Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Stability of Parallel Server Systems

Pascal Moyal
LMAC–UTC and IECL–Université de Lorraine, Vandoeuvre-les-Nancy, pascal.moyal@univ-lorraine.fr

Ohad Perry
Department of Industrial Engineering and Management Science, Northwestern University, ohad.perry@northwestern.edu

The fundamental problem in the study of parallel-server systems is that of finding and analyzing routing policies of arriving jobs to the servers that efficiently balance the load on the servers. The most well-studied policies are (in decreasing order of efficiency) *join the shortest workload* (JSW), which assigns arrivals to the server with the least workload; *join the shortest queue* (JSQ), which assigns arrivals to the smallest queue; the power-of-d (PW(d)), which assigns arrivals to the shortest among $d \geq 1$ queues that are sampled from the total of s queues uniformly at random; and uniform routing, under which arrivals are routed to one of the s queues uniformly at random.

In this paper we study the stability problem of parallel-server systems, assuming that routing errors may occur, so that arrivals may be routed to the "wrong" queue (not the smallest among the relevant queues) with a positive probability. We treat this routing mechanism as a probabilistic routing policy, named a **p**-allocation policy, that generalizes the PW(d) policy, and thus also the JSQ and uniform routing, where **p** is an s-dimensional vector whose components are the routing probabilities. Our goal is to study the (in)stability problem of the system under this routing mechanism, and under its "non-idling" version, which assigns new arrivals to an idle server, if such a server is available, and otherwise routs according to the **p**-allocation rule. We characterize a sufficient condition for stability, and prove that the stability region, as a function of the system's primitives and **p**, is in general smaller than the set $\{\rho < 1\}$. Our analyses build on representing the queue process as a continuous-time Markov chain in an ordered space of s-dimensional real-valued vectors, and employing a generalized form of the Schur-convex order.

1. Introduction

We consider a parallel-server system with $s \geq 2$ statistically-homogeneous servers, each providing service at rate μ, that is fed by a rate-λ Poisson arrival process of statistically identical jobs (or customers). For each server there is a dedicated infinite buffer in which jobs queue, waiting for their turn to be served. Upon arrival, a job is routed to one of the s servers according to some prespecified dispatching (routing) rule, with no jockeying between the queues allowed. In this setting, one seeks a “good” routing policy of jobs to the servers, e.g., a policy ensuring that steady state waiting times are minimized, or that the total throughput rate is maximized. If the workload at each queue can be computed, then it is natural to employ the Join the Shortest Workload (JSW) routing policy, under which an arriving job is routed to the server with the least workload among all s servers (together with some tie-breaking rule). However, if the workload is unknown, as is often the case in practice, one may opt to employ the Join-the-Shortest Queue (JSQ) control, which routes an arriving job to the server with the smallest number of jobs. Indeed, JSW was shown to minimize the workload process in [13], whereas JSQ has been shown to be throughput maximizing in terms of stochastic order, when the service-time distribution has a non-decreasing failure rate [50], and in particular, when the service times are exponentially distributed [53].

However, even the queue at each server is not always known: In some settings, the number of customers in each queue is estimated, either by the arriving customers who are free to choose which queue to join (as in a supermarket or security lanes in airports), or by a central dispatcher (as is often the case in passport-checking stations, for example). Even in automated settings the queue lengths may not be known. For example, information regarding the queues to each of the servers in web-server farms requires constant communication between the servers and the job dispatchers, slowing down the response time, and is thus not always available; e.g., see [33].

For this reason, other routing policies have been considered in the literature, most notably the “power-of-d” policy, which gives rise to the so-called “supermarket model” [37]. Under this policy, upon each arrival d servers are chosen uniformly at random, and that arrival is routed to the server with the smallest number of jobs among the d sampled queues, with ties broken uniformly at random. We denote this routing rule by PW(d) and note that $d = 1$ corresponds to uniform routing (i.e. any incoming job is sent to a queue that is chosen uniformly at random), whereas $d = s$ corresponds to JSQ.

1.1. Motivation and Goals

We are motivated by the fact that, unlike the idealized settings considered in the literature, routing errors can occur in practice. In this regard, our main goal is to gain an understanding of how the frequency at which such errors occur affects the overall system's stability. To this end, we study a particular form of error, under which arrivals are sent to the "wrong" queue (not the smallest) with a fixed probability, and show that the system might not be stable in this case, even if its total service rate is larger than the rate at which work arrives, i.e., if the traffic intensity to the system is smaller than 1.

Routing errors are likely to occur when JSW is employed, because the actual workload at each server can only be estimated, unless the server is idle. Similarly, such errors are likely to occur under JSQ when customers are free to choose which queue to join, or when a central dispatcher has only partial information about the queue lengths. Here we focus on the latter JSQ policy, since under appropriate distributional assumptions (Poisson arrival process and exponentially distributed service times), the queue process evolves as a continuous-time Markov chain (CTMC), whereas under JSW, the analysis of the queue process requires a continuous-space Markov representation. (Even under JSQ, exact analyses and steady-state computations of the queue are intractable, and most of the literature is concerned with asymptotic approximations; see Section 2 below.) The simulation examples in Section 6 suggest that our results extend to the JSW case.

Even though our main motivation is to study the impact of routing errors, we treat the allocation of jobs to servers as a probabilistic routing policy. We do this for mathematical convenience, as it allows us to treat PW(d), and therefore also JSQ and uniform routing, as a special case of the family of allocation policies we consider. Specifically, we assume that the dispatcher (or the arriving customer) chooses correctly the shortest queue with probability p_1, the second-shortest queue with probability p_2, and so forth. We also consider a **"non-idling"** version, in which routing errors are made only when all servers are busy, so that the dispatcher (or arriving customer) always chooses an idle server, if such a server is available, and otherwise makes errors as was just described. To show that such errors can lead to extreme departures from the desired behavior under JSQ, we characterize the stability region under the allocation policy as a function of the system's parameters and the error probabilities, and prove that the usual traffic condition $\rho := \lambda/(s\mu) < 1$ does not guarantee that the system is stable, *even in the non-idling case.*

1.2. Background: PW(d) and Related Routing Policies

Note that it is not immediately clear that the condition $\rho < 1$ does not imply that the system under a **p**-allocation policy is stable, especially under the non-idling mechanism, because such policies leave a lot of "room" for making routing errors, as can be seen by comparing a system operating under either one of the two extremes—JSQ and uniform routing. Clearly, uniform routing induces a lot of "avoidable" idleness in the system, because arrivals are often routed to busy servers even if there are idle servers present. Nevertheless, by symmetry, the rate at which jobs arrive at each server is the same under this policy, implying that the traffic intensity at each server separately is smaller than 1 whenever the traffic intensity ρ to the whole system is smaller than 1. When the arrival process to the system is Poisson, this follows directly from the splitting property of the Poisson process, which implies that each server operates as an $M/G/1$ queue independently of all other servers. Indeed, if service times are exponentially distributed, in addition to having a Poisson arrival process, so that the queue process evolves as a CTMC, the improvement that JSQ provides over uniform routing follows from existing results, which we now review.

Let $Q^{(d)}_{\Sigma}(t)$ denote the total number of jobs in the system at time $t \geq 0$ under PW(d). Theorem 4 in [47] implies that[1], if $d_1 > d_2$, then $Q^{(d_1)}_{\Sigma} \leq_{st} Q^{(d_2)}_{\Sigma}$, where $\leq_{st}$ denotes sample-path stochastic-order. (That is, there exists a coupling of the two processes, such that $Q^{(d_1)}_{\Sigma}(t) \leq Q^{(d_2)}_{\Sigma}(t)$ w.p.1 for all $t > 0$, provided that the inequality holds at time $t = 0$.) In particular, for $s > 2$,

$$Q^{(s)}_{\Sigma} \leq_{st} Q^{(d)}_{\Sigma} \leq_{st} Q^{(1)}_{\Sigma}, \quad 1 < d \leq s. \tag{1}$$

The stability of a parallel-server system under PW(d) readily follows. To state this result formally, we say that a parallel-server system is "Markovian" if its multi-dimensional queue process evolves as a CTMC. In particular, the arrival process is Poisson and the service times are independent and identically distributed (i.i.d.) exponentially distributed random variables, that are independent of the arrival process and of the state of the system.

COROLLARY 1. *For a Markovian parallel-server system with s servers operating under PW(d), $1 \le d \le s$, the condition $\rho := \lambda/(s\mu) < 1$ is necessary and sufficient in order for the queue process to be an ergodic CTMC.*

Proof. It is easy to see that $Q_\Sigma^{(d)}$ is an irreducible CTMC. If $\rho \ge 1$, then $Q_\Sigma^{(d)}$ is either null recurrent or transient, because it is bounded from below, in sample-path stochastic order, by the number-in-system process in an $M/M/1$ queue with arrival rate λ and service rate $s\mu$. On the other hand, if $\rho < 1$, then $Q_\Sigma^{(1)}$ is ergodic, because it evolves as s independent $M/M/1$ queues, each with arrival rate λ/s and service rate μ. In particular the empty state (zeroth vector) is positive recurrent for the CTMC $Q_\Sigma^{(1)}$, and, by virtue of (1), also for $Q_\Sigma^{(d)}$, $1 < d \le s$. □

A more quantitative analysis can be carried out asymptotically, by taking the number of servers s to infinity, assuming that the arrival rate grows proportionally to s. As was shown in [37, 49], the steady-state probability that an arrival is routed to a queue of length at least k is ρ^{d^k}, i.e., it is doubly exponential in k for $d \ge 2$, as opposed to exponential when $d = 1$ (which is tantamount to uniform routing). The dramatic differences between the *maximum* queue length in stationarity in the cases $d = 1$ and $d \ge 2$ is demonstrated in [34], which shows that the maximum queue length is of order $\ln(s)/\ln(1/\lambda)$ when $d = 1$, and of order $\ln\ln(s)/\ln(d)$ when $d \ge 2$ with probability converging to 1 as $s \longrightarrow \infty$. Further, heavy-traffic analysis shows that the performance under PW(d), for any fixed $d < s$, is substantially worse than under JSQ. In particular, considering a sequence of systems indexed by the number of servers s, and letting λ_s denote the arrival rate to system s, [16] and [17] analyze a system operating under JSQ and PW(d), respectively, in the heavy-traffic limiting regime, where $\lambda_s = s\mu - \Theta(\sqrt{s})$. It is proved in [16] that, under JSQ, only a negligible proportion (which converges to 0) of the customers encounter a queue upon arrival, and those customers that have to wait encounter only one customer in queue. Thus, asymptotically, no queue is larger than

2. (This result holds only after some transient period, because the initial condition may have many larger queues.) On the other hand, [17] proves that, in the supermarket model with $d > 1$, the fraction of queues that are of order $\log_d \sqrt{s}$ approaches 1 as $s \to \infty$.

To conclude, the dimensionality of the queue process, and the fact that it is not reversible, render exact analysis of parallel-server systems intractable, even under Markovian assumptions. Other than stability results and stochastic domination, as in (1), little can be said about the systems' dynamics and steady-state distributions. Nevertheless, the aforementioned asymptotic results suggest that JSQ is substantially more efficient than PW(d) for $d < s$, which, in turn, is substantially more efficient than uniform routing, namely, than PW(1).

Of course, the possibility of experiencing congestion collapse in parallel-server systems can nevertheless be considered a triviality for vacuous choices of the control. For example, if the arrival rate λ is larger than the service rate μ (but is smaller than $s\mu$), then the policy that routes all arrivals to the same server is clearly unstable. Here, however, we perform a refined analysis of the (in)stability region for the non-idling version of JSQ when routing errors occur with a nonnegligible probability.

1.3. Notation

We use $\mathbb{R}$ to denote the set of real numbers, with $\mathbb{R}_+ = [0, \infty)$, $\mathbb{Z}_+$ to denote the set of non-negative integers, and $\mathbb{Z}_+^* := \mathbb{Z}_+ - \{0\}$ the subset of (strictly) positive integers. For any $q \in \mathbb{Z}_+$ and all sets A, we denote by A^q the set of vectors of dimension q having elements in A, e.g., $\mathbb{R}^q$ is the set of q-dimensional real-valued vectors. Vectors are in general denoted by bold letters. For a vector $\mathbf{x} = (x_1, ..., x_q)$ in $\mathbb{R}^q$, we denote by $\mathcal{R}(\mathbf{x})$ the ordered version of $\mathbf{x}$, i.e. $\mathcal{R}(\mathbf{x}) = (x_{(1)}, x_{(2)}, \dots, x_{(q)})$ is any permutation of the elements of $\mathbf{x}$ such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(q)}$. The set of ordered vectors in A^q is denoted by $\mathcal{R}(A^q)$; for example, $\mathcal{R}(\mathbb{R}_+^q) := \{\mathbf{x} \in \mathbb{R}_+^s : x_1 \leq \cdots \leq x_q\}$.

We let $\mathbf{a} \circ \mathbf{x} \in \mathbb{R}^q$ denote the Hadamard product of two vectors $\mathbf{x} = (x_1, ..., x_q)$ and $\mathbf{y} = (y_1, ..., y_q)$ in $\mathbb{R}^q$, i.e., $\mathbf{y} \circ \mathbf{x} = (y_1 x_1, ..., y_q x_q)$. For $\mathbf{x} \in \mathbb{R}_+^q$, we define $n^+(\mathbf{x})$ to be the number of positive coordinates of $\mathbf{x}$, which is 0 if $\mathbf{x}$ is the zeroth vector $\mathbf{0} := (0, \dots, 0)$. Let $[\![p, q]\!] = \mathbb{Z}_+ \cap [p, q]$. For any

$i \in [\![1, q]\!]$, let $\mathbf{e}_i$ denote the vector having all coordinates equal to 0 except the ith coordinate, which is equal to 1, and let $\mathbf{e}$ denote the unit vector whose components are all equal 1; $\mathbf{e} := (1, \ldots, 1)$. For any $\mathbf{x} \in \mathbb{R}_+^q$ we denote by $\| \mathbf{x} \| = \sum_{i=1}^q x_i$ and $\| \mathbf{x} \|_2 = \sqrt{\sum_{i=1}^q x_i^2}$. For any two real numbers a and b, let $a \vee b$ and $a \wedge b$ denote the maximum and the minimum of a and b, respectively, and denote $a^+ := a \vee 0$.

1.4. Organization

The rest of the paper is organized as follows: We provide a detailed literature review in Section 2. The model, including the family of allocation policies, which we call **p**-allocation policies, is formally introduced in Section 3. In Section 4 we study a class of **p**-allocation policies for which the condition $\rho < 1$ implies that the system is stable. The insufficiency of this traffic condition to imply stability in general is demonstrated in Section 5. In Section 6 we present simulation results which suggest that our main results extend to workload-based routing policies. We Summarize in Section 7. Some of the technical proofs, together with auxiliary results, appear in an appendix.

2. Related Literature

Non-monotonic parallel queues. Under JSW, the dynamics of the system, as well as the sojourn time of jobs, coincide with those of a single-queue s-server system operating under the First In First Out (FIFO) service policy. In particular, that $\rho < 1$ is a necessary and sufficient condition for the stability of the system under JSW follows from from the basic stability theory of the $GI/GI/s$ queue, first proved in the seminal paper [27]. The sufficiency of the condition $\rho < 1$ for stability of the $G/G/s$ queue was generalized in [8] to the stationary ergodic framework, namely, when both the inter-arrival and service-time sequences are time-stationary and ergodic, but not necessarily independent; see also §2.2 of [3]. This general result was proved using a backwards scheme of the Loynes type [31], building on the fact that the (random) updating map of the stochastic recursive sequence representing the system is non-decreasing for the coordinate-wise vector ordering. For the same reason, JSW is the unique routing rule within the class of semi-cyclic policies introduced in [46], which renders the total workload to be a non-decreasing function of s at all times; see [39]. Therefore, the stability region under allocation policies *other than JSW* cannot simply be characterized via a Loynes-type construction, and we must therefore adopt a different approach.

JSQ systems. The JSQ policy was first introduced in [25] for a system with two servers, each having a different service rate. The first proof that the condition $\rho < 1$ is necessary and sufficient for a Markovian parallel-server system under JSQ to be stable (admit a steady state) appears in [28, Theorem 1] for a system with $s = 2$ servers, building on a straightforward Lyapunov stability argument. The main goal of [28] is to characterize the stationary distribution of the (stable) system via generating functions; an explicit computation of this distribution is provided in [19]. Reference [15] studies a system with finite buffers, and provides closed-form expressions for the loss probabilities. A non-idling version of JSQ was proposed and analyzed in [33] which considers systems with more than one dispatcher, and analyzes how to balance information regarding idle servers among those dispatchers.

There are several papers that study JSQ in asymptotic regimes. In addition to [16], which was discussed above, we mention [23], which identifies a mean-field limit, and shows the chaoticity of the system as N increases. An Ornstein-Uhlenbeck limit for the same model is obtained in [24].

In general, Lyapunov-stability arguments, as in [28], can be hard to generalize to higher-dimensions, because of the need to control the drifts of the process at all states outside some compact subset of the state space. Our proof of Theorem 1 below, that $\rho < 1$ implies that the system is stable for a certain subset of control parameters, is a generalization of [28, Theorem 1], both because it allows any number of servers s, and because it considers a larger family of routing policies, for which JSQ is a special case. In the latter regard, it also generalizes Corollary 1. Our proof is achieved by employing a certain partial-order relation (see Definition 2 in Section 4) in conjunction with a Lyapunov-stability argument.

Power-of-d allocations. The PW(d) policy was first studied in [49] and [37], which also coined the term “supermarket model” to describe a system operating under this control. The supermarket model has since received substantial attention due to its practical and theoretical significance. Both [17] and [11] study the supermarket model in heavy traffic, namely, as the traffic intensity approaches 1. The rate at which the equilibrium distribution of a typical queue converges to the limiting one in the total-variation distance is studied in [35], which also quantifies the chaotic behavior of the system, asymptotically, namely, the rate at which the joint distribution of any fixed number of queues converges to the limiting product-form distribution. We also mention a recent game-theoretic supermarket model in [54], which is analyzed asymptotically, as the number of servers and arrival rate increase to infinity.

It is significant that the asymptotic result regarding the doubly exponential decay rate of the queue size in equilibrium does not necessarily hold for general service-time distributions. Indeed, [6] shows that, for some power-law service-time distributions, the equilibrium queue sizes decay at an exponential, or even polynomial, rate, depending on the power-law exponent and the number of sampled queues d.

In a recent paper [2], the PW(d) policy is studied (together with other policies) in a time-varying setting and with non-homogeneous servers when both the arrival and service rates scale proportionally to n, as $n \to \infty$; in particular, the system need not be in heavy traffic, and the queues may be of fluid scale, at least some of the time. A sufficient condition is given, guaranteeing that the difference between the largest and smallest queue is subdiffusive (namely, is $o(\sqrt{n})$), a phenomenon known in the queueing literature as *state-space collapse* (SSC). (The authors in [2] reserve the term SSC for the heavy-traffic setting, and use the term *subdiffusivity of the deviation process* in their more general setting.) Under this condition, it is proved that PW(d) is asymptotically optimal in the sense that the diffusion-scaled nominal workload process under this policy may be larger than under any other policy by a random quantity that converges to 0 as $n \to \infty$; see Proposition 1 in this reference.

Robustness of Control. The dynamics of a system under a given control are typically studied in idealized settings, which do not fully hold in practice. In particular, even small deviations from the theoretical implementation of a control (due to, e.g., human or measurement errors, discretization of a continuous control process, delays in making or applying a decision, etc.) can in turn lead to substantial perturbations from theoretically predicted performance. Such discrepancies between theory and implementation constitute an important area of research in dynamical control theory (see, e.g., [26, §14] and [30]), but received little attention in the queueing literature. In [42] it is shown how the implementation of a control, that has theoretically desirable performance in a certain asymptotic regime, can lead to chattering of the queue process and, in turn, to *congestion collapse*, namely, to a severe overload that is solely due to the implementation of the control. We refer to [42, Section 9] for a detailed (informal) discussion on how small perturbations from idealized control settings can have substantial impacts on the performance of queueing systems.

Instability of Subcritical Systems. Congestion collapse is related to the more general research area regarding instability of subcritical networks, which initialized with the presentation of the (deterministic) Lu-Kumar network studied in [32], and its stochastic counterpart, the Rybko-Stolyar network [45]; see also [5] and [40] for applications and literature reviews. A non-idling policy is considered in [38], in which an arrival is routed to the queue having the 2nd smallest workload. A sufficient condition for stability, that is strictly stronger than $\rho < 1$, is provided, and it is conjectured that the latter condition is also necessary.

3. The Model

We consider the following class of parallel systems: There are s servers, each having its own infinite buffer for waiting jobs. Jobs arrive to the system following an homogeneous Poisson process with intensity λ, and join one of the servers according to a routing policy from a class of policies that will be formally defined immediately. If the server to which a job is routed is idle, that job enters service immediately; otherwise, it joins the end of the server's dedicated queue, waiting for its turn to be served (there is no jockeying between queues). All jobs are statistically equivalent, requiring i.i.d. service times that are exponentially distributed with mean $1/\mu$, regardless of the server. We let $\rho := \lambda/(s\mu)$ denote the traffic intensity to the system.

Recall that our goal is to study the possible impacts that departure from the idealized modeling assumptions that are taken in the analyses of load-balancing controls has on the systems' load. It is nevertheless analytically convenient to carry-out this study by treating the erroneous execution of the different policies as a control, since this allows us to study the different routing mechanisms (both in the "idealized" and in our "erroneous" settings) simultaneously. In particular, we study a probabilistic routing mechanism which we call a "**p**-allocation policy", where **p** is the *allocation probability vector* $\mathbf{p} = (p_1, p_2, ..., p_s)$. For example, if JSQ is exercised, then the controller sends each new arrival to the shortest queue with probability p_1, to the second shortest queue with probability p_2, and so on. Of course, this routing-with-error mechanism is mathematically equivalent to a controller that routes new arrivals according to the same **p**-allocation vector by choice. With this view, the PW(d) policy, and therefore also JSQ and uniform splitting, becomes a special case of the **p**-allocation policies; see (2)–(4) below.

Specifically, the class of allocation policies we consider depends only on the queue sizes (number of customers in service plus the number of customers waiting in line) of the servers. To determine the server allocations without ambiguity, we assume that the servers are re-labeled as $1, 2, ..., s$ upon each event (arrival or departure), such that $i < j$ if the queue size for server i is no larger than the queue for server j. Servers having the same queue size have consecutive labels; the labeling within each such group of servers can be arbitrary, but for concreteness, we assume that it is made uniformly at random. Therefore, with $Q_i(t)$ denoting the queue size of server i at time $t \geq 0$, the vector $Q(t) := (Q_1(t), ..., Q_s(t))$ is an element of $\mathcal{R}\left(\mathbb{Z}_+^s\right)$. We let $Q_\Sigma(t) = \sum_{i=1}^d Q_i(t)$ denote the total number of customers in the system at time t.

Let Π^s denote the family of probability vectors on $[0,1]^s$, namely, a vector $\mathbf{p} := (p_1, \ldots, p_s)$ is in Π^s if $p_i \in [0,1]$, $1 \leq i \leq s$, and $\sum_{i=1}^s p_i = 1$.

Definition 1. We call a routing policy a **p-allocation policy**, and call **p** the **allocation (probability) vector**, $\mathbf{p} \in \Pi^s$, if, upon arrival, a customer is sent to server i with probability p_i, independently of everything else. A **p**-allocation policy is said to be *non-idling* if an incoming job is routed to an idle server, whenever there is one upon that job's arrival, and is otherwise routed to server i with probability p_i, independently of everything else.

In particular, for each $\mathbf{p}$-allocation policy there is a corresponding non-idling version which uses the same allocation vector to route jobs that arrive when all servers are busy, and otherwise route the arrivals to one of the idle servers.

Observe that if two or more queues have equal size upon an arrival, a $\mathbf{p}$-allocation policy assigns the incoming customer to one of those queues with an equal probability. Indeed, if a customer enters the system at t and the consecutive indices $j, j+1, ..., k-1, k$ are such that $Q_{j-1}(t^-) < Q_j(t^-) = Q_{j+1}(t^-) = ...Q_{k-1}(t-) = Q_k(t-) < Q_{k+1}(t-)$, then by uniformity of the choice of labeling, server ℓ is chosen with the probability

$$\frac{1}{k-j+1}\sum_{i=j}^{k} p_i, \quad \text{for any } \ell \in [\![j,k]\!].$$

A particular class of $\mathbf{p}$-allocation policies is the PW(d) policy, and its special cases, uniform splitting and JSQ.

- For uniform splitting, the allocation vector is

$$\mathbf{p}^{(1)} := (1/s, ..., 1/s)\,. \tag{2}$$

- For JSQ, we have

$$\mathbf{p}^{(s)} := (1, 0, ..., 0). \tag{3}$$

- More generally, under PW(d) an arriving job is routed to server i if it is one of the d draws, and the other $d-1$ servers drawn have indices in $[\![i+1, s]\!]$. Then the allocation vector for this policy is (with ties broken uniformly at random)

$$\mathbf{p}^{(d)} := \left(p_1^{(d)}, ..., p_s^{(d)}\right) = \begin{cases} p_i^{(d)} = \binom{s-i}{d-1}/\binom{s}{d}, & i \in \{1, ..., s-d+1\}; \\ p_i^{(d)} = 0, & i \in \{s-d+2, \dots, s\}, \end{cases} \tag{4}$$

Observe that (2) and (3) are consistent with (4), and are achieved by taking $d = 1$ and $d = s$, respectively.

3.1. The Stability Regions of the Allocation Policies

It is immediate that for any probability vector $\mathbf{p} \in \Pi^s$, the process Q is an $\mathcal{R}\left(\mathbb{Z}_+^s\right)$-valued continuous-time Markov chain (CTMC). The *stability region* of the parallel-server system corresponding to the $\mathbf{p}$-allocation policy, which we denote by $\mathcal{S}(\mathbf{p})$, is then defined as the set of values of the traffic intensity $\rho = \lambda/(s\mu)$ under which Q is stable in the sense that it is a positive recurrent. Then for any $\mathbf{p}$-allocation vector we define

$$\mathcal{S}(\mathbf{p}) := \{\rho \in [0,1) : Q \text{ is positive recurrent under the } \mathbf{p}\text{-allocation policy}\};$$

$$\mathcal{S}^{\text{NI}}(\mathbf{p}) := \{\rho \in [0,1) : Q \text{ is positive recurrent under the } \textbf{non-idling } \mathbf{p}\text{-allocation policy}\}.$$

It is intuitively clear that the stability region under a non-idling $\mathbf{p}$-allocation policy cannot be smaller than the stability region under the same allocation vector when the policy is not non-idling. In other words, we have that

Proposition 1. $\mathcal{S}(\mathbf{p}) \subseteq \mathcal{S}^{\text{NI}}(\mathbf{p})$ *for all* $\mathbf{p} \in \Pi^s$.

The proof of Proposition 1 is given in Appendix A.

As an immediate consequence of Proposition 1 we see that, if stability is proved for given system's parameters and for a specific $\mathbf{p}$-allocation policy (a specific allocation vector $\mathbf{p}$), then the system is also stable under the non-idling version of that policy. On the other hand, a system is unstable if operated under a $\mathbf{p}$-allocation policy, if it is shown to be unstable under its non-idling version.

4. Maximal $\mathbf{p}$-Allocation Policies

In this section we identify a sub-class of $\mathbf{p}$-allocation policies under which the stability region is the interval $[0,1)$. We call such an allocation policy *maximal*, since its stability region is the largest possible. To this end, we introduce the following partial order on $\mathbb{R}_+^s$.

Definition 2. Let $\mathbf{a} = (a_1, ..., a_s)$ and $\mathbf{b} = (b_1, ..., b_s)$ be two elements of $\mathbb{R}_+^s$, $s \geq 1$. We say that $\mathbf{a}$ is smaller than $\mathbf{b}$ in the "generalized Schur-convex" order, and write $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$, if

$$\sum_{i=k}^{s} a_i \leq \sum_{i=k}^{s} b_i \text{ for all } k \leq s.$$

The relation "$\preceq_{\text{GSC}}$" defines a partial ordering on $\mathbb{R}_+^s$ that is a variant (for non-necessarily ordered vectors) of the partial semi-ordering "$\prec_{\text{CX}}$" introduced in Definition 3 of [39], which itself generalizes the well-known Schur-convex partial semi-ordering "$\prec_{\text{SCX}}$" (see e.g. [36]) to vectors of different total sums. Specifically, we have $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$ if and only if $\mathbf{a} \prec_{\text{CX}} \mathbf{b}$ for any $\mathbf{a},\mathbf{b} \in \mathcal{R}\left(\mathbb{R}_+^s\right)$, and $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$ if and only if $\mathbf{a} \prec_{\text{SCX}} \mathbf{b}$ for any $\mathbf{a},\mathbf{b} \in \mathcal{R}\left(\mathbb{R}_+^s\right)$ such that $\parallel \mathbf{a} \parallel = \parallel \mathbf{b} \parallel$. Observe that, for any random variables X and Y having respective probability mass functions $\mathbf{p}_X$ and $\mathbf{p}_Y$ in Π^s and values in $[\![1,s]\!]$, it holds that $X \leq_{st} Y$ if and only if $\mathbf{p}_X \preceq_{\text{GSC}} \mathbf{p}_Y$.

To state and prove the main result of this section, Theorem 1 below, we need the following property of the generalized Schur-convex order. We remark that further properties of this order are proved in Lemma 4 in Appendix A.

Lemma 1. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}_+^s$ *such that* $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$, *and let* $\mathbf{x} \in \mathcal{R}\left(\mathbb{R}_+^s\right)$. *Then,*

$$\mathbf{x} \circ \mathbf{a} \preceq_{\text{GSC}} \mathbf{x} \circ \mathbf{b}.$$

Proof. As $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$ and $\mathbf{x}$ is ordered, we have that, for any $k \leq s$,

$$\begin{aligned}\sum_{i=k}^{s} x_i a_i &= x_k a_k + \sum_{i=k+1}^{s}\sum_{j=k}^{i-1}(x_{j+1}-x_j)a_i + \sum_{i=k+1}^{s} x_k a_i \\ &= x_k \sum_{i=k}^{s} a_i + \sum_{i=k+1}^{s}(x_i - x_{i-1})\sum_{j=i}^{s} a_j \\ &\leq x_k \sum_{i=k}^{s} b_i + \sum_{i=k+1}^{s}(x_i - x_{i-1})\sum_{j=i}^{s} b_j = \sum_{i=k}^{s} x_i b_i. \quad \square\end{aligned}$$

Theorem 1. *If* $\mathbf{p}$ *satisfies*

$$\mathbf{p} \preceq_{\text{GSC}} \mathbf{p}^{(1)}, \tag{5}$$

for $\mathbf{p}^{(1)} = (1/s,...,1/s)$ *in (2), then* $\mathcal{S}(\mathbf{p}) = [0,1)$, *namely, the* $\mathbf{p}$*-allocation policy is maximal.*

Proof. For $n \geq 0$, let T_n denote the nth transition epoch of the CTMC Q, with $T_0 = 0$, and consider the embedded discrete-time Markov chain (DTMC) $\{Q_n : n \geq 0\}$ defined via $Q_n := Q\left(T_n\right)$. We prove the result via a Lyapunov stability argument, employing the Lyapunov function $V : \mathcal{R}\left(\mathbb{Z}_+^s\right) \longrightarrow \mathbb{R}+$ defined by $V(x) = \|\mathbf{x}\|_2^2$. Let

$$\mathcal{K} = \left\{ \mathbf{x} \in \mathcal{R}\left(\mathbb{Z}_+^s\right) \, : \, \sum_{i=1}^{s} x_i \leq \frac{s(\lambda + s\mu)}{2(s\mu - \lambda)} \right\}.$$

Then, for any $n \geq 1$ and $\mathbf{x} = (x_1, ..., x_s) \in \mathcal{K}^c \cap \mathcal{R}\left(\mathbb{Z}_+^s\right)$ we have

$$\begin{aligned}
&\mathbb{E}\left[V\left(Q_{n+1}\right) - V\left(Q_n\right) \mid Q_n = \mathbf{x}\right] \\
&\qquad = \sum_{i=1}^{s} \frac{\lambda}{\lambda + n^+(\mathbf{x})\mu} p_i \left((x_i+1)^2 - (x_i)^2\right) + \sum_{i=1}^{s} \frac{\mu}{\lambda + n^+(\mathbf{x})\mu} \left(((x_i-1)^+)^2 - (x_i)^2\right) \\
&\qquad = \frac{1}{\lambda + n^+(\mathbf{x})\mu} \left(2\left(\lambda \sum_{i=1}^{s} p_i x_i - \mu \sum_{i=1}^{s} x_i\right) + \lambda + n^+(\mathbf{x})\mu\right).
\end{aligned} \tag{6}$$

Applying Lemma 1 with $\mathbf{a} := \mathbf{p}$, $\mathbf{b} := \mathbf{p}^{(1)}$, where $\mathbf{p}^{(1)}$ is the uniform distribution on $[\![1, s]\!]$ in (2), and the ordered vector $\mathbf{x}$, we obtain that $\mathbf{x} \circ \mathbf{p} \preceq_{\text{GSC}} \mathbf{x} \circ \mathbf{p}^{(1)}$, and in particular, that $\sum_{i=1}^{s} p_i x_i \leq \frac{1}{s} \sum_{i=1}^{s} x_i$. As $n^+(\mathbf{x}) \leq s$, this entails that the last expression in (6) is less than or equal to

$$\frac{1}{\lambda + n^+(\mathbf{x})\mu} \left(2\left(\frac{\lambda}{s} - \mu\right) \sum_{i=1}^{s} x_i + \lambda + s\mu\right),$$

which is strictly negative for $\mathbf{x} \notin \mathcal{K}$. In particular, for all $\mathbf{x} = (x_1, ..., x_s) \in \mathcal{K}^c \cap \mathcal{R}\left(\mathbb{Z}_+^s\right)$ and all n,

$$\mathbb{E}\left[V\left(Q_{n+1}\right) - V\left(Q_n\right) \mid Q_n = \mathbf{x}\right] < 0.$$

We deduce from the Lyapunov-Foster Theorem (see, e.g., [10, §5.1]) that the DTMC $\{Q_n : n \geq 1\}$ is positive recurrent. In turn, this implies that the CTMC Q is positive recurrent as well, by Theorem 6.18 in [29], as the rate of the exponentially distributed holding time in each of the states is bounded from below by λ. □

As discussed in Section 2, the maximality of PW(d) follows from (1) which is proved via coupling arguments. Theorem 1 can be used to provide an independent proof of this result.

Corollary 2. *JSQ, uniform splitting, and PW(d), $d \geq 2$, are maximal allocation policies.*

Proof. Recall (2), (3) and (4). As $\mathbf{p}^{(s)} \preceq_{\text{GSC}} \mathbf{p}^{(1)}$ (and $\mathbf{p}^{(1)} \preceq_{\text{GSC}} \mathbf{p}^{(1)}$ by definition), both the JSQ and uniform splitting policies satisfy the assumptions of Theorem 1.

To prove the statement for PW(d) policies, $d \in [\![2, s-1]\!]$, fix such d and observe that, for any $k \leq s - d + 1$, the quantity $\sum_{i=k}^{s} p_i^{(d)}$ is the probability that the d uniformly drawn servers have indices in $[\![k, s]\!]$, which is equal to $\binom{s-k+1}{d} / \binom{s}{d}$. From this, we deduce that

$$\mathbf{p}^{(d)} \preceq_{\text{GSC}} \mathbf{p}^{(2)}. \tag{7}$$

Indeed, for any $k \geq s-d+2$ we have $\sum_{i=k}^{s} p_i^{(d)} = 0$, whereas for any $k \leq s-d+1$, we have that

$$\frac{\sum_{i=k}^{s} p_i^{(d)}}{\sum_{i=k}^{s} p_i^{(2)}} = \frac{\binom{s-k+1}{d}\binom{s}{2}}{\binom{s}{d}\binom{s-k+1}{2}} = \frac{(s-d)...(s-d-k+2)}{(s-2)...(s-2-k+2)} \leq 1,$$

whence (7). Now, $\sum_{i=s}^{s} p_i^{(2)} = 0$ and for all $k \leq s-1$, so that

$$\sum_{i=k}^{s} p_i^{(2)} = \frac{1}{\binom{s}{2}} \sum_{i=k}^{s} (s-i) = \frac{s-k}{s-1}\frac{s-k+1}{s} \leq \frac{s-k+1}{s} = \sum_{i=k}^{s} \frac{1}{s},$$

implying that $\mathbf{p}^{(2)} \preceq_{\text{GSC}} \mathbf{p}^{(1)}$. This, together with (7) and the transitivity of "$\preceq_{\text{GSC}}$", shows that $\mathbf{p}^{(d)} \preceq_{\text{GSC}} \mathbf{p}^{(1)}$. Thus, PW($d$) is maximal by Theorem 1. □

Theorem 1, Corollary 2 and Proposition 1 also imply

COROLLARY 3. $\mathcal{S}^{\text{NI}}(\mathbf{p}) = [0,1)$ *for any* $\mathbf{p}$ *satisfying (5). In particular, the non-idling versions of uniform splitting and PW(d) allocation policies are maximal.*

5. Insufficiency of the Condition $\rho < 1$

Theorem 1 requires, in addition to the usual traffic condition $\rho < 1$, that the allocation probability $\mathbf{p}$ is smaller, in the generalized Schur convex order, than the uniform probability distribution on $[\![1,s]\!]$. We now demonstrate that the latter condition is not futile, and that the traffic condition by itself does not imply stability of a system. To provide simple counter-examples, we consider $\mathbf{p}_{p,2}$-allocation probabilities, with $\mathbf{p}_{p,2} := (1-p, p, 0, ...0)$, for $0 < p < 1$. In other words, any arrival is routed to the shortest queue with probability $q := 1-p$, or to the second-shortest queue with probability p (ties broken by a uniform draw from the relevant queues.) We interpret p as the probability that the controller (or the arriving customer) is making an error in distinguishing between the shortest and the second shortest queue. We denote this $\mathbf{p}_{p,2}$-allocation policy by J2SQ(p), and its corresponding non-idling version by J2SQ$^{\text{NI}}$(p).

Under the non-idling version of the latter policy, the controller identifies idle servers, but otherwise has a probability p of making an error by sending an arrival to the second-shortest queue.

Thus, when all the servers are busy, errors are made according to a Bernoulli trial with a probability p of "success." Observe that, for $\mathbf{p}^{(1)}$ in (2),

$$\mathbf{p}_{p,2} \preceq_{\text{GSC}} \mathbf{p}^{(1)} \quad \text{if and only if} \quad p \leq 1 - 1/s. \tag{8}$$

For a given number of servers $s \geq 1$ and an error probability $p > 0$, let

$$V_{\text{cr}}(p) := \frac{s-1}{2s}\left(1 + \sqrt{1 + \frac{4}{p(s-1)}}\right). \tag{9}$$

We refer to $V_{\text{cr}}(p)$ as the *critical value* (for stability; see Theorem 2 below). Simple algebra shows that

Lemma 2. *For any $s \geq 2$ and any $p \in [0,1]$ we have that*

$$V_{cr}(p) < 1 \quad \textit{if and only if} \quad p > 1 - 1/s. \tag{10}$$

In this case, we have that

$$V_{cr}(p) > \frac{s-1}{sp}. \tag{11}$$

Moreover, $V_{cr}(p)$ is the only positive root of the polynomial $x \mapsto s^2px^2 - s(s-1)px - (s-1)$.

We can now state our main result regarding the insufficiency of the condition $\rho < 1$ to ensure stability.

Theorem 2. $\mathcal{S}^{\text{NI}}(\mathbf{p}_{p,2}) \subseteq [0, V_{cr}(p) \wedge 1)$ *for any* $p \in [0,1]$.

We defer the proof of Theorem 2 to §5.4. In view of (8) and (10), Theorems 1 and 2 immediately imply the following.

Corollary 4. *J2SQ*$^{\text{NI}}(p)$ *is maximal if and only if* $p \leq 1 - 1/s$.

In view of Proposition 1, Corollary 4 implies that the stability region under the $\mathbf{p}_{p,2}$-allocation policy is also characterized by the value of p.

Corollary 5. $\mathcal{S}(\mathbf{p}_{p,2}) \subseteq [0, V_{cr}(p) \wedge 1)$ *for all* $p \in [0,1]$. *In particular J2SQ*(p) *is maximal if and only if* $p \leq 1 - 1/s$.

5.1. $\mathcal{S}^{\mathrm{ni}}(\mathbf{p}_{p,2})$ and $\mathcal{S}(\mathbf{p}_{p,2})$ in Two-Server Systems

A characterization of the stability regions under J2SQ(p) and J2SQ$^{\mathrm{NI}}$(p) is difficult, because it requires controlling the drifts of the multi-dimensional CTMC corresponding to the queue process. However, we can characterize the stability regions of J2SQ(p) and J2SQ$^{\mathrm{NI}}$(p) in the special case $s = 2$. (Observe that in the special case $p = 1$, these policies then correspond to the *join the longest queue* policy when $s = 2$.) Corollaries 4 and 5 imply that both J2SQ$^{\mathrm{NI}}$(p) and J2SQ(p) are maximal if and only if $p \leq 1/2$. The following two propositions, whose proofs are deferred to Appendices B.1 and B.2, characterize the stability regions under these two policies.

Proposition 2. *For $s = 2$, it holds that*

$$\mathcal{S}^{\mathrm{NI}}(\mathbf{p}_{p,2}) = [0, V_{cr}(p) \wedge 1) \text{ for any } p \in [0,1]. \tag{12}$$

Proposition 3. *For $s = 2$, it holds that*

$$\mathcal{S}(\mathbf{p}_{p,2}) = \left[0, \frac{1}{2p} \wedge 1\right) \text{ for any } p \in [0,1].$$

In particular, when $p > 1/2$, a system operating under J2SQ$^{\mathrm{NI}}$(p) is stable if and only if $\rho < V_{\mathrm{cr}}(p) < 1$, and a system operating under J2SQ(p) is stable if and only if $\rho < 1/(2p) < 1$. It is easily checked that, in this case, $V_{\mathrm{cr}}(p) < 1/(2p)$, so that $\mathcal{S}^{\mathrm{NI}}(\mathbf{p}_{p,2}) \subsetneq \mathcal{S}(\mathbf{p}_{p,2})$. Thus, the containment in Proposition 1 cannot be replaced with an equality.

5.2. Join the 2nd Shortest Queue Allocation Policy

The proof of Theorem 2 involves some technical details that obscure the main intuition for the instability whenever the error probability p is greater than $1 - 1/s$. Simplicity is achieved by considering the special case $p = 1$, which is tantamount to having the allocation vector be $\mathbf{p}_{1,2} := (0, 1, 0, ..., 0)$. In this case, the routing policy is simply *join the second shortest queue*, which we denote by J2SQ; we denote its non-idling version by J2SQ$^{\mathrm{NI}}$. (As was mentioned above, this latter policy is also the join-the-longest-queue policy in the spacial case $s = 2$.) It follows from (10) that $V_{\mathrm{cr}}(1)$, defined in (9) with $p = 1$, satisfies $V_{\mathrm{cr}}(1) < 1$.

Proposition 4. $\mathcal{S}^{\text{NI}}(\mathbf{p}_{1,2}) \subset [0, V_{cr}(1))$. *In particular, J2SQ*$^{\text{NI}}$ *is non-maximal.*

Proof. Let

$$\mathcal{A} := \{x \in \mathbb{Z}_+^s : x_1 \in \{0,1\},\ x_i \geq 2,\ i \in [\![2,s]\!]\}, \tag{13}$$

and note that $Q \in \mathcal{A}$ if and only if queue 1 (the smallest queue) has no jobs waiting for service, whereas queue 2 (and thus all other queues) have waiting jobs.

Let $\mathbf{s} := (0,2,\dots,2) \in \mathcal{A}$, and for $k = 1,2,\dots$, define the time $t_k := \inf\{t \geq 0 : Q(t) = \mathbf{s}\}$, where the event $\{t_m = \infty\}$ for some $m \geq 1$ (and then for all $k \geq m$) may have a positive probability. We say that the kth *visit* (to $\mathcal{A}$) begins at time t_k and ends when Q exits the set $\mathcal{A}$, namely, at a random time $t_k + T_k$ such that $Q((t_k + T_k)-) \in \mathcal{A}$ and $Q(t_k + T_k) \notin \mathcal{A}$. We henceforth refer to T_k as the length of the kth visit.

We prove the result by making the contradictory assumption that Q is positive recurrent, and thus ergodic. Under this ergodicity assumption, $P(t_k < \infty) = 1$ for all $k \geq 1$, and the lengths of the visits $\{T_k : k \geq 1\}$ are i.i.d. by virtue of the strong Markov property, with $P(0 < T_1 < \infty) = 1$ and $E[T_1] < \infty$. Now, during the kth visit, namely, during the intervals $I_k := [t_k, t_k + T_k)$, the (ordered) queue process Q operates as follows: Any arrival is routed to server 1, if this server is idle. Otherwise, the arrival is routed to server 2. Hence, over each interval I_k, we can view server 1 as a single-server loss system (to which we refer as the *front server*), with the overflow from this front server constituting the arrival process to a system with $s-1$ homogeneous servers operating under the JSQ routing policy (to which we refer as the *back servers*).

If the first arrival during the kth visit finds the system in state $\mathbf{s}$, then that arrival is routed to server 1 (which is idle). Let A_k denote this latter event: with a_k denoting the time of the first arrival after time t_k, $A_k := \{Q(a_k-) = \mathbf{s}\}$. By the strong Markov property, the events $A_1, A_2, \dots$ are independent and have the same probability, and it clearly holds that $P(A_1) > 0$.

By Lemma 5 in Appendix C, the first arrival to a single-server loss system puts this system in steady state. In particular, on $[a_1, t_1 + T_1)$ the instantaneous probability that an arrival finds server 1 busy, and is therefore "overflowed" to the back system, is $\lambda/(\lambda + \mu)$. Thus, due to the

PASTA (Poisson Arrivals See Time Average) property, the "arrival rate" to the back servers during $[a_1, t_1 + T_1)$ is $\alpha := \lambda^2/(\lambda+\mu)$. It follows that the process $Q_{-1} := (Q_2, ..., Q_s)$ coincides in distribution with the ordered queue-length process of a JSQ system with $s-1$ servers and arrival rate α.

Next, observe that $V_{\text{cr}}(1) < 1$ by (10), and that $V_{\text{cr}}(1)$ is thus the only positive root of the polynomial $x \mapsto s^2x^2 - (s-1)sx - (s-1)$. It then readily follows that, for any $\rho > 0$,

$$\frac{(s\rho)^2}{1+s\rho} > (s-1) \quad \text{if and only if} \quad \rho > V_{\text{cr}}(1). \tag{14}$$

Therefore, if $\rho = \lambda/s\mu > V_{\text{cr}}(1)$, then $\alpha > (s-1)\mu$, and so the probability that the process Q_{-1} will never reach a state in which the smallest of the $s-1$ queues is equal to 1 is strictly positive, implying that $P(T_1 = \infty) > 0$. If $\alpha = (s-1)\mu$ (so that $\rho = V_{\text{cr}}(1)$), then Q_{-1} is null recurrent, and the expected time until a state with the smallest queue being 1 is reached is infinite. In either case, the expected length of a visit is infinite, namely, $E[I_1] = E[T_1] = \infty$, in contradiction to the assumed ergodicity of Q. □

The proof of Proposition 4 makes the reason for the instability of the system we consider apparent: Eventually, the system must split into a front loss single-server system whose overflow process constitutes the arrival process to a back $(s-1)$ parallel-server system operating under the JSQ policy. If the overflow process is larger than the service capacity of the "back servers", then the system as a whole is unstable, because the expected time for it to exit this split structure is infinite. In particular, once the system splits, the expected time until Q reaches states that are not in the set $\mathcal{A}$ defined in (13) is infinite. In fact, the regenerative structure of Q implies that, if the traffic intensity is *strictly larger* than the critical value, i.e., if $\rho > V_{\text{cr}}(p, s)$, then $P(T_k = \infty \text{ for some } k \geq 1) = 1$ and $\|Q(t)\| \longrightarrow \infty$ w.p.1 as $t \to \infty$.

Remark 1. We note that the (in)stability of the back system is solely determined by the arrival rate to that system and mean service time μ, and is independent of any other distributional assumptions; in particular, it does not rely on the service time distribution. Furthermore, the blocking probability of a loss system is insensitive to the service-time distribution, so that the overflow rate from the front server *at stationarity* is $\alpha = \lambda^2/(\lambda+\mu)$ regardless of the assumption that service times are exponentially distributed. Thus, a generalization of Proposition 4 can be proved for a system with general service time distributions having a finite mean μ.

5.3. Join the m-Shortest Queue Allocation Policy

The arguments in the proof of Proposition 4 can be easily extended to the case in which there are several "front servers" instead of just one such server, a scenario which arises when the p-allocation policy follows the "join the mth shortest queue" assignment rule, corresponding to the allocation vector $\mathbf{p}_{1,m} = (0,...,0,\underbrace{1}_{m},0,...,0)$. Under this allocation policy, which we denote by JmSQ, an incoming customer is routed to the mth shortest queue $(2 \le m \le s)$ with probability 1. The non-idling version of this policy is denoted by JmSQ$^{\text{NI}}$.

For $m \in [\![2,s]\!]$, define

$$\mathscr{G}(m) := \left\{ \rho \in (0,1) : \frac{s\rho\,(s\rho)^{m-1}/(m-1)!}{\sum_{i=0}^{m-1}(s\rho)^i/i!} < (s-m+1) \right\}; \tag{15}$$

$$V_{\text{cr}}(1,m) := \sup \mathscr{G}(m). \tag{16}$$

Note that the set $\mathscr{G}(m)$ is not empty, since it contains all the positive numbers that are smaller than $(s-m+1)/s$. In particular, $V_{\text{cr}}(1,m)$ is finite. Further, the inequality in the definition of $\mathscr{G}(m)$ reduces to (14) when $m=2$, so that $V_{\text{cr}}(1,2) \equiv V_{\text{cr}}(1)$, for $V_{\text{cr}}(1)$ in (9).

LEMMA 3. *$V_{cr}(1,m) < 1$ for all $m \in [\![2,s]\!]$.*

The proof of Lemma 3 appears in Appendix B.3. Given Lemma 3, the following result generalizes Proposition 4.

PROPOSITION 5. *$\mathcal{S}^{\text{NI}}(\mathbf{p}_{1,m}) \subset [0, V_{cr}(1,m))$; In particular, JmSQ$^{\text{NI}}$ is non-maximal.*

Proof. Fix $m \in [\![2,s]\!]$ and let

$$\mathcal{A}_m := \{x \in \mathbb{Z}_+^s : x_i \in \{0,1\},\ i \in [\![1,m-1]\!],\quad \text{and } x_j \ge 2,\ j \in [\![m,s]\!]\}.$$

As in the proof of Proposition 4, the statistical homogeneity of the s servers implies that any vector $\mathbf{x} \in \mathbb{Z}_+^s$ that has exactly $m-1$ coordinates with values in $\{0,1\}$ can be considered in $\mathcal{A}_m$ since $\mathcal{R}(\mathbf{x}) \in \mathcal{A}_m$. Further, as long as the system is in $\mathcal{A}_m$, it is essentially split into two systems: the first $m-1$ servers operate like an $M/M/(m-1)$ loss system, and the remaining $s-m+1$ servers operate like a parallel system under the JSQ routing policy, whose arrival process is the overflow from the first $m-1$ "front servers." Let $\mathbf{s} = \left(\underbrace{0,\dots,0}_{m-1},\underbrace{2,\dots,2}_{s-m+1}\right)$. We say that a *visit* begins when the system transitions into state $\mathbf{s}$, and ends when it exists the set $\mathcal{A}_m$, namely, when the splitting into a front and back servers ends.

Let $L_m := \{L_m(t) : t \ge 0\}$ denote the number-in-system process in the $M/M/(m-1)$ loss system, and let $L_m(\infty)$ denote a random variable having the stationary distribution of L, which we denote by π_m, i.e., $\pi_m(j) := P(L_m(\infty) = j)$. Note that, during a visit, the number of busy servers in the aforementioned $m-1$ front-servers is distributed like L_m. By Lemma 6 in Appendix C, there exists a random time τ, such that $L_m(t) \overset{\mathrm{d}}{=} L_m(\infty)$ for all $t \ge \tau$, and therefore, the number of busy servers among those front servers is also distributed like $L_m(\infty)$ for all $t \ge \tau_k$ on the event $E_k := \{\tau_k < T_k\}$, where T_k denotes the length of the kth visit, and $\{\tau_k : k \ge 1\}$ are i.i.d. with $\tau_1 \overset{\mathrm{d}}{=} \tau$. By the strong Markov property, all the visits are i.i.d. and $P(E_1) > 0$. Therefore, $\{E_k : k \ge 1\}$ must occur infinitely often, unless one of the visits is infinite, i.e., finitely-many E_k's will occur if and only if $T_k = \infty$, for some $k \ge 1$.

Now, if E_k occurs for the kth visit, then the overflow process from the front servers, which is the arrival process into the back servers, has rate $\lambda\pi_m(m-1)$ after time τ_k, due to PASTA. If $\rho \ge V_{\mathrm{cr}}(1,m)$, then $\lambda\pi^m(m-1) \ge \mu(s-m+1)$, i.e. the arrival rate to the "back servers" is larger than the maximum total service rate of those $s-m+1$ servers after time τ_k as long as the kth visit is in process. Therefore, $P(T_k = \infty) > 0$ on the event E_k. We conclude that

$$P(T_k = \infty \text{ for some } k \ge 1) = 1,$$

so that Q is either transient or null recurrent. □

5.4. Proof of Theorem 2

The proofs of Propositions 4 and 5 build on the fact that each time a splitting of the system occurs, the front "loss system" has a positive probability of reaching stationarity in finite time, after which PASTA is employed to characterize the overflow rate into the "back servers." In the setting of Theorem 2 with $p<1$ the splitting is as follows: There is one "front server" and $s-1$ "back servers", as in the proof of Proposition 4. However, the front server does not operate as a loss system. Instead, during each "visit" (splitting event), the front server operates as an $M/M/1$ queue with an infinite buffer, having a Poisson arrival process with rate λ. Each arrival to this $M/M/1$ queue enters service if the server is idle, and otherwise joins its queue with probability $1-p$, and the back servers with probability p, independently of everything else. In particular, the arrival process to the $s-1$ back servers constitutes all the arrival who did not join the front server. For the particular $M/M/1$ queue we obtain during a splitting event, the time to reach stationarity is infinite, so that PASTA cannot be directly employed as in the proofs of Propositions 4 and 5.

Proof of Theorem 2. Consider $p\in(1-1/s,1]$, and fix λ,μ such that $\rho=\lambda/s\mu\in[V_{\mathrm{cr}}(p,s),1)$. Let $Y^{\mathrm{F}}(t)\in\mathbb{Z}_+$ be the number of customers in the front server at time t, and for $i\in[\![1,s-1]\!]$, let $Y_i^{\mathrm{NI}}(t)$ be the size of the ith queue among the back servers, in the increasing order of queue lengths. It is easily seen that both processes Y^{F} and $Y:=\left(Y^{\mathrm{F}},Y_1^{\mathrm{B}},...,Y_{s-1}^{\mathrm{B}}\right)$ (as functions of t) are CTMCs on $\mathbb{Z}_+$ and $\mathbb{Z}_+^{s-1}$, respectively. In particular, Y^{F} is a Birth and Death (BD) process on $\mathbb{Z}_+$ with respective birth and death rates λ and 0 at state 0, and $\lambda(1-p)$ and μ at all other states. By the assumed values of p and ρ, Y^{F} is ergodic with stationary distribution

$$\begin{aligned}\pi^{\mathrm{F}}(0)&=\frac{\mu-\lambda+\lambda p}{\mu+\lambda p};\\ \pi^{\mathrm{F}}(i)&=\left(\frac{\lambda(1-p)}{\mu}\right)^{i-1}\frac{\lambda}{\mu}\pi^{\mathrm{F}}(0),\, i\geq 2.\end{aligned}$$

In particular the stationary probability that the front server is busy is

$$\pi^{\mathrm{F}}\left(\mathbb{Z}_+^*\right)=1-\pi^{\mathrm{F}}(0)=\frac{\lambda}{\mu+\lambda p}=\frac{s\rho}{1+s\rho p}. \tag{17}$$

Now, it is well-known that an ergodic BD process with birth and death rates that are uniformly bounded is exponentially ergodic; e.g., see [48, §4]. Then letting $\|\cdot\|_{TV}$ denote the total-variation norm (e.g., see [1]),

$$\|P(Y^{\mathrm{F}}(t)\in\cdot)-\pi(\cdot)\|_{TV}<C_0e^{-\beta t},\quad t\geq 0, \tag{18}$$

for some $C_0\in[0,\infty)$ that depends on the initial condition only, and for some $\beta>0$ that is independent of the initial condition.

For a given $y\in\mathbb{Z}_+$, Let P_t^y denote the one-dimensional marginal distribution of the random variable $Y^{\mathrm{F}}(t)$ when $Y^{\mathrm{F}}(0)=y$. It follows from (18) that, for any $\epsilon>0$, there exists a $T_\epsilon^y<\infty$ that depends on the initial condition y, such that

$$\|P_t^y-\pi^{\mathrm{F}}\|_{TV}<\epsilon\quad\text{for all } t>T_\epsilon^y. \tag{19}$$

Consider the $\mathbb{Z}_+^2$-CTMC $X(t):=\{(Y^{\mathrm{F}}(t),N_p(t)):t\geq 0\}$, where N_p is a Bernoulli splitting of the Poisson arrival process to the system. In particular, each arrival to the system is an event in N_p with probability p, independently of all other events and of time. Next, define $f:\mathbb{Z}_+^2\times\mathbb{Z}_+^2\longrightarrow\mathbb{R}$ via

$$f((i,j),(i',j')):=\mathbf{1}_{\{i>0,i=i',j=j+1\}}. \tag{20}$$

It follows from Lévy's formula (e.g., Equation (2.2) in [9, p.5]) that, for f in (20),

$$E\left[\sum_{s\leq u\leq t}f(X(u-),X(u))\right]=\lambda pE\left[\int_s^t\mathbf{1}_{\{Y^{\mathrm{F}}(u)>0\}}\right]. \tag{21}$$

Now, As in (14), one can easily check that $\rho>V_{\mathrm{cr}}(p,2)$ if and only if $\lambda p\pi^{\mathrm{F}}(\mathbb{Z}_+^*)>(s-1)\mu$, so that we can take $\epsilon>0$ for which $\lambda p\left(\pi^{\mathrm{F}}\left(\mathbb{Z}_+^*\right)-\epsilon\right)>(s-1)\mu$. Let $N_{\mathrm{OF}}(a,b]$ denoting the overflow process from the front server (which is the arrival process to the back servers) over the time interval $(a,b]$, $0\leq a<b$. Then (19) and (21) imply that, for f in (20) and for all $t>0$,

$$t^{-1}E\left[N_{\mathrm{OF}}(T_\epsilon,T_\epsilon+t]\right]=t^{-1}E\left[\sum_{T_\epsilon\leq u\leq T_\epsilon+t}f(X(u-),X(u))\right]=\lambda p\int_{T_\epsilon}^{T_\epsilon+t}P_u^y du>\lambda p\left(\pi^{\mathrm{F}}\left(\mathbb{Z}_+^*\right)-\epsilon\right). \tag{22}$$

The rest of the proof is similar to the arguments in the proof of Proposition 4: Taking the (contradictory) assumption that Q is ergodic, a splitting to a forward and backward servers must occur infinitely often. Letting a visit begin when, during such a splitting, the front server first reaches the empty state, we have that the visits are i.i.d. and each lasts for at least T_ϵ time units with a strictly positive probability, for any ϵ satisfying the inequality in (22). (Note that, since a visit begins at a fixed state, we can choose the same T_ϵ in (19) for all the visits.) More specifically, with I_k denoting the time interval during the kth visit beginning when the front server is empty and ending when the visit ends, we have that $P(I_k > T_\epsilon) > 0$, so that $\{I_k > T_\epsilon\}$, $k \geq 1$, must occur i.o. However, since the overflow process from the front server is guaranteed to be larger than the total service rate $\mu(s-1)$ of the back servers after time T_ϵ, there is a positive probability that a visit will never end, contradicting the ergodicity assumption. The proposition is proved. □

6. Simulation Experiments for Workload-Based Allocation Policies

As discussed in Section 1.1, our results and analyses provide insights for systems operating under allocation policies that are based on the workload (as opposed to the queue length). Indeed, it is intuitively clear from the proofs of our main results that a system under JSW also experiences random "splitting" into front and back subsystems, and that the back subsystem may be unstable (so that the whole system is unstable) even if $\rho < 1$. In this section we present simulation experiments to support this intuition. In fact, the simulations indicate that the bounds we obtained for the stability regions in Theorem 2 and Propositions 4 and 5, are tight estimates of the stability regions for the corresponding workload-based allocation policies, which are formally defined below.

Fix an integer $m \in [\![2, s]\!]$, and Let $W(t) := (W_1(t), \ldots, W_s(t))$, $t \geq 0$, denote the *ordered* workload process, namely, $W_i(t)$ is the workload at time t at queue i, $1 \leq i \leq s$, and $W_1(t) \leq W_2(t) \leq \cdots \leq W_s(t)$. For $m \in [\![1, s]\!]$ and $p \in [0, 1]$, we say that the allocation policy is *Join the mth shortest workload with probability* p, denoted by JmSW(p), if each arrival is sent to the queue having the smallest workload with probability $1 - p$ (i.e., to the server having workload $W_1(t)$ at the arrival time t), and is otherwise sent to the queue with the mth smallest workload (i.e., to the server having workload $W_m(t)$) with probability p. In the non-idling version of JmSW(p), denoted by JmSW$^{\text{NI}}$(p), an arrival is sent to an idle server w.p.1, if such a server is available, and is otherwise routed to a server according to JmSW(p).

Cases Considered. We simulated a system with 4 servers, each providing exponentially distributed service with mean 1, that is operating under J2SW$^{\text{NI}}$(p) (join the second-smallest workload with probability p), where $p \in \{0.8, 0.9, 1\}$. In addition, we simulated the system when it is operating under J3SW$^{\text{NI}}$(1), namely, $m = 3$ and $p = 1$. For each of these four systems we simulated the corresponding embedded DTMC over 10^7 arrivals for two values of the traffic intensity ρ, one that is slightly above, and the other slightly below, the critical values $V_{\text{cr}}(p)$ (for J2SW$^{\text{NI}}$(p)) and $V_{\text{cr}}(1,3)$ (for the system under J3SW$^{\text{NI}}$(1)). The critical values are computed via (9) and (15)–(16), respectively. In particular, for each of the four examples we considered a traffic intensity that is larger than the critical value of ρ by $2/10^3 = 0.002$, and a traffic intensity that is smaller than the corresponding critical value by 0.002. We emphasize that the critical values are for the same system operating under J2SQ$^{\text{NI}}$(p) and J3SW$^{\text{NI}}$(1), and so we do not know whether they are also the critical values for the system under the simulated scenarios.

In Figure 1 we show a sample path of the most loaded server (in terms of workload) for each of the six cases considered for the system under J2SW$^{\text{NI}}$(p), namely, two examples, each with a different ρ for each of the three different values of p, as described above. Two sample paths simulated for the system operating under J3SW$^{\text{NI}}$(1), one for each value of ρ, are shown in Figure 2.

We remark that, whenever ρ is equal to its critical value, the queue process is null recurrent, and it is therefore hard to determine from simulation whether a system is stable when ρ is "too close" to its critical value. (For any value of ρ in a small-enough neighborhood of the critical value, the stochastic fluctuations are large, and one may observe a return to the empty state over any finite time interval, even in the transient case.) Nevertheless, for each of the four simulated routing policies, the system seems to be unstable for the larger value of ρ, and to be stable for the smaller value of ρ. This, together with the fact that the difference between the two traffic intensities is

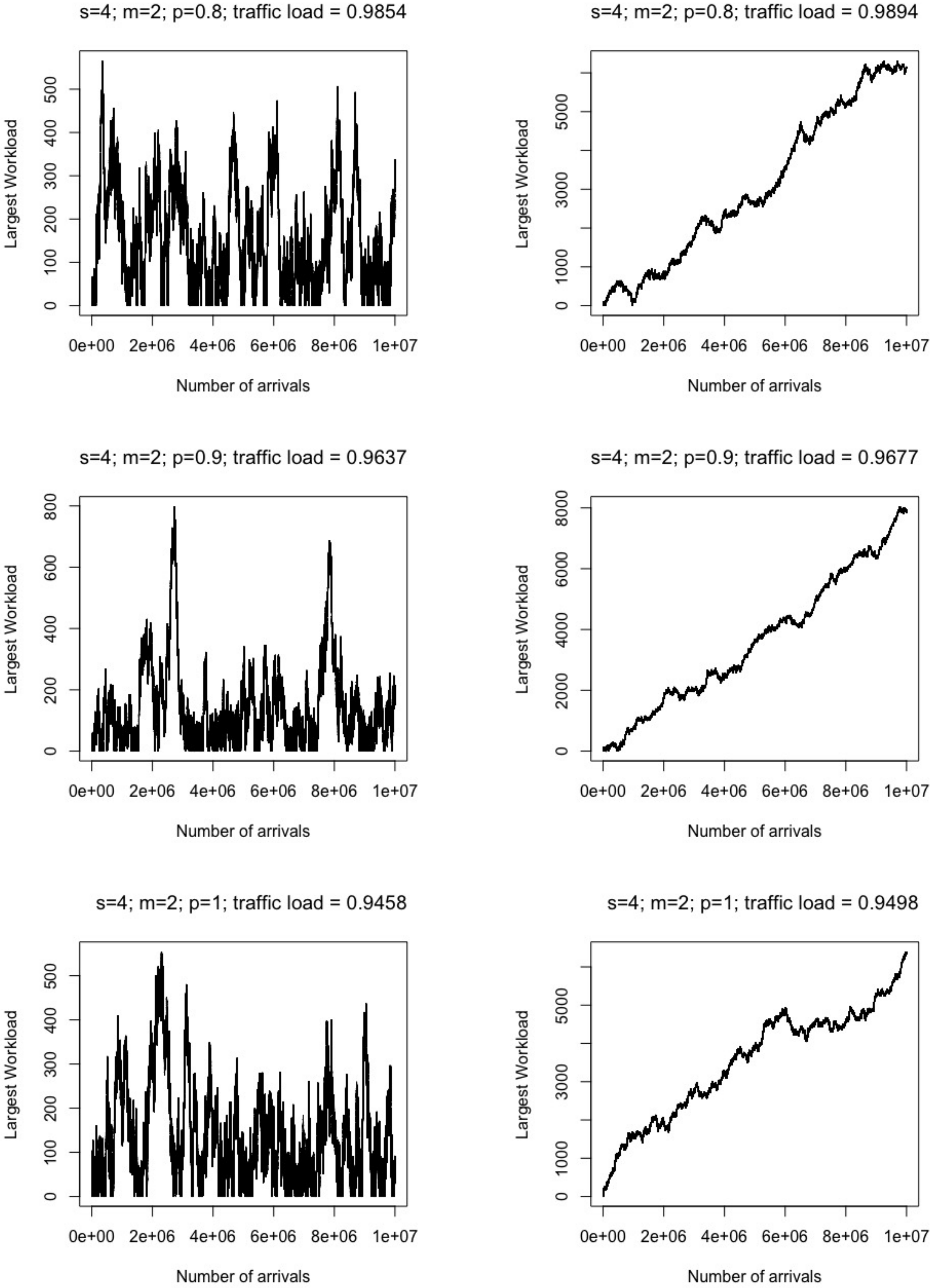

Figure 1 Sample paths of the largest workload process generated for 10^7 arrivals of a system with four servers operating under $\mathrm{J2SW}^{\mathrm{NI}}(p)$. The two figures in each row depict one value of p, with the left figure having $\rho = V_{\mathrm{cr}}(p) + 0.002$, and the right figure having $\rho = V_{\mathrm{cr}}(p) - 0.002$. **Upper panel:** a system operating under $\mathrm{J2SW}^{\mathrm{NI}}(0.8)$, for which $V_{\mathrm{cr}}(0.8) \approx 0.9874$. **Middle panel:** a system operating under $\mathrm{J2SW}^{\mathrm{NI}}(0.9)$, for which $V_{\mathrm{cr}}(0.9) \approx 0.9657$. **Lower panel:** a system operating under $\mathrm{J2SW}^{\mathrm{NI}}(1)$, for which $V_{\mathrm{cr}}(1) \approx 0.9478$.

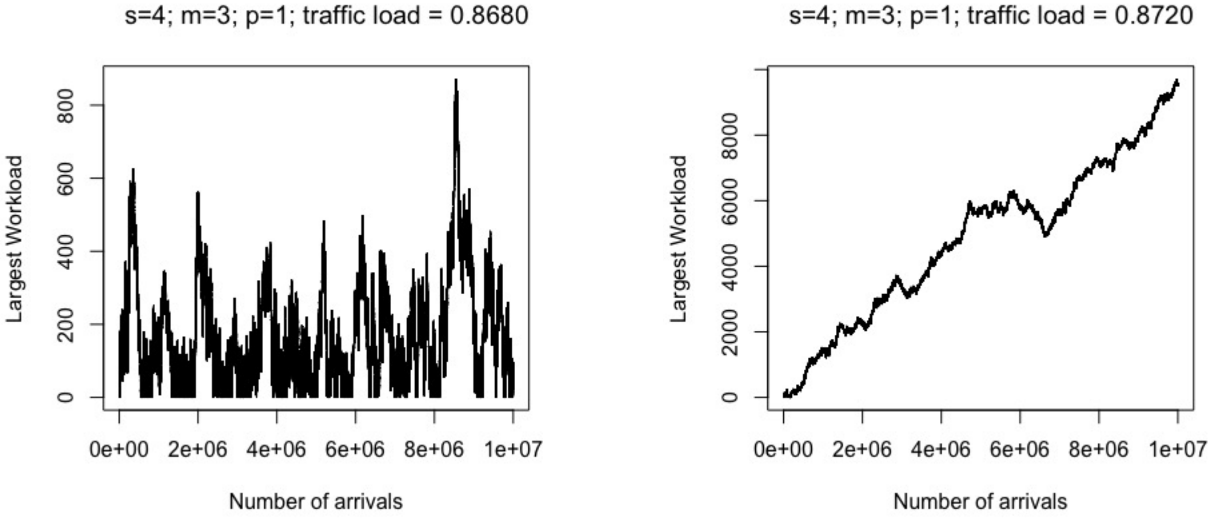

Figure 2 Sample paths of the largest workload process generated for 10^7 arrivals of a system with four servers operating under J3SW$^{\text{NI}}(1)$, for which $V_{\text{cr}}(1,3) = 0.87$. The left figure depicts a sample path when $\rho = V_{\text{cr}}(1,3) - 0.002$, and the right figure depicts a sample path when $\rho = V_{\text{cr}}(1,3) + 0.02$.

just 0.004, suggests that the critical value of ρ for the system operating under the queue-based allocation policy is very close (and may be equal) to critical value of ρ for the system operating under the corresponding workload-based allocation policy.

7. Summary

Stability of a queueing system is the cruder performance measure and it is therefore among the simplest performance measures to characterize. On the other hand, while more refined performance measures, such as those corresponding to the queue length and waiting times, can be effectively estimated via simulation, estimating the stability region of a stochastic system via simulation is difficult, even for fixed parameters of the system's primitives. (Of course, estimating the stability region of a system *as a function of these parameters* is clearly harder.)

In this paper we considered the (in)stability problem of parallel server systems with $s > 1$ statistically homogeneous servers, to which jobs are routed upon arrival according to a family of random-assignment rules, which we named **p**-allocation policies. That family of policies includes the PW(d) routing rule, and its special cases JSQ and uniform routing, as well as their "non-idling" versions, under which an arrival is always routed to an idle server, if one is available at that arrival time. Our motivation for this study was the fact that in practice, and unlike the ideal settings that are typically considered in the literature, routing errors are likely to occur, so that jobs are not necessarily routed to the shortest among the relevant queues.

We started by characterizing a sufficient condition for stability (Theorem 1) which, in addition to the usual traffic condition $\rho < 1$, requires the $\mathbf{p}$-allocation vector to be smaller, in the generalized Schur convex order, than the uniform distribution on $[\![1, s]\!]$. In particular, under this latter assumption on $\mathbf{p}$, the $\mathbf{p}$-allocation policy (and its non-idling version) is guaranteed to be maximal.

We then demonstrated that the condition $\rho < 1$ by itself does not guarantee that the system is stable, even when a non-idling $\mathbf{p}$-allocation policy is employed. Specifically, we considered the stability region of the policy J2SQ$^{\text{NI}}(p)$, under which arrivals are always routed to an idle server, if one is present, and are otherwise routed to the shortest queue with probability $1 - p$, and to the second shortest queue with an "error probability" p. Theorem 2 proves that ρ must be smaller than a positive number V_{cr}, which is strictly smaller than 1 for a range of values of p, implying that the stability region under the control may be strictly contained in $[0, 1)$. Corollary 5 proves that p must satisfy $p \leq 1 - 1/s$ in order for J2SQ$^{\text{NI}}(p)$ to be maximal.

Finally, simulation examples in §6 demonstrate that our results are insightful also for systems operating under JSW, for which routing errors are more likely to occur, even in automated environments, because the actual workload in each queue can typically only be estimated. We conjecture that the stability regions under JSQ and JSW are the same.

Further Implications of the Results. The fact that the $\mathbf{p}$-allocation policy may not be maximal has important implications well beyond the possibility of experiencing congestion collapse. Indeed, even though the risk of instability caused by erroneous routing decisions is small when the probability of making an error is small, or when the number of servers is large, routing errors cause any system to effectively be in "heavier traffic" than planned. Thus, if the system is designed to operate in heavy traffic, so that $\rho \approx 1$, even a small probability of making routing errors may lead to harmful departures from the desired performance, and may even lead to instability. In particular, SSC as in [2] and [43], may not hold asymptotically, even if it should hold under idealized modeling assumptions (that ignore erroneous routing decisions). As a result, the goal of balancing the load among the servers may not be achieved, even if the system is stable. We again refer to Sections 1 and 9 in [42] for a general discussion on congestion collapse caused by SSC-inducing controls.

Acknowledgements

We thank Area Editor Amy Ward, AE and two anonymous referees for carefully reading the paper and making many useful comments and suggestions that helped improve the paper. In particular, the use of Lévy's formula in the proof of Theorem 2 was proposed by one of the referees. Ohad Perry received support from National Science Foundation (NSF) Grant CMMI 1763100.

Appendix.

The appendix is organized as follows: We prove Proposition 1 in §A, after establishing several properties of the generalized Schur-convex ordering in Lemma 4 below. We prove the Remaining results from Section 5—Propositions 2 and 3, and Lemma 3— in §B. Finally, we state and prove two auxiliary results in §C.

A. Proof of Proposition 1

In this section we prove Proposition 1, building on the next lemma.

LEMMA 4. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors of* $\mathcal{R}\left(\mathbb{R}_+^s\right) \cap \mathbb{Z}_+^s$ *be such that* $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$. *Then,*

1. *for any* $i' \leq i \leq s$ *we have that*

$$\mathcal{R}\left(\mathbf{a} + \boldsymbol{e}_{i'}\right) \preceq_{\text{GSC}} \mathcal{R}\left(\mathbf{b} + \boldsymbol{e}_i\right);$$

2. *for any* $i \leq s$ *such that* $a_i \geq 1$ *and* $b_i \geq 1$, *we have that*

$$\mathcal{R}\left(\mathbf{a} - \boldsymbol{e}_i\right) \preceq_{\text{GSC}} \mathcal{R}\left(\mathbf{b} - \boldsymbol{e}_i\right);$$

3. *for any* $i \leq s$ *such that* $a_i = 0$ *and* $b_i > 0$,

$$\mathbf{a} \preceq_{\text{GSC}} \mathcal{R}\left(\mathbf{b} - \boldsymbol{e}_i\right).$$

Proof. The proof is reminiscent of the arguments in the proofs of Lemma 2 in [39] and Lemma A.15 in [14]. Fix $\mathbf{a}$ and $\mathbf{b} \in \mathbb{Z}_+^s \cap \mathcal{R}\left(\mathbb{R}_+^s\right)$ such that $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$.

1. Fix $i' \le i \le s$, and let $\tilde{\mathbf{a}} := \mathcal{R}(\mathbf{a} + \mathbf{e}_{i'})$ and $\tilde{\mathbf{b}} := \mathcal{R}(\mathbf{b} + \mathbf{e}_i)$. Fix $k \le s$. We need to show that $\sum_{j=k}^{s} \tilde{a}_j \le \sum_{j=k}^{s} \tilde{b}_j$ for all $k \le s$. The key relations are the following: for all $k \le s$,

$$\sum_{j=k}^{s} \tilde{a}_j = \left(\sum_{j=k}^{s} a_j\right) \vee \left(a_{i'} + 1 + \sum_{j=k+1}^{s} a_j\right); \tag{23}$$

$$\sum_{j=k}^{s} \tilde{b}_j = \left(\sum_{j=k}^{s} b_j\right) \vee \left(b_i + 1 + \sum_{j=k+1}^{s} b_j\right). \tag{24}$$

Consequently, for any $k \le i'$, we get that

$$\sum_{j=k}^{s} \tilde{a}_j = \sum_{j=k}^{s} a_j + 1 \le \sum_{j=k}^{s} b_j + 1 = \sum_{j=k}^{s} \tilde{b}_j,$$

and whenever $i' < i$, for any $i' < k \le i$, we obtain that

$$\sum_{j=k}^{s} \tilde{a}_j \le \sum_{j=k}^{s} a_j + 1 \le \sum_{j=k}^{s} b_j + 1 = \sum_{j=k}^{s} \tilde{b}_j.$$

Only the case where $k > i$ remains to be treated. We have the following alternatives:

(i) if $a_{i'} < a_k$, then it immediately follows from (23)–(24) that

$$\sum_{j=k}^{s} \tilde{a}_j = \sum_{j=k}^{s} a_j \le \sum_{j=k}^{s} b_j \le \sum_{j=k}^{s} \tilde{b}_j.$$

(ii) if $a_{i'} = a_k$, we have three sub-cases:

(iia) If there exists $\ell \in [\![i', k]\!]$ such that $b_\ell > a_\ell = a_k = a_{i'}$ (take the first such ℓ in increasing order), then (23)–(24) imply that

$$\sum_{j=k}^{s} \tilde{a}_j = a_{i'} + 1 + \sum_{j=k+1}^{s} a_j \le b_\ell + \sum_{j=k+1}^{s} b_j \le \sum_{j=k}^{s} b_j \le \sum_{j=k}^{s} \tilde{b}_j.$$

(iib) If $a_j \ge b_j$ for all $j \in [\![i', k]\!]$, and there exists $\ell \in [\![i', k]\!]$ such that $b_\ell < a_\ell = a_k$, then, we have that $\sum_{j=\ell}^{k-1} a_j > \sum_{j=\ell}^{k-1} b_j$ and thus, as $\sum_{j=\ell}^{s} a_j \le \sum_{j=\ell}^{s} b_j$, we must have that $\sum_{j=k}^{s} a_j < \sum_{j=k}^{s} b_j$. Therefore,

$$\sum_{j=k}^{s} \tilde{a}_j = a_{i'} + 1 + \sum_{j=k+1}^{s} a_j \le \sum_{j=k}^{s} a_j + 1 \le \sum_{j=k}^{s} b_j \le \sum_{j=k}^{s} \tilde{b}_j.$$

(iic) If $a_j = b_j$ for all $j \in [\![i', k]\!]$, then $b_i = b_k = a_k = a_{i'}$, and so

$$\sum_{j=k}^{s} \tilde{a}_j = a_{i'} + 1 + \sum_{j=k+1}^{s} a_j \le \sum_{j=k}^{s} a_j + 1 \le \sum_{j=k}^{s} b_j + 1 = b_i + \sum_{j=k+1}^{s} b_j + 1 = \sum_{j=k}^{s} \tilde{b}_j.$$

We conclude that $\tilde{\mathbf{a}} \preceq_{\text{GSC}} \tilde{\mathbf{b}}$, and the first assertion follows.

2. To prove the second assertion, let $\hat{\mathbf{a}} := \mathcal{R}(\mathbf{a} - \mathbf{e}_i)$ and $\hat{\mathbf{b}} := \mathcal{R}(\mathbf{b} - \mathbf{e}_i)$. First, for any $k > i$ we easily get that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k}^{s} a_j \leq \sum_{j=k}^{s} b_j = \sum_{j=k}^{s} \hat{b}_j.$$

Now, for any $k \leq i$, we have that

$$\sum_{j=k}^{s} \hat{a}_j = \left(\sum_{j=k}^{s} a_j - 1 \right) \vee \left(\sum_{j=k-1;\, j\neq i}^{s} a_j \right); \tag{25}$$

$$\sum_{j=k}^{s} \hat{b}_j = \left(\sum_{j=k}^{s} b_j - 1 \right) \vee \left(\sum_{j=k-1;\, j\neq i}^{s} b_j \right). \tag{26}$$

Then there are two sub-cases to consider:

(i) If $a_i > a_{k-1}$, we deduce from (25)–(26) that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k}^{s} a_j - 1 \leq \sum_{j=k}^{s} b_j - 1 \leq \sum_{j=k}^{s} \hat{b}_j.$$

(ii) If $a_i = a_{k-1}$, we also have that $a_i = a_k$, and it follows from (25) that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k-1;\, j\neq i}^{s} a_j = \sum_{j=k}^{s} a_j.$$

We are thus in the following alternative:

(iia) If $b_{k-1} = b_i$, then from (26) we get that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k}^{s} a_j \leq \sum_{j=k}^{s} b_j = \sum_{j=k}^{s} \hat{b}_j.$$

(iib) If $\sum_{j=k}^{s} a_j < \sum_{j=k}^{s} b_j$, then we obtain that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k}^{s} a_j \leq \sum_{j=k}^{s} b_j - 1 \leq \sum_{j=k}^{s} \hat{b}_j.$$

(iic) If $\sum_{j=k}^{s} a_j = \sum_{j=k}^{s} b_j$ and $b_{k-1} < b_i$, then observe, first, that it must be the case that $b_{k-1} \geq a_{k-1}$. Indeed, $b_{k-1} < a_{k-1}$ would imply that $\sum_{j=k-1}^{s} a_j > \sum_{j=k-1}^{s} b_j$, a contradiction to the assumption that $\mathbf{a} \preceq_{\text{GSC}} \mathbf{b}$. Recalling that $\mathbf{a}$ and $\mathbf{b}$ are ordered, this implies, first, that $b_j \geq b_{k-1} \geq a_{k-1} = a_i = a_j$ for all $j \in [\![k, i-1]\!]$ (whenever $i > k$), and second, that $b_i > b_{k-1} \geq a_{k-1} = a_i$. We thus obtain that

$$\sum_{j=k}^{s} \hat{a}_j = \sum_{j=k}^{s} a_j = \sum_{j=k}^{i-1} a_j + a_i + \sum_{j=i+1}^{s} a_j \leq \sum_{j=k}^{i-1} b_j + b_i - 1 + \sum_{j=i+1}^{s} b_j = \sum_{j=k}^{s} b_j - 1 = \sum_{j=k}^{s} \hat{b}_j,$$

where sums over the set $[\![k, i-1]\!]$, $k = i$, are defined to equal 0.

This shows that $\hat{a} \preceq_{\text{GSC}} \hat{b}$.

3. Regarding the third assertion, fix $i \le s$ such that $a_i = 0$ and $b_i > 0$, and denote again $\hat{\mathbf{b}} = \mathcal{R}(\mathbf{b} - \mathbf{e}_i)$. Then, for any $k > i$ we clearly have that

$$\sum_{j=k}^{s} a_j \le \sum_{j=k}^{s} b_j = \sum_{j=k}^{s} \hat{b}_j,$$

whereas for $k \le i$, as $\sum_{j=i+1}^{s} a_j \le \sum_{j=i+1}^{s} b_j$, we have

$$\sum_{j=k}^{s} a_j = \sum_{j=i+1}^{s} a_j \le \sum_{j=k}^{i} b_j - 1 + \sum_{j=i+1}^{s} b_j = \sum_{j=k}^{s} b_j - 1 \le \sum_{j=k}^{s} \hat{b}_j,$$

where we used (26) in the last inequality.

□

Proof of Proposition 1. Consider two s-server systems, one operating under a stable $\mathbf{p}$-allocation policy, and the other operating under its non-idling counterpart; let Q and Q^{NI} denote the ordered queue processes (CTMCs) under the corresponding control. We couple the two systems (to which we refer as the "idling" and "non-idling" system) as follows: First, we feed both systems by the same arrival process. Second, upon each arrival, a common draw of the distribution $\mathbf{p}$ determines, independently of everything else, the targeted queue of the incoming customer in the idling system, and in the non-idling system only if no idle server is present. If there is an idle server in the non-idling system, then that arrival is routed according to the realization of $\mathbf{p}$ in the idling system, but to the idle server in the non-idling system.

Finally, we couple the service times in both systems, so as to satisfy the following property: for any $i \le s$, at each time point t such that $Q_i(t) > 0$ and $Q_i^{\text{NI}}(t) > 0$, the remaining service time at server i is equal in the two systems. In particular, ongoing services at the servers with the same indices in both systems are synchronized. To this end, it suffices to reset the service times of the customers in service at server i using a common realization of the exponential service times whenever there is a change concerning server i in either system, e.g., a re-ordering of the queues, or an arrival to server i at a time when server i is idling in one system but not in the other. Note that resetting the service times does not change the distribution of the service times, due to the memoryless property, and does not impact the overall distribution of the systems, due to their strong Markov property.

Denote by $\hat{Q}$ and $\hat{Q}^{\text{NI}}$ the coupled ordered CTMCs of the two systems. From the construction above it is clear that both these CTMCs are defined on the same probability space and that $\hat{Q} \overset{\text{d}}{=} Q$ and $\hat{Q}^{\text{NI}} \overset{\text{d}}{=} Q^{\text{NI}}$, although the joint distributions of $(\hat{Q}, \hat{Q}^{\text{NI}})$ is different than the joint distribution of the original systems. (In fact, the original systems may not have any specified joint distribution.) Take $\hat{Q}(0)=\hat{Q}^{\text{NI}}(0)$. We argue that

$$\hat{Q}^{\text{NI}}(t)\preceq_{\text{GSC}}\hat{Q}(t),\qquad \text{for all } t\geq 0\quad w.p.1. \tag{27}$$

As both processes are constant between event times, it suffices to show that (27) holds at event times (arrivals and departures). Therefore, for $\bar{T}_0=0$, and $\bar{T}_n$ denoting the time of the nth event, we need to show $\hat{Q}^{\text{NI}}(\bar{T}_n)\preceq_{\text{GSC}}\hat{Q}(\bar{T}_n)$ for all $n\geq 0$ w.p.1. We prove this by induction on n. This is true by assumption for $n=0$, and if this is true at a given $n\in\mathbb{Z}_+$, then we are in the following alternatives:

(i) If $\bar{T}_{n+1}$ is an arrival time in both systems and the common draw following the distribution $\mathbf{p}$ draws index i, then:

(ia) if server 1 is busy in the non-idling system (i.e. the first coordinate of $\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)$ is non-zero), then in view of the induction assumption, and by applying assertion 2 of Lemma 4 to $i'=i$, we have

$$\hat{Q}^{\text{NI}}\left(\bar{T}_{n+1}\right)=\mathcal{R}\left(\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)+\mathbf{e}_i\right)\preceq_{\text{GSC}}\mathcal{R}\left(\hat{Q}\left(\bar{T}_n\right)+\mathbf{e}_i\right)=\hat{Q}\left(\bar{T}_{n+1}\right).$$

(ib) If server 1 is idling in the non-idling system (that is, $\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)=0$), then from the induction assumption, and by applying assertion 2 of Lemma 4 to $i'=1$, we have

$$\hat{Q}^{\text{NI}}\left(\bar{T}_{n+1}\right)=\mathcal{R}\left(\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)+\mathbf{e}_1\right)\preceq_{\text{GSC}}\mathcal{R}\left(\hat{Q}\left(\bar{T}_n\right)+\mathbf{e}_i\right)=\hat{Q}\left(\bar{T}_{n+1}\right).$$

(ii) If $\bar{T}_{n+1}$ is a departure time from server i in both systems, then in view of the induction assumption and applying assertion 1 of Lemma 4, we have

$$\hat{Q}^{\text{NI}}\left(\bar{T}_{n+1}\right)=\mathcal{R}\left(\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)-\mathbf{e}_i\right)\preceq_{\text{GSC}}\mathcal{R}\left(\hat{Q}\left(\bar{T}_n\right)-\mathbf{e}_i\right)=\hat{Q}\left(\bar{T}_{n+1}\right).$$

(iii) If $\bar{T}_{n+1}$ is a departure from server i in the non-idling system, and if server i is idling at this time in the idling system, then

$$\hat{Q}^{\text{NI}}\left(\bar{T}_{n+1}\right)=\mathcal{R}\left(\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)-\mathbf{e}_i\right)\preceq_{\text{GSC}}\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)\preceq_{\text{GSC}}\hat{Q}\left(\bar{T}_n\right),$$

where we utilized the inductive assumption in the last inequality.

(iv) If $\bar{T}_{n+1}$ is a departure time from server i in the idling system, and server i is idling at this time in the non-idling system, then necessarily, $\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)_i=0$ and $\hat{Q}\left(\bar{T}_n\right)_i>0$. Thus, using the induction assumption together with assertion 3 of Lemma 4, we have

$$\hat{Q}^{\text{NI}}\left(\bar{T}_{n+1}\right)=\hat{Q}^{\text{NI}}\left(\bar{T}_n\right)\preceq_{\text{GSC}}\mathcal{R}\left(\hat{Q}\left(\bar{T}_n\right)-\mathbf{e}_i\right)=\hat{Q}\left(\bar{T}_{n+1}\right).$$

Therefore, (27) holds, and in turn, $\sum_{i=1}^{s}Q^{\text{NI}}(t)_i\leq_{st}\sum_{i=1}^{s}Q_i(t)$ for all $t\geq 0$. Thus, if Q is positive recurrent, then so is Q^{NI}. □

B. Remaining Proofs of Results in Section 5

In this section, we prove Propositions 2 and 3, and Lemma 3.

B.1. Proof of Proposition 2

We now turn to the proof of Proposition 2. We consider a J2SQ$^{\text{NI}}(p)$ system with $s=2$ servers. Thanks to Proposition 4 and Theorem 1, only the right inclusion in (12) for $p>1/2$ needs to be proved. To this end, assume that $p>1/2$ and $\rho<V_{\text{cr}}(p)$. It is useful in this case to label the two servers, say server 1 and server 2, and to consider the CTMC $\tilde{Q}(t):=\left(\tilde{Q}_1(t),\tilde{Q}_2(t)\right)$, $t\geq 0$, where for all $t\geq 0$, $\tilde{Q}_i(t)$ denotes the queue at server i at time t, $i=1,2$. (In particular, $\tilde{Q}$ is not an $\mathcal{R}\left(\mathbb{Z}_+^s\right)$-valued process.) Let $\{\tilde{Q}_n\}$ denote the embedded DTMC, i.e., the process $\left\{\left(\tilde{Q}_1(T_n^-),\tilde{Q}_2(T_n^-)\right):n\geq 1\right\}$, where T_n is the time of the nth event of the CTMC $\tilde{Q}$.

Under the J2SQ$^{\text{NI}}(p)$ policy, the planar chain $\{\tilde{Q}_n\}$ has the following transitions on the positive quadrant:

$$\begin{cases}
\text{Origin:} & P_{(0,0),(0,1)} &= 1/2, & P_{(0,0),(1,0)} &= 1/2, & \\
x\text{-axis:} & P_{(x,0),(x-1,0)} &= \frac{\mu}{\lambda+\mu}, & P_{(x,0),(x,1)} &= \frac{\lambda}{\lambda+\mu}, & x\in\mathbb{Z}_+^*, \\
y\text{-axis:} & P_{(0,y),(0,y-1)} &= \frac{\mu}{\lambda+\mu}, & P_{(0,y),(1,y)} &= \frac{\lambda}{\lambda+\mu}, & y\in\mathbb{Z}_+^*, \\
\text{Interior:} & P_{(x,y),(x-1,y)} &= \frac{\mu}{\lambda+2\mu}, & P_{(x,y),(x,y-1)} &= \frac{\mu}{\lambda+2\mu}, & x,y\in\mathbb{Z}_+^*, \\
& P_{(x,y),(x,y+1)} &= \frac{\lambda(1-p)}{\lambda+2\mu}, & P_{(x,y),(x+1,y)} &= \frac{\lambda p}{\lambda+2\mu}, & x,y\in\mathbb{Z}_+^*;\, x>y, \\
& P_{(x,y),(x,y+1)} &= \frac{\lambda p}{\lambda+2\mu}, & P_{(x,y),(x+1,y)} &= \frac{\lambda(1-p)}{\lambda+2\mu}, & x,y\in\mathbb{Z}_+^*;\, x<y, \\
& P_{(x,x),(x,x+1)} &= \frac{\lambda/2}{\lambda+2\mu}, & P_{(x,x),(x+1,x)} &= \frac{\lambda/2}{\lambda+2\mu}, & x\in\mathbb{Z}_+^*.
\end{cases} \tag{28}$$

As the above transitions are not space-homogeneous, the DTMC $\{\tilde{Q}_n\}$ is not directly amenable to the ergodicity criteria in Theorem 3.3.1 of [18]. To circumvent this difficulty we consider two auxiliary DTMCs $\{\tilde{Q}_n^1\}$ and $\{\tilde{Q}_n^2\}$, having the respective sets of transitions P^1 and P^2 defined via

$$\begin{cases}
\text{Origin:} & P^1_{(0,0),(0,1)} &= P^2_{(0,0),(0,1)} = 1, & & & \\
x\text{-axis:} & P^1_{(x,0),(x-1,0)} &= P^2_{(x,0),(x-1,0)} = \frac{\mu}{\lambda+\mu}, & P^1_{(x,0),(x,1)} &= P^2_{(x,0),(x,1)} = \frac{\lambda}{\lambda+\mu}, & x\in\mathbb{Z}_+^*, \\
y\text{-axis:} & P^1_{(0,y),(0,y-1)} &= P^2_{(0,y),(0,y-1)} = \frac{\mu}{\lambda+\mu}, & P^1_{(0,y),(1,y)} &= P^1_{(0,y),(1,y)} = \frac{\lambda}{\lambda+\mu}, & y\in\mathbb{Z}_+^*, \\
\text{Interior:} & P^1_{(x,y),(x-1,y)} &= P^2_{(x,y),(x,y-1)} = \frac{\mu}{\lambda+2\mu}, & P^1_{(x,y),(x,y-1)} &= P^2_{(x,y),(x-1,y)} = \frac{\mu}{\lambda+2\mu}, & x,y\in\mathbb{Z}_+^*, \\
& P^1_{(x,y),(x,y+1)} &= P^2_{(x,y),(x+1,y)} = \frac{\lambda(1-p)}{\lambda+2\mu}, & P^1_{(x,y),(x+1,y)} &= P^2_{(x,y),(x,y+1)} = \frac{\lambda p}{\lambda+2\mu}, & x,y\in\mathbb{Z}_+^*.
\end{cases} \tag{29}$$

The transitions of the three chains $\{\tilde{Q}_n\}$, $\{\tilde{Q}_n^1\}$ and $\{\tilde{Q}_n^2\}$ are represented in Figure 3.

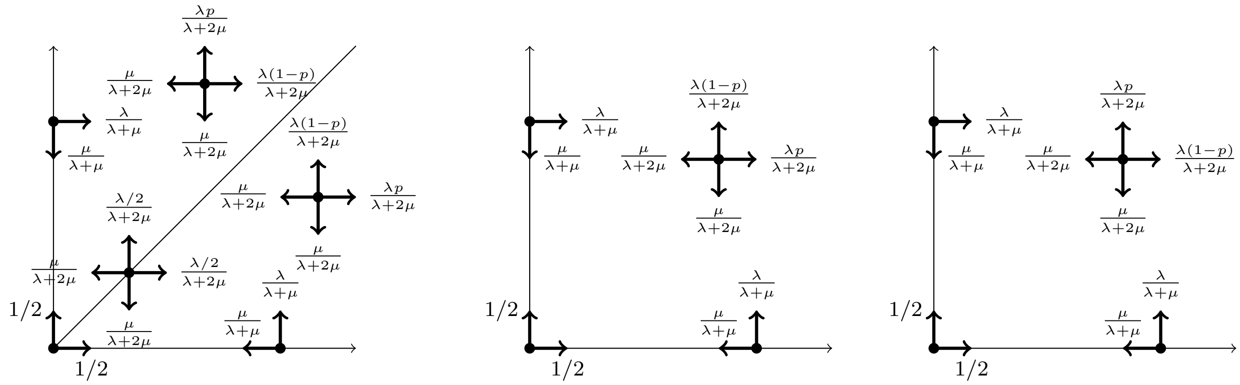
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\lambda/2}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda/2}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
1/2
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
1/2
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
1/2

Figure 3 Transitions of the Planar chains $\{\tilde{Q}_n\}$ (left) $\{\tilde{Q}^1_n\}$ (middle) and $\{\tilde{Q}^2_n\}$ (right) for the J2SQ$^{\text{NI}}(p)$ system.

For a planar Markov chain $\{U_n\}=\{(U^x_n,U^y_n)\}$, the mean horizontal (respectively, vertical) drift at $u=(u^x,u^y)$ is defined to be $\mathbb{E}\left[U^x_{n+1}-U^x_n \mid U_n=u\right]$ (respectively, $\mathbb{E}\left[U^y_{n+1}-U^y_n \mid U_n=u\right]$). Denote by (Δ^1_x,Δ^1_y), $(\Delta^{1'}_x,\Delta^{1'}_y)$ and $(\Delta^{1''}_x,\Delta^{1''}_y)$ the mean (horizontal/vertical) drifts of the chain $\{\tilde{Q}^1_n\}$, starting from, respectively, the interior, the x-axis and the y-axis of the quarter plan. Similarly, denote by (Δ^2_x,Δ^2_y), $(\Delta^{2'}_x,\Delta^{2'}_y)$ and $(\Delta^{2''}_x,\Delta^{2''}_y)$ the mean (horizontal/vertical) drifts of the chain $\{\tilde{Q}^2_n\}$, respectively, in the interior, the x-axis, and the y-axis. It follows from (29) that these drifts are as follows.

$$\{\tilde{Q}^1_n\}:\begin{cases}x\text{-axis:} & \Delta^{1'}_x=-\frac{\mu}{\lambda+\mu},\ \Delta^{1'}_y=\frac{\lambda}{\lambda+\mu};\\ y\text{-axis:} & \Delta^{1''}_x=\frac{\lambda}{\lambda+\mu},\quad \Delta^{1''}_y=-\frac{\mu}{\lambda+\mu};\\ \text{Interior:} & \Delta^1_x=\frac{\lambda p-\mu}{\lambda+2\mu},\quad \Delta^1_y=\frac{\lambda(1-p)-\mu}{\lambda+2\mu}.\end{cases}\qquad \{\tilde{Q}^2_n\}:\begin{cases}x\text{-axis:} & \Delta^{2'}_x=-\frac{\mu}{\lambda+\mu},\quad \Delta^{2'}_y=\frac{\lambda}{\lambda+\mu};\\ y\text{-axis:} & \Delta^{2''}_x=\frac{\lambda}{\lambda+\mu},\quad \Delta^{2''}_y=-\frac{\mu}{\lambda+\mu};\\ \text{Interior:} & \Delta^2_x=\frac{\lambda(1-p)-\mu}{\lambda+2\mu},\ \Delta^2_y=\frac{\lambda p-\mu}{\lambda+2\mu}.\end{cases}$$

As $\lambda(1-p)-\mu<\lambda/2-\mu<0$, we have that $\Delta^1_y=\Delta^2_x<0$. Recalling (11), there are two sub-cases:

Case 1: $\boldsymbol{\rho<1/(2p)}$. In this case we also have that $\Delta^1_x=\Delta^2_y<0$. Then we have that

$$\Delta^1_x\Delta^{1'}_y-\Delta^1_y\Delta^{1'}_x=\Delta^2_y\Delta^{2''}_x-\Delta^2_x\Delta^{2''}_y=\frac{\mu^2}{(\lambda+\mu)(\lambda+2\mu)}\left(4p\rho^2-2p\rho-1\right), \tag{30}$$

$$\Delta^1_y\Delta^{1''}_x-\Delta^1_x\Delta^{1''}_y=\Delta^2_x\Delta^{2'}_y-\Delta^2_y\Delta^{2'}_x=\frac{\mu^2}{(\lambda+\mu)(\lambda+2\mu)}\left(4(1-p)\rho^2-2(1-p)\rho-1\right). \tag{31}$$

Now, the right-hand side in (30) is strictly negative, due to the facts that $0<\rho<V_{\text{cr}}(p)$, and that $V_{\text{cr}}(p)$ is the only positive root of the polynomial $x\mapsto 4px^2-2px-1$, as mentioned in Lemma 2. Similarly, by replacing p with $(1-p)$ in Lemma 2, we see that the right-hand side of (31) is also strictly negative. It follows from Assertion (a-i) in Theorem 3.3.1 of [18] that the DTMC's $\{\tilde{Q}^1_n\}$ and $\{\tilde{Q}^2_n\}$ are both positive recurrent.

More specifically, applying the argument leading to assertion (3.15) in [18] shows that, for any $\epsilon < 0$, there exist three real numbers u, v, w, such that $u, v > 0$, $w^2 < 4uv$ and

$$\begin{cases} 2u\Delta_x^1 + w\Delta_y^1 = 2u\Delta_y^2 + w\Delta_x^2 < -\epsilon\,; \\ 2v\Delta_y^1 + w\Delta_x^1 = 2v\Delta_x^2 + w\Delta_y^2 < -\epsilon\,; \\ 2u\Delta_x^{1'} + w\Delta_y^{1'} = 2u\Delta_y^{2''} + w\Delta_x^{2'''} < -\epsilon\,; \\ 2v\Delta_y^{1''} + w\Delta_x^{1''} = 2v\Delta_x^{2'} + w\Delta_y^{2'} < -\epsilon. \end{cases} \tag{32}$$

It is then clear that (32) also holds when replacing both u and v by $u \vee v$ throughout. Consequently, defining the Lyapunov function $F : (x, y) \mapsto \sqrt{(u \vee v)x^2 + (u \vee v)y^2 + wxy}$, Lemma 3.3.3 in [18] implies that, for some compact set K in the positive quadrant, for some $\epsilon' > 0$, for any $(x, y) \notin K$ and any $n \in \mathbb{Z}_+$,

$$\left(\mathbb{E}\left[F(\tilde{Q}^1_{n+1}) - F(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x, y)\right]\right) \vee \left(\mathbb{E}\left[F(\tilde{Q}^2_{n+1}) - F(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x, y)\right]\right) < -\epsilon'.$$

Fix $(x, y) \notin K$ and $n \in \mathbb{Z}_+$, and recall (28–29). If $x > y \geq 0$, we get that

$$\mathbb{E}\left[F(\tilde{Q}_{n+1}) - F(\tilde{Q}_n) \mid \tilde{Q}_n = (x, y)\right] = \mathbb{E}\left[F(\tilde{Q}^1_{n+1}) - F(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x, y)\right] < -\epsilon';$$

if $0 \leq x < y$, we get

$$\mathbb{E}\left[F(\tilde{Q}_{n+1}) - F(\tilde{Q}_n) \mid \tilde{Q}_n = (x, y)\right] = \mathbb{E}\left[F(\tilde{Q}^2_{n+1}) - F(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x, y)\right] < -\epsilon'.$$

Finally, if $x > 0$, we obtain

$$\begin{aligned}
\mathbb{E}\left[F(\tilde{Q}_{n+1}) - F(\tilde{Q}_n) \mid \tilde{Q}_n = (x, x)\right] &= \frac{1}{\lambda + 2\mu}\bigg(\frac{\lambda}{2}\left(F(x+1, x) - F(x, x)\right) + \frac{\lambda}{2}\left(F(x, x+1) - F(x, x)\right) \\
&\quad + \mu\left(F(x-1, x) - F(x, x)\right) + \mu\left(F(x, x-1) - F(x, x)\right)\bigg) \\
&= \frac{1}{2(\lambda + 2\mu)}\bigg(\lambda p\left(F(x+1, x) - F(x, x)\right) + \lambda(1-p)\left(F(x, x+1) - F(x, x)\right) \\
&\quad + \mu\left(F(x-1, x) - F(x, x)\right) + \mu\left(F(x, x-1) - F(x, x)\right)\bigg) \\
&\quad + \frac{1}{2(\lambda + 2\mu)}\bigg(\lambda(1-p)\left(F(x+1, x) - F(x, x)\right) + \lambda p\left(F(x, x+1) - F(x, x)\right) \\
&\quad + \mu\left(F(x-1, x) - F(x, x)\right) + \mu\left(F(x, x-1) - F(x, x)\right)\bigg) \\
&= \frac{1}{2}\mathbb{E}\left[F(\tilde{Q}^1_{n+1}) - F(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x, x)\right] + \frac{1}{2}\mathbb{E}\left[F(\tilde{Q}^2_{n+1}) - F(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x, x)\right] \\
&< -\epsilon'.
\end{aligned}$$

It follows from the Lyapunov-Foster Theorem that the DTMC $\{\tilde{Q}_n\}$ is positive recurrent, and in turn, so is the CTMC $\tilde{Q}$ by, e.g., [29, Theorem 6.18].

Case 2: $\boldsymbol{\frac{1}{2p} \leq \rho < V_{\mathrm{cr}}(p)}$. In this case $\Delta^1_y = \Delta^2_x \geq 0$. Moreover, (30) still holds, so both chains are again positive recurrent (applying respectively cases (b-i) and (c-i) of Lemma 3.3.1 in [18]). Again, (32) is satisfied for some $\epsilon > 0$ and some u, v, w such that $u, v > 0$ and $w^2 < 4uv$. One can easily check that (32) still holds when replacing u by $u \wedge v$ and v by $u \vee v$. Therefore, by [18, Lemma 3.3.3], there exist Lyapunov functions $F^1 : (x,y) \mapsto \sqrt{(u\wedge v)x^2 + (u \vee v)y^2 + wxy}$ and $F^2 : (x,y) \mapsto \sqrt{(u\vee v)x^2 + (u \wedge v)y^2 + wxy}$ for $\{\tilde{Q}^1_n\}$ and $\{\tilde{Q}^2_n\}$, respectively, two compact sets K^1 and K^2, and an $\epsilon' > 0$, such that for all $(x,y) \notin K^1 \cup K^2$ and all n,

$$\left(\mathbb{E}\left[F^1(\tilde{Q}^1_{n+1}) - F^1(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x,y)\right]\right) \vee \left(\mathbb{E}\left[F^2(\tilde{Q}^2_{n+1}) - F^2(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x,y)\right]\right) < -\epsilon'. \quad (33)$$

Let

$$F : (x,y) \mapsto F^1(x,y)\mathbf{1}_{\{x \geq y\}} + F^2(x,y)\mathbf{1}_{\{x<y\}}.$$

It follows from (33) that for all $n \geq 1$ and $(x,y) \notin K^1 \cup K^2$,

$$\mathbb{E}\left[F(\tilde{Q}_{n+1}) - F(\tilde{Q}_n) \mid \tilde{Q}_n = (x,y)\right] = \begin{cases} \mathbb{E}\left[F^1(\tilde{Q}^1_{n+1}) - F^2(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x,y)\right] < -\epsilon' & \text{if } 0 \leq x < y; \\ \mathbb{E}\left[F^2(\tilde{Q}^2_{n+1}) - F^2(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x,y)\right] < -\epsilon' & \text{if } 0 \leq y < x. \end{cases}$$

Only the case of a starting point (x,x), for $x \in \mathbb{Z}^*_+$, remains to be treated. Then, simple algebra shows that, for some positive constant C,

$$F^2(x+1,x) - F^1(x+1,x) = C\left((u\vee v) - (u \wedge v)\right)(2x+1) \geq 0,$$

and we obtain similarly that the quantities $F^1(x,x+1) - F^2(x,x+1)$, $F^2(x,x-1) - F^1(x,x-1)$ and $F^1(x-1,x) - F^2(x-1,x)$ are non-negative. Therefore, if $(x,x) \notin K^1 \cup K^2$, it follows from (33) that

$$\begin{aligned}
&\mathbb{E}\left[F(\tilde{Q}_{n+1}) - F(\tilde{Q}_n) \mid \tilde{Q}_n = (x,x)\right] \\
&= \frac{1}{\lambda+2\mu}\Bigg(\frac{\lambda}{2}\left(F^1(x+1,x) - F(x,x)\right) + \frac{\lambda}{2}\left(F^2(x,x+1) - F(x,x)\right) \\
&\quad + \mu\left(F^2(x-1,x) - F(x,x)\right) + \mu\left(F^1(x,x-1) - F(x,x)\right)\Bigg) \\
&= \frac{1}{2(\lambda+2\mu)}\Bigg(\lambda p\left(F^1(x+1,x) - F^1(x,x)\right) + \lambda(1-p)\left(F^1(x+1,x) - F^2(x,x)\right) \\
&\quad + \lambda(1-p)\left(F^2(x,x+1) - F^1(x,x)\right) + \lambda p\left(F^2(x,x+1) - F^2(x,x)\right) \\
&\quad + \mu\left(F^2(x-1,x) - F^1(x,x)\right) + \mu\left(F^2(x-1,x) - F^2(x,x)\right) \\
&\quad + \mu\left(F^1(x,x-1) - F^1(x,x)\right) + \mu\left(F^1(x,x-1) - F^2(x,x)\right)\Bigg) \\
&\le \frac{1}{2(\lambda+2\mu)}\Bigg(\lambda p\left(F^1(x+1,x) - F^1(x,x)\right) + \lambda(1-p)\left(F^1(x,x+1) - F^1(x,x)\right) \\
&\quad + \mu\left(F^1(x-1,x) - F^1(x,x)\right) + \mu\left(F^1(x,x-1) - F^1(x,x)\right)\Bigg) \\
&\quad + \frac{1}{2(\lambda+2\mu)}\Bigg(\lambda(1-p)\left(F^2(x+1,x) - F^2(x,x)\right) + \lambda p\left(F^2(x,x+1) - F^2(x,x)\right) \\
&\quad + \mu\left(F^2(x-1,x) - F^2(x,x)\right) + \mu\left(F^2(x,x-1) - F^2(x,x)\right)\Bigg) \\
&= \frac{1}{2}\mathbb{E}\left[F^1(\tilde{Q}^1_{n+1}) - F^1(\tilde{Q}^1_n) \mid \tilde{Q}^1_n = (x,x)\right] + \frac{1}{2}\mathbb{E}\left[F^2(\tilde{Q}^2_{n+1}) - F^2(\tilde{Q}^2_n) \mid \tilde{Q}^2_n = (x,x)\right] \\
&< -\epsilon',
\end{aligned}$$

which, by virtue of the Lyapunov-Foster Theorem, implies the result. □

B.2. Proof of Proposition 3

We use the same notation as in the proof of Proposition 2. Without the non-idling assumption, the chain $\{\tilde{Q}_n\}$ has mostly the same transitions as in (28), except for

$$\begin{cases} x\text{-axis: } P_{(x,0),(x-1,0)} = \frac{\mu}{\lambda+\mu},\ P_{(x,0),(x,1)} = \frac{\lambda(1-p)}{\lambda+\mu},\ P_{(x,0),(x+1,0)} = \frac{\lambda p}{\lambda+\mu},\ x \in \mathbb{Z}^*_+, \\ y\text{-axis: } P_{(0,y),(0,y-1)} = \frac{\mu}{\lambda+\mu},\ P_{(0,y),(1,y)} = \frac{\lambda(1-p)}{\lambda+\mu},\ P_{(0,y),(0,y+1)} = \frac{\lambda p}{\lambda+\mu},\ x \in \mathbb{Z}^*_+. \end{cases}$$

The transitions of the three chains $\{\tilde{Q}_n\}$, $\{\tilde{Q}^1_n\}$ and $\{\tilde{Q}^2_n\}$ are then represented in Figure 4.

Then the interior drifts Δ^1_x, Δ^1_y, Δ^2_x and $\Delta^2_y < 0$ are all strictly negative, and the equalities in (30-31) become

$$\Delta^1_x \Delta^{1'}_y - \Delta^1_y \Delta^{1'}_x = \Delta^2_y \Delta^{2''}_x - \Delta^2_x \Delta^{2''}_y = \frac{\mu}{(\lambda+\mu)(\lambda+2\mu)}\left(\lambda p - \mu\right), \tag{34}$$

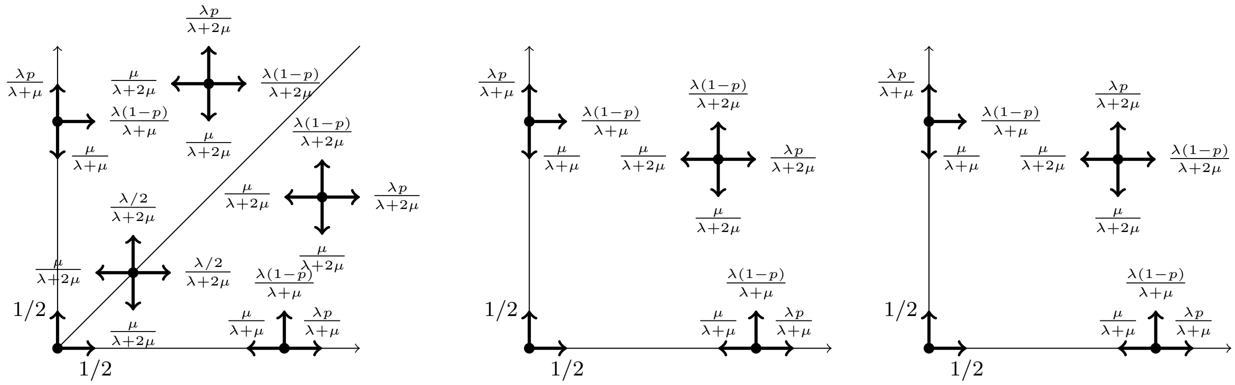
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\lambda p}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\lambda/2}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda/2}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
$\frac{\lambda p}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
1/2
$\frac{\lambda p}{\lambda+\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
$\frac{\lambda p}{\lambda+\mu}$
1/2
$\frac{\lambda p}{\lambda+\mu}$
$\frac{\lambda p}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+2\mu}$
$\frac{\mu}{\lambda+2\mu}$
$\frac{\lambda(1-p)}{\lambda+\mu}$
1/2
$\frac{\mu}{\lambda+\mu}$
$\frac{\lambda p}{\lambda+\mu}$
1/2

Figure 4 Transitions of the Planar chains $\{\tilde{Q}_n\}$ (left) $\{\tilde{Q}^1_n\}$ (middle) and $\{\tilde{Q}^2_n\}$ (right) for the J2SQ(p) system.

$$\Delta^1_y\Delta^{1''}_x - \Delta^1_x\Delta^{1''}_y = \Delta^2_x\Delta^{2'}_y - \Delta^2_y\Delta^{2'}_x = \frac{\mu}{(\lambda+\mu)(\lambda+2\mu)}\left(\lambda p - \mu\right). \tag{35}$$

If $\rho < \frac{1}{2p}$, then the quantities in (34) are strictly negative; as in Case 1 in the proof of Proposition 2, this implies that the queue process is positive recurrent. If $\rho \geq \frac{1}{2p}$, then we are in case (a-ii) in [18, Theorem 3.3.1], implying that the DTMC $\{\tilde{Q}_n\}$ cannot be positive recurrent. Specifically, by [18, Theorem 3.3.2], if $\rho = \frac{1}{2p}$ (respectively, if $\rho > \frac{1}{2p}$), then the embedded DTMC is null recurrent (respectively, transient), and so is the queue process $\tilde{Q}$. □

B.3. Proof of Lemma 3

For $m \in [\![2, s]\!]$ let $\pi_{\rho,m}$ denote the loss probability of a $M/M/m-1/0$ queue (a loss system with $m-1$ servers), having traffic intensity $s\rho = \lambda/\mu$; then

$$\pi_{\rho,m} := \frac{(s\rho)^{m-1}/(m-1)!}{\sum_{i=0}^{m-1}(s\rho)^i/i!}.$$

Observe that $\rho \in \mathscr{G}(m)$, for $\mathscr{G}(m)$ in (16), is equivalent to $s\rho\pi_m < (s-m+1)$. Also, we clearly have that

$$\frac{1}{\pi_{\rho,m+1}} = 1 + \frac{m}{s\rho\pi_{\rho,m}},\ m = 2, ..., s-1. \tag{36}$$

First, $V_{\mathrm{cr}}(1,2) = \sup\mathscr{G}(2) < 1$ from (10). We then proceed by induction. Suppose that $\sup\mathscr{G}(m) < 1$ for some $m \in [\![2, s]\!]$. Let $\rho \in \mathscr{G}(m+1)$. If $\rho \geq \frac{(s-m)(s+1)}{(s-m+1)s}$, then we have that

$$s\rho\pi_{\rho,m+1} < (s-m) \leq s\rho\frac{s-m+1}{s+1}$$

which, after an immediate computation using (36), is equivalent to $s\rho\pi_{\rho,m} < s-m+1$, i.e. $\rho \in \mathscr{G}(m)$. By the induction assumption, this implies that

$$\sup \mathscr{G}(m+1) \leq \left(\sup \mathscr{G}(m) \vee \frac{(s-m)(s+1)}{(s-m+1)s}\right) < 1,$$

which concludes the proof. □

C. Auxiliary results

Let $L_1 := \{L(t) : t \geq 0\}$ denote the queue process in an $M/M/1/0$ queue (one-server loss system) having a Poisson arrival process with rate λ and service rate μ. The proof of the following lemma is a simple application of a standard coupling argument which we bring here for completeness.

Lemma 5. *Consider the process L_1, and let τ_1 denote the time of the first event after time* 0 *(arrival or departure). Then L_1 is stationary for all $t \geq \tau_1$; in particular, $P(L_1(t)) = 0) = 1 - P(L_1(t) = 0) = \mu/(\lambda+\mu)$, $t \geq \tau_1$.*

Proof. Let $L_e := \{L_e(t) : t \geq 0\}$ denote a stationary version of the process L_1, namely, $P(L_e(0) = 0) = 1 - P(L_e(0) = 1) = \mu/(\lambda+\mu)$. Let T denote the first time L_1 and L_e are equal; $T := \inf\{t \geq 0 : L(t) = L_e(t)\}$, and define the process

$$L_0(t) := \begin{cases} L_1(t) \ t < T, \\ L_e(t) \ t \geq T. \end{cases} \tag{37}$$

Since T is a stopping time that is finite w.p.1, the strong Markov property implies that $L_0 \overset{\mathrm{d}}{=} L_1$. The coupling inequality (e.g., [1, VII 2a] gives

$$\|P(L_1(t) \in \cdot) - \pi(\cdot)\|_{TV} \leq P(T > t).$$

Clearly, L_0 and L_e are equal when the first event (arrival or departure) in either of the two processes occurs, and in particular, when the first event in L_0 occurs. □

Similarly to the proof of Lemma 5 we can prove the following result. Recall that $L_m : -\{L_m(t) : t \geq 0\}$ denotes the number-in-system process in an $M/M/(m-1)/0$ queue–a loss system with $m-1$ servers and no buffer. Let $\tau_m := \inf\{t \geq 0 : L_m(t) = m-1\}$, namely, τ_m is the first time instant in which all servers are busy. Note that τ_m is a proper random variable, i.e., $P(\tau_m < \infty) = 1$.

Lemma 6. *If $L_m(0)=0$, then L_m is stationary for all $t\geq\tau_m$; in particular, for all $t\geq\tau_m$,*

$$P(L(t)=k)=\pi^{m-1}:=\frac{\rho^k/k!}{\sum_{j=0}^{m-1}\rho^j/j!},\quad k\in[\![1,m-1]\!].$$

Proof. Let L_∞ denote the stationary version of L_m, namely, $L_\infty(0)\overset{\mathrm{d}}{=}\pi^m$, for π^m in the statement of the lemma. We couple L_m and L_∞ on the same probability space and allow them to evolve independently of each other until they couple, after which the two processes follow the path of L_∞ (similarly to the construction of L_0 in the proof of Lemma 5). Since $L_m(0)=0$, the two processes must have coupled by τ_m, and so the result follows from the strong Markov property. □

References

[1] S. Asmussen. (2003). *Applied probability and queues.* Springer Verlag.

[2] R. Atar, I. Keslassy, and G. Mendelson. (2019). Subdiffusive load balancing in time-varying queueing systems. *Operations Research*, **67**(6), 1678–1698.

[3] F. Baccelli and P. Brémaud. (2002). *Elements of Queueing Theory* (2nd ed.). Springer.

[4] M. Bramson. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4**(2), 414–431.

[5] M. Bramson. (2008). *Stability of queueing networks.* Springer.

[6] M. Bramson, Y. Lu, and B. Prabhakar. (2010). Randomized load balancing with general service time distributions. *ACM SIGMETRICS performance evaluation review* 38(1), 275–286.

[7] M. Bramson. (2011) Stability of join the shortest queue networks. *The Annals of Applied Probability*, 21(4), 1568–1625.

[8] A. Brandt. (1985). On stationary waiting times and limiting behavior of queues with many servers I: the general G/G/m/∞ case. Elektron. Inform. u. Kybernet. **21**, 47–64.

[9] P. Brémaud. (1981). *Point Processes and Queues: Martingale Dynamics.* Springer, new York.

[10] P. Brémaud. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Texts Appl. Math. **31**. Springer, new York.

[11] G. Brightwell and M. Luczak (2012). The supermarket model with arrival rate tending to one. arXiv preprint arXiv:1201.5523.

[12] J.G. Dai. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, 5(1), 49–77.

[13] D.J. Daley. (1987). Certain optimality properties of the first-come first-served discipline for $G/G/s$ queues. *Stochastic Processes and their Applications*, 25, 301-308.

[14] L. Decreusefond and P. Moyal. (2012). *Stochastic Modeling and Analysis of Telecom Networks.* ISTE Wiley.

[15] P.S. Dester, C. Fricker and D. Tibi. (2017). Stationary analysis of the shortest queue problem. *Working paper.* Available at: arXiv: 1704.066442v3.

[16] P. Eschenfeldt and D. Gamarnik. (2015). Join the shortest queue with many servers. The heavy traffic asymptotics. *Working paper.* Available at: arXiv:1502.00999.

[17] P. Eschenfeldt and D. Gamarnik. (2016). Supermarket queueing system in the heavy traffic regime. Short queue dynamics. *Working paper.* Available at: arXiv:1610.03522.

[18] Fayolle, G., Malyshev, V. A. And Menshikov, M. (1995). Topics in the Constructive Theory of Countable Markov Chains. Cambridge University Press.

[19] L. Flatto and H.P. Mc Kean. (1977). Two queues in parallel. *Comm. Pure Appl. Math.*, **15**, 255-263.

[20] G.J. Foschini and J. Salz. (1978). A basic routing problem and diffusion. *IEEE Trans. on Comm.* **26**, 320–327.

[21] S. Foss. (1981). Comparison of service disciplines in multichannel service systems. *Siberian Math. Zh.*, **22**(1), 190–197.

[22] S. Foss and N. Chernova. (1998). On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems*, **29**(1), 55–73.

[23] C. Graham. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journ. Appl. Prob.* **37**, 198-211.

[24] C. Graham. (2005). Functional central theorems for a large network in which customers join the shortest among several queues or a queueing network with selection of the shortest of several queues. *Probab. Theory Relat. Fields.* **131**, 97-120.

[25] F.A. Haight. (1958). Two queues in parallel. *Biometrika* 45, 401–410.

[26] H.K. Khalil. (2002). *Nonlinear Systems.* Prentice Hall, New Jersey.

[27] J. Kiefer and J. Wolfowitz. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78**, 1–18.

[28] J.F.C. Kingman. (1961). Two Similar Queues in Parallel. *The Annals of Mathematical Statistics* **32**(4), 1314–1323.

[29] V.G. Kulkarni. (2017). *Modeling and analysis of stochastic systems.* Chapman and Hall/CRC.

[30] Liberzon, D. (2003). *Switching in Systems and Control.* Birkäuser.

[31] R.M. Loynes. (1962). The stability of queues with non-independent interarrivals and service times. *Proceedings of the Cambridge Philosophical Society*, **58**, 497–520.

[32] S.H. Lu and P.R. Kumar. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control.* **36**(12), 1406–1416.

[33] Y. Lu, Q. Xie, G. Kliot, A. Geller, J.R. Larus and A. Greenberg. (2011). Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11), 1056–1071.

[34] M.J. Luczak and C. McDiarmid. (2006). On the maximum queue length in the supermarket model. *The Annals of Probability*, 34(2), 493–527.

[35] M.J. Luczak and C. McDiarmid. (2007). Asymptotic distributions and chaos for the supermarket model. *Electronic Journal of Probability* 12, 75–99.

[36] A.W. Marshall and I. Olkin. (1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

[37] M. Mitzenmacher. (1996). *The Power of Two Choices in Randomized Load Balancing*, PhD thesis, Univ. of California, Berkeley.

[38] P. Moyal. (2017). On the Stability of non-monotonic systems of parallel queues. *Discrete Events Dynamic Systems*, 27(1), 85–107.

[39] P. Moyal. (2017). A pathwise comparison of parallel queues. *Discrete Events Dynamic Systems*, 27(3), 573–584.

[40] P. Moyal and O. Perry. (2017). On the instability of matching queues. *The Annals of Applied Probability*, 27(6), pp. 3385-3434.

[41] G. Pang, R. Talreja and W. Whitt. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* 4, pp. 193–267.

[42] O. Perry and W. Whitt. (2016). Chattering and Congestion Collapse in an Overload Switching Control. *Stochastic Systems*, 6(1), pp. 132–210.

[43] M.I. Reiman. (1984). Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology*, 207–240. Springer, Berlin, Heidelberg.

[44] P. Robert. (2003). *Stochastic networks and queues.* Springer-Verlag.

[45] A.N. Rybko and A.L. Stolyar. (1992). Ergodicity of stochastic processes describing the operations of open queueing networks. *Problems Inform. Transmission* **28**, 3–26 (in Russian).

[46] A. Scheller-Wolf. (2003). Necessary and sufficient conditions for delay moments in FIFO multiserver queues with and application comparing s slow servers with one fast one. *Operations Research* **51**(5): 748–758.

[47] S.R. Turner. (1998). The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences*, 12(1), 109–124.

[48] R.L. Tweedie. (1981). Criteria for ergodicity, exponential ergodicity and strong ergodicity of Markoc processes. *Journal of Applied Probability*, 18(1), 122–130.

[49] N.D. Vvedenskaya, R.L.V. Dobrushin, and F.I. Karpelevich. (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1), 20–34.

[50] R.W. Weber. (1978). On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* 15(2), 406–413.

[51] W. Whitt. (1985). Deciding which queue to join: some counterexamples. *Operations Research*, 34(1), 55–62.

[52] W. Whitt. (2002). *Stochastic Process Limits*, Springer, New York.

[53] W. Winston. (1977). Optimality of the shortest line discipline. *Journal of Applied Probability* 14(1), 181–189.

[54] J. Xu and B. Hajek. (2013). The supermarket game, *Stochastic Systems*, 3(2), 405–441.