Check for updates

J. Appl. Probab. 1–10 (2026)
*doi:*10.1017/jpr.2026.10130

CONVERGENCE RATE OF THE JOIN-THE-SHORTEST-QUEUE SYSTEM

YUANZHE MA,* *Columbia University*
SIVA THEJA MAGULURI,** *Georgia Institute of Technology*

Abstract

The join-the-shortest-queue (JSQ) policy is among the most widely used load-balancing algorithms and has been extensively studied. However, an exact characterization of the system behavior remains challenging. Most prior research has focused on analyzing its performance in the steady state in certain asymptotic regimes, such as the heavy-traffic regime. However, convergence to the steady state in these regimes is often slow, so steady-state and heavy-traffic characterizations may be less informative over practical time horizons. To address this limitation, we provide a finite-time convergence rate analysis of a JSQ system with two symmetric servers. In sharp contrast to the existing literature, we directly study the original system rather than an approximate limiting system such as a diffusion approximation. Our results demonstrate that for such a system, the convergence rate to its steady state, measured in the total variation distance, is $O((1/(1-\rho)^3)(1/t))$, where $\rho \in (0, 1)$ is the traffic intensity.

Keywords: Join-the-shortest-queue; load-balancing system; convergence rate analysis; coupling

2020 Mathematics Subject Classification: Primary 60K25; 68M20; 90B22
Secondary 60H99

1. Introduction

Efficient load balancing is central to optimizing resource allocation and minimizing job delays in modern service systems and data centers. The join-the-shortest-queue (JSQ) policy is a popular load-balancing algorithm due to its simplicity and strong performance under various criteria. Since its introduction in [21], it has received significant attention over the decades. However, analysis of the JSQ system remains highly nontrivial [19]. Due to the complexity of the problem, most prior works consider a JSQ system only in steady state. Further, researchers typically conduct analyses in some asymptotic regimes, such as the heavy-traffic regime. For example, there are various bounds on the queue length of a JSQ system in its steady state, such as [15, Theorem 10.2.3]. However, traffic patterns in modern data centers tend to be time-varying [2]. This necessitates a careful finite-time analysis of JSQ performance, rather

Received 27 August 2024 UTC; accepted 7 August 2026 UTC.
* Postal address: Department of Industrial Engineering and Operations Research, Columbia University, 500 W 120th St, New York, NY 10027, USA. Email: ym2865@columbia.edu
** Postal address: H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive, NW, Atlanta, GA 30332, USA. Email: siva.theja@gatech.edu

© The Author(s), 2026. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

than relying only on steady-state behavior. Motivated by this, we focus on analyzing the *transient behavior* of a JSQ system. As a starting point, we consider a system consisting of two servers with both interarrival and service times following exponential distributions. We analyze its non-asymptotic convergence rate toward its steady state.

As discussed, prior works focus on analyzing the JSQ policy under the heavy-traffic regime and in its steady state. Many prior results rely on state-space collapse phenomena in heavy traffic and do not directly yield explicit finite-time bounds for light traffic. In this paper, we address this challenge and obtain an explicit upper bound on the difference between the transient distribution of a JSQ system and its steady state for any traffic intensity $\rho \in (0, 1)$.

We focus on analyzing the distribution $\mathbb{P}^t_{\boldsymbol{x}}$ of the queue length vector at time t for a JSQ system with an initial queue length vector $\boldsymbol{x}$. In particular, we consider its total variation (TV) distance to the steady state π at every time $t \geq 0$. The total variation distance is a common metric quantifying the convergence rate of a stochastic system and is known to be closely connected to many other metrics. For example, bounds on TV distances imply bounds on Hellinger distances. In addition to the TV distance, we derive the convergence rate of the mean queue length. In our coupling approach, the key challenge is to obtain an upper bound on the expected hitting times or first-passage times of the JSQ system. For general continuous-time Markov chains (CTMCs), analyzing the hitting time usually requires extensive exploitation of the structure of the CTMC [5]. We use several structural properties of the JSQ system and show that the problem can be reduced to providing an upper bound on $\mathbb{E}[T_0]$, where T_0, to be defined in (3.1), is the time it takes for a JSQ system's queue length vector to hit the origin starting from state (1, 0).

1.1. Main contribution

Our main result is for a continuous-time JSQ system with two symmetric queues, where both interarrival and service times follow exponential distributions. We derive an upper bound on $d_{\mathrm{TV}}(\mathbb{P}^t_{\boldsymbol{x}}, \pi)$ and, as a corollary, an upper bound on the difference between the mean queue length at time t and its steady-state value. These two quantities converge to zero as $t \to \infty$ for any traffic intensity $\rho \in (0, 1)$. In particular, for the TV distance, we establish in Theorem 1 that the convergence rate is $O((1/(1-\rho)^3)(1/t))$. In contrast to prior work based on diffusion limits, we derive transient bounds by directly studying the original JSQ system, thus avoiding approximation errors. To the best of our knowledge, we are the first to provide a convergence rate analysis of a JSQ system for any $\rho \in (0, 1)$.

The rest of the paper is organized as follows. We first discuss related work in Section 1.2 and notation in Section 1.3. We present the model and state our main result in Section 2. We provide all proofs in Section 3. Finally, we provide concluding remarks and promising future directions in Section 4.

1.2. Related work

Since the pioneering work of [21], the JSQ policy has been extensively studied in the literature [3, 19]. Transient analyses are generally complicated, even for simple systems such as M/M/1 queues [1, 16] and M/M/n queues [7]. There are numerical results for transient analysis of JSQ systems [8], but few theoretical results are known regarding their convergence rates. In this work, we analyze the transient behavior through a hitting-time analysis (see, e.g., [9]). Though there are some results related to hitting times or first-passage times for JSQ systems [17, 22], they cannot be used directly for our problem. For a simple system such as an M/M/1 system, we can obtain the moment-generating function (MGF) of the hitting time

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

for the queue length to hit state 0 using a construction of an exponential martingale and the optional stopping theorem [14, Proposition 5.4]. One of the key reasons this works is the special property of an M/M/1 queue: starting from state $x \geq 1$, the queue length must first hit states $x-1, x-2, \ldots, 1$ before hitting state 0. However, this property is not true for a JSQ system. For example, consider a JSQ system with an initial queue length (x,y), where $x, y \geq 1$. It is possible that the queue length vector never hits state $(x-1, y-1)$ before hitting $(0, 0)$. This is one of the technical barriers to analyzing a JSQ system. One possible approach is to use the exponential martingale constructed in [17]; however, that work focuses on the hitting time to (K,K), $K \geq 1$, whereas our interest is in the hitting time to $(0, 0)$, which is fundamentally different, making those results inapplicable.

The convergence rates of stochastic systems and hitting-time analyses based on Foster–Lyapunov drift functions have been extensively studied in the literature; see, e.g., [6]. For example, [11, Theorem 2] provides a general framework for obtaining exponential convergence rates of general stochastic systems. However, applying it requires a suitable drift function, and it is unclear how to construct one for the JSQ system. Relatedly, researchers have discussed transient analysis for queues under various limiting regimes or for simpler systems [7, 20]. In contrast to these works, we analyze the JSQ system in a fixed system setting.

1.3. Notation

In this section we introduce the notation that will be used throughout this paper. We use $\mathbb{P}(A)$ to denote the probability that event A occurs, and use $\mathbb{E}[X]$ to denote the expected value of the random variable X. For a vector $\boldsymbol{x}$, we use x_i to denote its ith element. Given two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, we write $\boldsymbol{x} \geq \boldsymbol{y}$ if $x_i \geq y_i$ for every $i \in \{1, \ldots, n\}$. For a stochastic process $(X_t)_{t\geq 0}$ with steady state π, we use $\mathbb{E}_\pi[X]$ and $\mathbb{P}_\pi(\cdot)$ to denote the expectation and distribution of X_t with respect to the steady state. For two random variables A and B, we write that A and B are equal in distribution, or $A \overset{\mathrm{d}}{=} B$, if $\mathbb{P}(A \leq x) = \mathbb{P}(B \leq x)$ for all $x \in \mathbb{R}$. We define A to be stochastically dominated by B, or $A \preceq B$, if and only if $\mathbb{P}(A > x) \leq \mathbb{P}(B > x)$ for all $x \in \mathbb{R}$. For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ defined on the sigma-algebra $\mathcal{F}$, we denote the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ as $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A\in\mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$.

2. Model and results

2.1. Model

Throughout the paper, we consider a continuous-time load-balancing system operating under the JSQ policy. It consists of two single-server queues, and each has an infinite buffer size. Jobs arrive according to a Poisson process with rate λ, and service times for each queue are independent exponential random variables with mean $1/\mu$. We use $q_i(t)$, $i \in \{1, 2\}$, to denote the number of jobs in the ith queue at time t. Let $\boldsymbol{q}(t) \in \mathbb{Z}_+^2 = \{0, 1, 2, 3, \ldots\}^2$ be the queue length process with elements $(q_1(t), q_2(t))$. The JSQ policy sends each arriving job to the queue with the fewest jobs, and picks the one with the smaller index when there is a tie. We define $i^\star(\boldsymbol{q}) = \min\left\{\arg\min_{i\in\{1,2\}}\{q_i\}\right\}$. The queue length process $\{\boldsymbol{q}(t)\colon t \geq 0\}$ is a CTMC with generator matrix G defined as

$$G_{\boldsymbol{q},\boldsymbol{q}'} = \sum_{i=1}^{2} \mu 1_{\{q_i>0 \text{ and } \boldsymbol{q}'=\boldsymbol{q}-\boldsymbol{e}^{(i)}\}} + \lambda 1_{\{\boldsymbol{q}'=\boldsymbol{q}+\boldsymbol{e}^{(i^\star(\boldsymbol{q}))}\}} - \left(\lambda + \mu \sum_{i=1}^{2} 1_{\{q_i>0\}}\right) 1_{\{\boldsymbol{q}=\boldsymbol{q}'\}},$$

where $\boldsymbol{e}^{(i)}$ is the ith unit vector.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

2.2. Main result

We now present the main result of this paper: the convergence rate for the JSQ system described in Section 2.1.

Theorem 1. *Consider a JSQ system as described in Section 2.1 with $\rho = \lambda/(2\mu) < 1$. Let its initial queue length vector be $\boldsymbol{x}$. Let $\mathbb{P}^t_{\boldsymbol{x}}$ be the distribution of its queue length vector $\boldsymbol{q}(t)$ at time t, and let π be its steady state. Then, for all $t > 0$,*

$$d_{\mathrm{TV}}(\mathbb{P}^t_{\boldsymbol{x}}, \pi) \le \frac{1}{t} C_1(\lambda, \mu, \boldsymbol{x}), \tag{2.1}$$

where

$$C_1(\lambda, \mu, \boldsymbol{x}) = \frac{2-\rho}{2\mu(1-\rho)^3}((x_1 + x_2)(1-\rho) + \rho(2-\rho)). \tag{2.2}$$

Our result in Theorem 1 can be further refined for light-traffic cases (small ρ). Specifically, we obtain a multiplicative factor $K(\rho) \le 1$ in the bound (2.1). This factor provides a more precise characterization of convergence rates in light-traffic scenarios. For a more detailed discussion, we refer the reader to the technical report version of this paper [12]. Next, in Corollary 1, we establish a convergence rate result for the difference between the transient and steady-state mean queue lengths. In particular, the bound (2.3) implies that the mean queue length converges at rate $O((1/(1-\rho)^3)(1/\sqrt{t}))$; the proof is provided in [12]. Combining this corollary with existing results on the stationary queue length of the JSQ system, such as [4, Proposition 4], yields an explicit transient upper bound on the mean queue length, as stated in (2.4).

Corollary 1. *Consider a JSQ system as described in Theorem 1 with $\rho = \lambda/(2\mu) < 1$ and an initial queue length vector $\boldsymbol{x}$. Then*

$$|\mathbb{E}[q_1(t) + q_2(t)] - \mathbb{E}_\pi[q_1 + q_2]| \le \frac{1}{\sqrt{t}}\big(2C_1(\lambda, \mu, \boldsymbol{x}) + C_2(\lambda, \mu, \boldsymbol{x})\big) + \frac{2}{t} C_1(\lambda, \mu, \boldsymbol{x}) \tag{2.3}$$

and

$$\mathbb{E}[q_1(t) + q_2(t)] \le \frac{\rho(2-\rho)}{1-\rho} + \frac{1}{\sqrt{t}}\big(2C_1(\lambda, \mu, \boldsymbol{x}) + C_2(\lambda, \mu, \boldsymbol{x})\big) + \frac{2}{t} C_1(\lambda, \mu, \boldsymbol{x}), \tag{2.4}$$

where $C_1(\lambda, \mu, \boldsymbol{x})$ is defined in (2.2) and

$$C_2(\lambda, \mu, \boldsymbol{x}) = (x_1^2 + x_2^2) + 2(x_1 + x_2)\frac{\rho}{1-\rho} + \frac{4\rho(1+\rho)}{(1-\rho)^2}. \tag{2.5}$$

3. Proof of Theorem 1

In the proof, we use the following hitting time, denoted by T_0 in (3.1). Consider a JSQ system as described in Section 2.1 with queue length process $\boldsymbol{q}(t)$. We define

$$T_0 = \inf\{t \ge 0 \colon \boldsymbol{q}(t) = (0, 0) \mid \boldsymbol{q}(0) = (1, 0)\} \tag{3.1}$$

as the time for the queue length process to hit state (0, 0) starting from (1, 0). In the proof, we use the two coupled processes defined in the following subsection.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

3.1. Coupling with arrival and service

Consider two JSQ systems with queue length vectors $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ as described in Section 2.1. We call them coupled with arrival and service if they can be constructed using $\mathcal{N}_{\mathrm{a}}$ and $\mathcal{N}_{\mathrm{d}}^{i}$, $i = 1, 2$, where $\mathcal{N}_{\mathrm{a}}$ is a Poisson process with rate λ, and $\mathcal{N}_{\mathrm{d}}^{1}, \mathcal{N}_{\mathrm{d}}^{2}$ are two independent Poisson processes, each with rate μ. We allow the initial queue length vectors of $\boldsymbol{p}$ and $\boldsymbol{q}$ to be different. We also assume $\mathcal{N}_{\mathrm{a}}$, $\mathcal{N}_{\mathrm{d}}^{1}$, and $\mathcal{N}_{\mathrm{d}}^{2}$ are independent of each other. We interpret $\mathcal{N}_{\mathrm{a}}$ as the arrival process and $\mathcal{N}_{\mathrm{d}}^{1}, \mathcal{N}_{\mathrm{d}}^{2}$ as the potential service processes. Accordingly, we can write the expressions for the dynamics of the queue length vectors in the two JSQ systems as

$$\begin{aligned}
\mathrm{d}p_1(t) &= 1_{\{p_1(t-)\le p_2(t-)\}}\mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{p_1(t-)\ge 1\}}\mathcal{N}_{\mathrm{d}}^{1}(\mathrm{d}t),\\
\mathrm{d}p_2(t) &= 1_{\{p_1(t-)> p_2(t-)\}}\mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{p_2(t-)\ge 1\}}\mathcal{N}_{\mathrm{d}}^{2}(\mathrm{d}t),\\
\mathrm{d}q_1(t) &= 1_{\{q_1(t-)\le q_2(t-)\}}\mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{q_1(t-)\ge 1\}}\mathcal{N}_{\mathrm{d}}^{1}(\mathrm{d}t),\\
\mathrm{d}q_2(t) &= 1_{\{q_1(t-)> q_2(t-)\}}\mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{q_2(t-)\ge 1\}}\mathcal{N}_{\mathrm{d}}^{2}(\mathrm{d}t),
\end{aligned} \tag{3.2}$$

where $p_1(t-)$ is the value of p_1 evaluated just before time t.

3.2. Proof of Theorem 1

In this section we present the proof of Theorem 1, which builds on Proposition 1 and Lemma 1.

Proposition 1 ([9, 10].) *Let $(X_t)_{t\ge 0}$ and $(Y_t)_{t\ge 0}$ be two copies of the same continuous-time, discrete-valued Markov chain on a common probability space with $X_0 = \boldsymbol{x}$ and $Y_0 \sim \pi$, where π is the steady state of this Markov chain. We assume X_t and Y_t take values in a countable set $\mathcal{X}$, and let $T = \inf\{t > 0 \colon X_t = Y_t\}$ denote the first time that they meet. Then*

$$d_{\mathrm{TV}}(\mathbb{P}_{\boldsymbol{x}}^{t}, \pi) \le \mathbb{P}(T \ge t).$$

Lemma 1. *Consider two coupled JSQ systems with queue length processes $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ as governed by (3.2). Assume $\boldsymbol{p}(0) = \boldsymbol{x}$ and $\boldsymbol{q}(0) \overset{\mathrm{d}}{=} \pi$, where π is the steady state of $\boldsymbol{p}$. Let T be the first time that the two systems have the same queue length vector. We have*

$$\mathbb{P}(T \ge t) \le \frac{1}{t} C_1(\lambda, \mu, \boldsymbol{x}), \tag{3.3}$$

where $C_1(\lambda, \mu, \boldsymbol{x})$ is defined in (2.2).

Now we are ready to present the proof of Theorem 1.

Proof of Theorem 1. Applying Proposition 1 yields $d_{\mathrm{TV}}(\mathbb{P}_{\boldsymbol{x}}^{t}, \pi) \le \mathbb{P}(T \ge t)$, where T is defined in Lemma 1. By (3.3) in Lemma 1, $d_{\mathrm{TV}}(\mathbb{P}_{\boldsymbol{x}}^{t}, \pi) \le \mathbb{P}(T \ge t) \le C_1(\lambda, \mu, \boldsymbol{x})/t$ and we complete the proof. □

3.3. Proof of Lemma 1

To prove Lemma 1, we first establish a monotonicity property for coupled JSQ systems in Lemma 2 with the proof given in Section 3.4. Next, we discuss Lemma 3, proved in Section 3.5, which provides an upper bound on the hitting time to the origin for a JSQ system starting

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

from an arbitrary state $\boldsymbol{x}$, in terms of T_0 defined in (3.1). We then bound $\mathbb{E}[T_0]$ in Lemma 4, with the proof provided in Section 3.6. Together, these results imply Lemma 1.

Lemma 2. *Consider two coupled JSQ systems (3.2) with queue length processes $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$. If $p_i(0) \geq q_i(0)$ for $i \in \{1, 2\}$ then $p_i(t) \geq q_i(t)$ for $i \in \{1, 2\}$ and for every $t \geq 0$. In particular, for any $i \in \{1, 2\}$ and $t \geq 0$, $p_i(t) = 0 \Rightarrow q_i(t) = 0$. Further, define $r_i(t) = p_i(t) - q_i(t) \geq 0$ and $r(t) = r_1(t) + r_2(t)$. Then $r(t) \leq r(0)$ for any $t \geq 0$.*

Lemma 3. *Consider a JSQ system with queue length process $\boldsymbol{q}(t)$. Define the random variable $\mathcal{T}(\boldsymbol{x})$ as the time for $\boldsymbol{q}(t)$ to first hit state $(0, 0)$ with $\boldsymbol{q}(0) = (x_1, x_2)$:*

$$\mathcal{T}(\boldsymbol{x}) = \mathcal{T}(x_1, x_2) = \inf\{t \geq 0 \colon \boldsymbol{q}(t) = (0, 0) \mid \boldsymbol{q}(0) = (x_1, x_2)\}.$$

Then we have $\mathcal{T}(x_1, x_2) \preceq \sum_{i=1}^{x_1+x_2} T_{0i}$, where T_{0i} are independent and identically distributed (i.i.d.) copies of T_0 (3.1) and the relation $\preceq$ is defined in Section 1.3.

Lemma 4. *Consider T_0 (3.1) and assume $\rho = \lambda/(2\mu) < 1$. Then*

$$\mathbb{E}[T_0] \leq \frac{1}{\lambda} \frac{\rho(2-\rho)}{(1-\rho)^2}. \tag{3.4}$$

To prove Lemma 1, we also use Proposition 2 from the literature, which provides an upper bound on $\mathbb{E}_\pi[q_1 + q_2]$ for the JSQ system.

Proposition 2. ([4, Proposition 4 and Remark 2].) *For the JSQ system described in Theorem 1 with $\rho = \lambda/(2\mu) < 1$, $\mathbb{E}_\pi[q_1 + q_2] \leq \rho(2-\rho)/(1-\rho)$.*

Although [4] uses uniform random tie-breaking, its total-queue-length bound remains valid here because the sorted queue length process is independent of the tie-breaking rule.

Proof of Lemma 1. We consider a JSQ system with queue length process $\boldsymbol{s}$, driven by the same arrival and potential service processes as $\boldsymbol{p}$ and $\boldsymbol{q}$, with $s_j(0) = \max\{q_j(0), p_j(0)\}$, $j \in \{1, 2\}$. We couple the three systems $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{s}$ similarly to Section 3.1. Formally, let $\boldsymbol{s}$ evolve according to

$$\begin{aligned} \mathrm{d}s_1(t) &= 1_{\{s_1(t-) \leq s_2(t-)\}} \mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{s_1(t-) \geq 1\}} \mathcal{N}_{\mathrm{d}}^1(\mathrm{d}t), \\ \mathrm{d}s_2(t) &= 1_{\{s_1(t-) > s_2(t-)\}} \mathcal{N}_{\mathrm{a}}(\mathrm{d}t) - 1_{\{s_2(t-) \geq 1\}} \mathcal{N}_{\mathrm{d}}^2(\mathrm{d}t). \end{aligned}$$

From Lemma 3, the time $\boldsymbol{s}$ hits 0, denoted by T', is stochastically dominated by $\sum_{i=1}^{s_1(0)+s_2(0)} T_{0i}$, where T_{0i} are i.i.d. copies of T_0 (3.1). Since $\boldsymbol{s}(0) \geq \boldsymbol{p}(0)$ and $\boldsymbol{s}(0) \geq \boldsymbol{q}(0)$ componentwise, Lemma 2 (applied to the coupled pairs $(\boldsymbol{s}, \boldsymbol{p})$ and $(\boldsymbol{s}, \boldsymbol{q})$) implies that, for all $t \geq 0$, $\boldsymbol{s}(t) \geq \boldsymbol{p}(t)$ and $\boldsymbol{s}(t) \geq \boldsymbol{q}(t)$ componentwise. Thus, for any $t \geq 0$, $\boldsymbol{s}(t) = (0, 0)$ implies $\boldsymbol{p}(t) = (0, 0)$ and $\boldsymbol{q}(t) = (0, 0)$, and hence $T \leq T'$ almost surely. Since $T' \preceq \sum_{i=1}^{s_1(0)+s_2(0)} T_{0i}$, we have

$$\begin{aligned} \mathbb{E}[T] \leq \mathbb{E}[T'] &= \mathbb{E}_{\boldsymbol{s}(0)}[\mathbb{E}[T' \mid \boldsymbol{s}(0)] \\ &\leq \mathbb{E}_{\boldsymbol{s}(0)}[(s_1(0) + s_2(0))\mathbb{E}[T_0]] \\ &= \mathbb{E}_{\boldsymbol{q}(0)}\big[\max\{q_1(0), p_1(0)\} + \max\{q_2(0), p_2(0)\}\big]\mathbb{E}[T_0]. \end{aligned} \tag{3.5}$$

Using $\max\{x, y\} \leq x + y$ for $x, y \geq 0$ and $\boldsymbol{q}(0) \overset{\mathrm{d}}{=} \pi$, (3.5) implies that

$$\mathbb{E}[T] \leq \mathbb{E}[T_0]\big((x_1 + x_2) + \mathbb{E}_\pi[q_1 + q_2]\big).$$

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

Using $\mathbb{E}_\pi[q_1+q_2] \le \rho(2-\rho)/(1-\rho)$ from Proposition 2 and recalling that

$$\mathbb{E}[T_0] \le \frac{1}{\lambda}\frac{\rho(2-\rho)}{(1-\rho)^2} = \frac{2-\rho}{2\mu(1-\rho)^2}$$

from (3.4) in Lemma 4, we obtain $\mathbb{E}[T] \le C_1(\lambda, \mu, \boldsymbol{x})$, (2.2). Applying Markov's inequality $\mathbb{P}(T \ge t) \le \mathbb{E}[T]/t$ for $t > 0$, we obtain $\mathbb{P}(T \ge t) \le (1/t)C_1(\lambda, \mu, \boldsymbol{x})$ and we complete the proof of Lemma 1. □

3.4. Proof of Lemma 2

Proof. Consider the processes $\mathcal{N}_{\mathrm{a}}(t)$, $\mathcal{N}_{\mathrm{d}}^1(t)$, and $\mathcal{N}_{\mathrm{d}}^2(t)$ in (3.2). Fix a sample path. Let t_1 be the first time that either an arrival or a potential service completion event occurs. In words, t_1 is the smallest time t such that $\mathcal{N}_{\mathrm{a}}(t) = 1$ or $\mathcal{N}_{\mathrm{d}}^1(t) = 1$ or $\mathcal{N}_{\mathrm{d}}^2(t) = 1$. We will show that $p_i(t_1) \ge q_i(t_1)$.

(i) If the first event is a potential service completion, then as we assume the two systems are coupled, we have $\mathrm{d}p_j(t_1) = -1_{\{p_j(0)\ge 1\}}$ and $\mathrm{d}q_j(t_1) = -1_{\{q_j(0)\ge 1\}}$, where j is the index such that $\mathcal{N}_{\mathrm{d}}^j(t_1) = 1$. It is easy to verify that $p_i(t_1) \ge q_i(t_1)$ for $i \in \{1, 2\}$ given $p_i(0) \ge q_i(0)$.

(ii) Next, suppose that the first event is an arrival. For contradiction, assume that $p_j(t_1) < q_j(t_1)$ for some $j \in \{1, 2\}$. Since $p_j(0) \ge q_j(0)$ and only one arrival occurs at time t_1, we must have

$$p_j(0) = q_j(0), \tag{3.6}$$

and at time t_1, the arrival must be routed to queue $3-j$ in the p-system and to queue j in the q-system. Thus, the JSQ rule implies that $p_{3-j}(0) \le p_j(0)$ and $q_j(0) \le q_{3-j}(0)$. Combining these inequalities with $\boldsymbol{p}(0) \ge \boldsymbol{q}(0)$ and (3.6), we obtain

$$p_{3-j}(0) \ge q_{3-j}(0) \ge q_j(0) = p_j(0) \ge p_{3-j}(0).$$

Every inequality in the above chain must therefore be an equality. Thus, $p_1(0) = p_2(0)$ and $q_1(0) = q_2(0)$. By the deterministic tie-breaking rule, at t_1 both systems must route the arrival to queue 1, contradicting the fact that the arrival is routed to queue j in one system and to queue $3-j$ in the other.

Combining the above two cases, we show that $p_i(t_1) \ge q_i(t_1)$ for $i \in \{1, 2\}$. Proceeding in the same way, we show that $p_i(t) \ge q_i(t)$ for any $t \ge 0$ and $i \in \{1, 2\}$.

Next, we will show that $r(t) \le r(0)$ holds for every $t \ge 0$. We fix a sample path. Consider the processes $\mathcal{N}_{\mathrm{a}}(t)$, $\mathcal{N}_{\mathrm{d}}^1(t)$, and $\mathcal{N}_{\mathrm{d}}^2(t)$ in (3.2). Let t denote the first time that either an arrival or a potential service completion event occurs.

(i) If the first event is an arrival, then it is obvious that $r(t) = r(0)$.

(ii) If the first event is a potential service completion, then we let i denote the index such that $\mathcal{N}_{\mathrm{d}}^i(t) = 1$. Since $\boldsymbol{p}(0) \ge \boldsymbol{q}(0)$, we have $p_i(t) \ge q_i(t)$ as shown above. Thus, there are two cases:

- $1_{\{p_i(t-)>0\}} = 1_{\{q_i(t-)>0\}}$, which implies $r(t) = r(0)$.
- $p_i(t-) > 0$ and $q_i(t-) = 0$, and $\mathrm{d}p_i(t) = -1$ and $\mathrm{d}q_i(t) = 0$, which implies that $r(t) \le r(0)$.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

Combining the two cases, we show that $r(t) \le r(0)$ for the t defined above. Using the same argument, we show that $r(t) \le r(0)$ holds for every $t \ge 0$. □

3.5. Proof of Lemma 3

Proof. Consider two coupled JSQ systems with queue length processes $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ as governed by (3.2). Suppose $\boldsymbol{p}(0) = (x+1, y)$ and $\boldsymbol{q}(0) = (x, y)$ for some $x \ge 0$, $y \ge 0$. From Lemma 2, for any $t \ge 0$, $\boldsymbol{q}(t) = (0, 0)$ implies that $\boldsymbol{p}(t) \in \{(1, 0), (0, 1), (0, 0)\}$.

Next, we show that $T(0, 1) \overset{\mathrm{d}}{=} T(1, 0) \overset{\mathrm{d}}{=} T_0$. We couple two copies of the JSQ system, with queue-length processes $\boldsymbol{q}^{(1)}$ and $\boldsymbol{q}^{(2)}$, started from (1, 0) and (0,1), respectively. We use the same arrival process in both systems and couple the service-completion times of their unique initial jobs to be identical. Thus, if the service completion occurs before the first arrival, both systems move to (0, 0); if an arrival occurs first, both systems move to (1,1). Hence, under this coupling, the two copies are in the same post-jump state at the first event time. From that time onward, we couple them using the same arrival and potential service processes, so that they evolve identically. Thus, $T(0, 1) \overset{\mathrm{d}}{=} T(1, 0) \overset{\mathrm{d}}{=} T_0$.

Combining $\boldsymbol{q}(t) = (0, 0) \Rightarrow \boldsymbol{p}(t) \in \{(1, 0), (0, 1), (0, 0)\}$ for all $t \ge 0$, $T(0, 1) \overset{\mathrm{d}}{=} T(1, 0) \overset{\mathrm{d}}{=} T_0$, and the strong Markov property, we have $\mathcal{T}(x+1, y) \preceq \mathcal{T}(x, y) + T_0'$, where T_0' is an independent copy of T_0. Similarly, we can show that $\mathcal{T}(x, y+1) \preceq \mathcal{T}(x, y) + T_0'$. Iterating these one-step bounds yields $T(x_1, x_2) \preceq \sum_{i=1}^{x_1+x_2} T_{0i}$. □

3.6. Proof of Lemma 4

Proof. Consider two independent M/M/1 queues with arrival rates $\lambda/2$ and service rates μ, and denote their queue lengths as $q_1^{\mathrm{M}}(t)$ and $q_2^{\mathrm{M}}(t)$. Then their Cartesian product, $\boldsymbol{q}^{\mathrm{M}}(t) = (q_1^{\mathrm{M}}(t), q_2^{\mathrm{M}}(t))$, is also a CTMC. Since each marginal queue is irreducible and positive recurrent under the condition $\lambda/2 < \mu$, it follows that the joint process $\boldsymbol{q}^{\mathrm{M}}(t)$ is also irreducible and positive recurrent, with the stationary distribution given by the product of the two marginal stationary distributions. Under stationarity, $\mathbb{P}(q_1^{\mathrm{M}} = k) = \mathbb{P}(q_2^{\mathrm{M}} = k) = \rho^k(1-\rho)$ for all $k \in \mathbb{Z}_+$. Since they are independent, $\mathbb{P}_{\pi^{\mathrm{M}}}\left(q_1^{\mathrm{M}} = q_2^{\mathrm{M}} = 0\right) = (1-\rho)^2$, where π^{M} is the stationary distribution of $\boldsymbol{q}^{\mathrm{M}}$.

Let $m_{(0,0)}^{\mathrm{M}}$ denote the expected time for $\boldsymbol{q}^{\mathrm{M}}$ to return to (0, 0) when starting from (0, 0), including the holding time in (0, 0). Applying [13, Theorem 3.5.3] to $\boldsymbol{q}^{\mathrm{M}}$, and noting that the total transition rate out of state (0, 0) is $\nu_{(0,0)}^{\mathrm{M}} = \lambda$, we obtain

$$m_{(0,0)}^{\mathrm{M}} = \frac{1}{\nu_{(0,0)}^{\mathrm{M}} \mathbb{P}_{\pi^{\mathrm{M}}}\left(q_1^{\mathrm{M}} = q_2^{\mathrm{M}} = 0\right)} = \frac{1}{\lambda(1-\rho)^2}. \tag{3.7}$$

Starting from state (0, 0), the process $\boldsymbol{q}^{\mathrm{M}}$ leaves (0, 0) after an Exp(λ) time and moves to either (1, 0) or (0,1). Let $T_{(1,0)}^{(0,0)}$ be the time for $\boldsymbol{q}^{\mathrm{M}}$ to hit state (0, 0) starting from state (1, 0). By symmetry,

$$m_{(0,0)}^{\mathrm{M}} = \frac{1}{\lambda} + \mathbb{E}\left[T_{(1,0)}^{(0,0)}\right].$$

Combining the above equation with (3.7), we have

$$\mathbb{E}\left[T_{(1,0)}^{(0,0)}\right] = \frac{1}{\lambda}\left(\frac{1}{(1-\rho)^2} - 1\right) = \frac{1}{\lambda}\frac{\rho(2-\rho)}{(1-\rho)^2}. \tag{3.8}$$

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

Next, consider a two-queue JSQ system $\boldsymbol{q}^{\mathrm{J}}(t) = (q_1^{\mathrm{J}}(t), q_2^{\mathrm{J}}(t))$ (with an arrival rate λ and service rate μ at each queue) and the product chain $\boldsymbol{q}^{\mathrm{M}}(t) = (q_1^{\mathrm{M}}(t), q_2^{\mathrm{M}}(t))$ of the two independent M/M/1 queues defined previously, both started from the same initial state $\boldsymbol{q}^{\mathrm{J}}(0) = \boldsymbol{q}^{\mathrm{M}}(0) = (1, 0)$. By [18, Theorem 4], there exists a coupling of $\{\boldsymbol{q}^{\mathrm{J}}(t)\}_{t\geq 0}$ and $\{\boldsymbol{q}^{\mathrm{M}}(t)\}_{t\geq 0}$ such that

$$\sum_{i=1}^{2} \left[q_i^{\mathrm{J}}(t) - x\right]_+ \leq \sum_{i=1}^{2} \left[q_i^{\mathrm{M}}(t) - x\right]_+ \quad \text{for all } x \geq 0,\ t \geq 0, \tag{3.9}$$

where $[x]_+ = \max\{x, 0\}$. In particular, taking $x = 0$ in (3.9) yields $\sum_{i=1}^{2} q_i^{\mathrm{J}}(t) \leq \sum_{i=1}^{2} q_i^{\mathrm{M}}(t)$ for all $t \geq 0$. It follows that

$$T_0 = \inf\{t \geq 0 \colon \boldsymbol{q}^{\mathrm{J}}(t) = (0, 0)\} \leq \inf\{t \geq 0 \colon \boldsymbol{q}^{\mathrm{M}}(t) = (0, 0)\} = T_{(1,0)}^{(0,0)} \quad \text{almost surely.}$$

Using the above inequality and (3.8), we have shown that

$$\mathbb{E}[T_0] \leq \mathbb{E}\big[T_{(1,0)}^{(0,0)}\big] = \frac{1}{\lambda} \frac{\rho(2-\rho)}{(1-\rho)^2},$$

and we have completed the proof. □

4. Conclusion and future work

In this paper we present a convergence rate result for a load-balancing system with two queues under the JSQ policy, which is valid for any $\rho \in (0, 1)$. Our method is based on a novel and simple coupling construction and on computing an upper bound on the expected hitting time. Instead of relying on approximations of the stochastic systems in relevant limits, we directly analyze the original system, yielding clean insights regarding the convergence rate and how it depends on ρ.

We now consider some interesting future directions. Recall that we have established a convergence rate that is of order $O((1/(1-\rho)^3)(1/t))$. However, based on the results for a single-server queue [14] and also based on the scaling used in diffusion approximations, we expect an exponential convergence rate of the form $O(\mathrm{e}^{-\alpha\varepsilon^2 t})$ with $\varepsilon = 1 - \rho$ for an appropriate constant α. Establishing such a strong bound on convergence is a future research direction, and we believe it requires adopting novel methodological approaches, particularly by exploiting state-space collapse results from the literature. Another possible approach is to compute an upper bound on the moment-generating function of the hitting time. An alternative direction is to extend our analysis to a system with $n \geq 3$ servers and heterogeneous service rates. We also hope the technique we have developed in this paper can spur further analysis of more complicated stochastic systems.

Acknowledgements

The authors would like to thank the editor, the editor-in-chief, and the three referees for their continued consideration and helpful feedback, which greatly improved the quality of this paper. We would also like to thank Sushil Varma and Daniela Hurtado Lange for helpful discussions during the early stages of this work.

Funding information

This work was partially supported by NSF grants EPCN-2144316 and CMMI-2140534.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ABATE, J. AND WHITT, W. (1987). Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems* **2**, 41–65.

[2] BENSON, T., AKELLA, A. AND MALTZ, D. A. (2010). Network traffic characteristics of data centers in the wild. In *Proc. 10th ACM SIGCOMM Conf. Internet Measurement*, pp. 267–280.

[3] CONOLLY, B. W. (1984). The autostrada queueing problem. *J. Appl. Prob.* **21**, 394–403.

[4] DESTER, P. S., FRICKER, C. AND TIBI, D. (2017). Stationary analysis of the shortest queue problem. *Queueing Systems* **87**, 211–243.

[5] DI CRESCENZO, A. AND MARTINUCCI, B. (2008). A first-passage-time problem for symmetric and similar two-dimensional birth–death processes. *Stoch. Models* **24**, 451–469.

[6] DOWN, D., MEYN, S. P. AND TWEEDIE, R. L. (1995). Exponential and uniform ergodicity of Markov processes. *Ann. Prob.* **23**, 1671–1691.

[7] GAMARNIK, D. AND GOLDBERG, D. A. (2013). On the rate of convergence to stationarity of the M/M/*N* queue in the Halfin–Whitt regime. *Ann. Appl. Prob.* **23**, 1879–1912.

[8] GRASSMANN, W. K. (1980). Transient and steady state results for two parallel queues. *Omega* **8**, 105–112.

[9] LEVIN, D. A. AND PERES, Y. (2017). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI.

[10] LINDVALL, T. (2002). *Lectures on the Coupling Method*. Dover Publications, Garden City, NY.

[11] LUND, R. B., MEYN, S. P. AND TWEEDIE, R. L. (1996). Computable exponential convergence rates for stochastically ordered Markov processes. *Ann. Appl. Prob.* **6**, 218–237.

[12] MA, Y. AND MAGULURI, S. T. (2025). Convergence rate of the join-the-shortest-queue system. Preprint, arXiv:2503.15736 [math.PR].

[13] NORRIS, J. R. (1997). *Markov Chains*. Cambridge University Press.

[14] ROBERT, P. (2003). *Stochastic Networks and Queues*. Springer, New York.

[15] SRIKANT, R. AND YING, L. (2014). *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press.

[16] TAKÁCS, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press.

[17] TIBI, D. (2019). Martingales and buffer overflow for the symmetric shortest queue model. *Queueing Systems* **93**, 153–190.

[18] TURNER, S. R. E. (1998). The effect of increasing routing choice on resource pooling. *Prob. Eng. Inf. Sci.* **12**, 109–124.

[19] VAN DER BOOR, M., BORST, S. C., VAN LEEUWAARDEN, J. S. H. AND MUKHERJEE, D. (2022). Scalable load balancing in networked systems: A survey of recent advances. *SIAM Rev.* **64**, 554–622.

[20] VAN LEEUWAARDEN, J. S. H. AND KNESSL, C. (2011). Transient behavior of the Halfin–Whitt diffusion. *Stoch. Process. Appl.* **121**, 1524–1545.

[21] WINSTON, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14**, 181–189.

[22] YAO, H. AND KNESSL, C. (2008). Some first passage time problems for the shortest queue model. *Queueing Systems* **58**, 105–119.

https://doi.org/10.1017/jpr.2026.10130 Published online by Cambridge University Press