrossRef]
11. van Doorn, E.A. Rate of convergence to stationarity of the system M/M/N/N+ R. *Top* **2011**, *19*, 336–350. [CrossRef]
12. Jin, X.; Pang, G.; Xu, L.; Xu, X. An Approximation to the Invariant Measure of the Limiting Diffusion of $G/Ph/n + GI$ Queues in the Halfin–Whitt Regime and Related Asymptotics. *Math. Oper. Res.* **2024**, *50*, 783–812. [CrossRef]
13. Halfin, S.; Whitt, W. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **1981**, *29*, 567–588. [CrossRef]
14. Zeifman, A.I. Some estimates of the rate of convergence for birth and death processes. *J. Appl. Probab.* **1991**, *28*, 268–277. [CrossRef]
15. Zeifman, A.I. Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. *Stoch. Proc. Appl.* **1995**, *59*, 157–173. [CrossRef]
16. Granovsky, B.L.; Zeifman, A.I. The decay function of nonhomogeneous birth-death processes, with application to mean-field models. *Stoch. Proc. Appl.* **1997**, *72*, 105–120. [CrossRef]
17. Zeifman, A.; Satin, Y.; Korolev, V.; Usov, I. Continuous-time Markov chains and models: Bounding via forward Kolmogorov system. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2025; Volume 52, pp. 179–210. [CrossRef]
18. Daleckii, J.L.; Krein, M.G. *Stability of Solutions of Differential Equations in Banach Space (No. 43)*; American Mathematical Society (AMS): Providence, RI, USA, 2002.
19. Söderlind, G. The logarithmic norm. History and modern theory. *BIT Numer. Math.* **2006**, *46*, 631–652. [CrossRef]
20. Gnedenko, B.V.; Makarov, I.P. Properties of the solutions of a problem with losses in the case of periodic intensities. *Differ. Uravn.* **1971**, *7*, 1696–1698.
21. Van Doorn, E. Stochastic Monotonicity and Queueing Applications of Birth-Death Processes. In *Lecture Notes in Statistics*; Springer: New York, NY, USA, 1981.

22. Varma, S.M.; Castro, F.; Maguluri, S.T. Power-of-d Choices Load Balancing in the Sub-Halfin-Whitt Regime. *Stoch. Syst.* **2025**, *15*, 291–336. [CrossRef]
23. van Doorn, E.A.; Zeifman, A.I. On the speed of convergence to stationarity of the Erlang loss system. *Queueing Syst.* **2009**, *63*, 241. [CrossRef]
24. Van Doorn, E.A. Representations for the decay parameter of a birth-death process based on the Courant–Fischer Theorem. *J. Appl. Probab.* **2015**, *52*, 278–289. [CrossRef]
25. Zhao, Z.; Banerjee, S.; Mukherjee, D. Many-Server Asymptotics for Join-the-Shortest-Queue in the Super-Halfin-Whitt Scaling Window. *Math. Oper. Res.* **2025**, *51*, 2073–2124. [CrossRef]
26. Zhu, J.; Liu, J.; Li, W. Stein's method for ergodic rates of stochastically monotone Markov chains. *Stat. Probab. Lett.* **2026**, *237*, 110822. [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.