Asymptotic optimality of power-of-d load balancing in large-scale systems

Asymptotic Optimality of Power-of-d Load Balancing in Large-Scale Systems

Debankur Mukherjee[*1], Sem C. Borst[1,2],
Johan S.H. van Leeuwaarden[1] , Philip A. Whiting[3]

[1]*Eindhoven University of Technology, The Netherlands*
[2]*Nokia Bell Labs, Murray Hill, NJ, USA*
[3]*Macquarie University, North Ryde, NSW, Australia*

December 5, 2016

Abstract

We consider a system of N identical server pools and a single dispatcher where tasks arrive as a Poisson process of rate $\lambda(N)$. Arriving tasks cannot be queued, and must immediately be assigned to one of the server pools to start execution, or discarded. The execution times are assumed to be exponentially distributed with unit mean, and do not depend on the number of other tasks receiving service. However, the experienced performance (e.g. in terms of received throughput) does degrade with an increasing number of concurrent tasks at the same server pool. The dispatcher therefore aims to evenly distribute the tasks across the various server pools. Specifically, when a task arrives, the dispatcher assigns it to the server pool with the minimum number of tasks among $d(N)$ randomly selected server pools. This assignment strategy is called the JSQ($d(N)$) scheme, as it resembles the power-of-d version of the Join-the-Shortest-Queue (JSQ) policy, and will also be referred to as such in the special case $d(N) = N$.

We construct a stochastic coupling to bound the difference in the system occupancy processes between the JSQ policy and a scheme with an arbitrary value of $d(N)$. We use the coupling to derive the fluid limit in case $d(N) \to \infty$ and $\lambda(N)/N \to \lambda$ as $N \to \infty$, along with the associated fixed point. The fluid limit turns out to be insensitive to the exact growth rate of $d(N)$, and coincides with that for the JSQ policy. We further leverage the coupling to establish that the diffusion limit corresponds to that for the JSQ policy as well, as long as $d(N)/\sqrt{N}\log(N) \to \infty$, and characterize the common limiting diffusion process. These results indicate that the JSQ optimality can be preserved at the fluid-level and diffusion-level while reducing the overhead by nearly a factor $O(N)$ and $O(\sqrt{N}/\log(N))$, respectively.

---

[*]d.mukherjee@tue.nl

Contents

1 Introduction

In the present paper we establish asymptotic optimality for a broad class of randomized load balancing strategies. While the specific features of load balancing policies may considerably differ, the principal purpose is to distribute service requests or tasks among servers or distributed resources in parallel-processing systems. Well-designed load balancing schemes provide an effective mechanism for improving relevant performance metrics experienced by users while achieving high resource utilization levels. The analysis and design of load balancing policies has attracted strong renewed interest in the last several years, mainly motivated by significant challenges involved in assigning tasks (e.g. file transfers, compute jobs, data base look-ups) to servers in large-scale data centers, see for instance [20].

Load balancing schemes can be broadly categorized as static (open-loop), dynamic (closed-loop), or some intermediate blend, depending on the amount of real-time feedback or state information (e.g. queue lengths or load measurements) that is used in assign-

ing tasks. Within the category of dynamic policies, one can further distinguish between push-based and pull-based approaches, depending on whether the initiative resides with a dispatcher actively collecting feedback from the servers, or with the servers advertizing their availability or load status. The use of state information naturally allows dynamic policies to achieve better performance and greater resource pooling gains, but also involves higher implementation complexity and potentially substantial communication overhead. The latter issue is particularly pertinent in large-scale data centers, which deploy thousands of servers and handle massive demands, with service requests coming in at huge rates.

In the present paper we focus on a basic scenario of N identical parallel server pools and a single dispatcher where tasks arrive as a Poisson process. Incoming tasks cannot be queued, and must immediately be dispatched to one of the server pools to start execution, or discarded. The execution times are assumed to be exponentially distributed, and do not depend on the number of other tasks receiving service, but the experienced performance (e.g. in terms of received throughput or packet-level delay) does degrade in a convex manner with an increasing number of concurrent tasks. These characteristics pertain for instance to video streaming sessions and various interactive applications. In contrast to elastic data transfers or computing-intensive jobs, the duration of such sessions is hardly affected by the number of contending service requests. The perceived performance in terms of video quality or packet-level latency however strongly varies with the number of concurrent tasks, creating an incentive to distribute the incoming tasks across the various server pools as evenly as possible.

Specifically, adopting the usual time scale separation assumption, suppose that the task-perceived performance at a particular server pool can be described as some function $F(x)$ of the instantaneous number of concurrent tasks x, and let $X = (X_1, \dots, X_N)$, with X_n the number of active tasks at the nth server pool. Then $G(X) = \sum_{n=1}^{N} X_n F(X_n) / \sum_{n=1}^{N} X_n$, provides a proxy for the instantaneous overall system performance. In many situations, the function $G(\cdot)$ tends to be either Schur-convex of Schur-concave. For example, if $F(\cdot)$ is convex increasing (for instance average packet-level delay), then $G(\cdot)$ is Schur-convex, i.e., $G(X) \leqslant G(Y)$ if X is majorized by Y, i.e., X is 'more balanced' than Y with $\sum_{n=1}^{N} X_n = \sum_{n=1}^{N} Y_n$. Likewise, if $F(X_n) = U(1/X_n)$, with $U(\cdot)$ is concave increasing (for instance throughput utility), then $G(\cdot)$ is Schur-concave, i.e., $G(X) \geqslant G(Y)$ if X is majorized by Y.

The so-called Join-the-Shortest-Queue (JSQ) policy has primarily been considered for load balancing among parallel *single-server queues* where it furnishes several strong optimality guarantees [3, 28, 31]. Its fundamental ability of optimally balancing tasks across parallel resources also translates however into crucial optimality properties with respect to the performance criterion $G(\cdot)$ in the present context with infinite-server dynamics. In particular, let $X^{\Pi}(t) = (X_1^{\Pi}(t), \dots, X_N^{\Pi}(t))$, with $X_n^{\Pi}(t)$ denoting the number of active tasks at the nth server pool at time t under a task assignment scheme Π. Then, given the same initial conditions and in the absence of any blocking, $\{X^{\text{JSQ}}(t)\}_{t \geqslant 0}$ is majorized by $\{X^{\Pi}(t)\}_{t \geqslant 0}$ for any non-anticipating task assignment scheme Π [12, 13, 22, 23, 24]. Thus $G(X^{\text{JSQ}}(t))$ is either stochastically smaller or larger than $G(X^{\Pi}(t))$ at all times t for any task assignment scheme Π, depending on whether the function $G(\cdot)$ is Schur-convex or Schur-concave. In a scenario where each server pool can only accommodate a maximum of $B < \infty$ simultaneous tasks, the JSQ policy belongs to the class of policies that

stochastically minimize the total cumulative number of blocked tasks over any time interval $[0, t]$ [22, 25].

In order to implement the JSQ policy, a dispatcher requires instantaneous knowledge of the numbers of tasks at all the server pools, which may give rise to a substantial communication burden, and may not be scalable in scenarios with large numbers of server pools. The latter issue has motivated consideration of so-called JSQ(d) policies, where the dispatcher assigns an incoming task to a server pool with the minimum number of active tasks among d randomly selected server pools. Mean-field limit theorems in Mitzenmacher [14] and Vvedenskaya *et al.* [27] indicate that even a value as small as $d = 2$ yields significant performance improvements in a single-server queueing regime with $N \to \infty$, in the sense that the tail of the queue length distribution at each individual server falls off much more rapidly compared to a strictly random assignment policy ($d = 1$). This is commonly referred to as the "power-of-two" effect. Work of Turner [26] and recent papers by Mukhopadhyay *et al.* [17, 18] and Xie *et al.* [32] have shown similar power-of-two properties for loss probabilities in a *blocking* scenario with infinite-server dynamics as described above.

As illustrated by the above, the diversity parameter d induces a fundamental trade-off between the amount of communication overhead and the performance in terms of blocking probabilities or throughputs. For example, a strictly random assignment policy can be implemented with zero overhead, but for any finite buffer capacity $B < \infty$ the blocking probability does *not* fall to zero as $N \to \infty$. In contrast, a nominal implementation of the JSQ policy (without maintaining state information at the dispatcher) involves $O(N)$ overhead per task, but it can be shown that for any subcritical load, the blocking probability vanishes as $N \to \infty$. As mentioned above, JSQ(d) strategies with a fixed parameter $d \geqslant 2$ yield significant performance improvements over purely random task assignment while reducing the overhead by a factor $O(N)$ compared to the JSQ policy. However, the blocking probability does *not* vanish in the limit, and in that sense a fixed value of d is not sufficient to achieve asymptotically optimal performance.

In order to gain further insight in the trade-off between performance and communication overhead as governed by the diversity parameter d, we also consider a regime where the number of servers N grows large, but allow the value of d to depend on N, and write $d(N)$ to explicitly reflect that. For convenience, we assume a Poisson arrival process of rate $\lambda(N)$ and unit-mean exponential service requirements.

We construct a stochastic coupling to bound the difference in the system occupancy processes between the JSQ policy and a scheme with an arbitrary value of $d(N)$. We exploit the coupling to obtain the fluid limit in case $\lambda(N)/N \to \lambda < B$ and $d(N) \to \infty$ as $N \to \infty$, along with the associated fixed point. As it turns out, the fluid limit is insensitive to the exact growth rate of $d(N)$, and in particular coincides with that for the ordinary JSQ policy. This implies that the overhead of the JSQ policy can be reduced by almost a factor $O(N)$ while maintaining fluid-level optimality.

We further consider the diffusion limit of the system occupancy states, and consider the infinite-server dynamics analog of the Halfin-Whitt regime. We leverage the above-mentioned coupling to prove that the diffusion limit in case $d(N)/(\sqrt{N} \log(N)) \to \infty$ corresponds to that for the ordinary JSQ policy, and characterize the common limiting diffusion process. This indicates that the overhead of the JSQ policy can be reduced by almost a factor $O(\sqrt{N}/\log(N))$ while retaining diffusion-level optimality.

The above results mirror fluid-level and diffusion-level optimality properties reported in the companion paper [16] for the power-of-d(N) strategies in a scenario with single-server queues. As it turns out, however, the infinite-server dynamics in the present paper require a fundamentally different coupling argument to establish asymptotic equivalence. In particular, for the single-server dynamics, first the servers are ordered according to the number of active tasks, and the departures at the ordered servers under two different policies are then coupled. In contrast, for the infinite-server dynamics, the departure rate at the ordered server pools can vary depending on the exact number of active tasks. Therefore, the departure processes under two different policies cannot be coupled as before, which necessitates the construction of a novel stochastic coupling. Specifically, one can think of the coupling for the single-server dynamics as one-dimensional (depending only upon the ordering of the servers), while the coupling we introduce in this paper is two-dimensional, with the server ordering as one coordinate and the number of tasks as the other, as will be explained in greater detail later. We further elaborate on the necessity and novelty of the coupling methodology developed in the current paper, and reflect on the contrast with the stochastic optimality results for the JSQ policy in the existing literature and the coupling technique in [16] in Remarks 4.3 and 4.4. In addition, the fluid- and diffusion-limit results in the infinite-server scenario are also notably different from those in [16]. More specifically, we extend the fluid-limit result in [16, Theorem 4.1] to a more general class of assignment probabilities and departure rate functions, and depending on whether the scaled arrival rate converges to an integer or not, obtain a qualitatively different behavior of the occupancy state process on diffusion scale. Furthermore, the diffusion limit result in [16, Theorem 2.4] characterizes the diffusion-scale behavior only in the transient regime, whereas in the current paper we are able to analyze the steady-state behavior as well.

The remainder of the paper is organized as follows. In Section 2 we present a detailed model description, and provide an overview of the main results. In Section 3 we explain the proof outline and introduce a notion of asymptotic equivalence of two assignment schemes. Section 4 introduces a stochastic coupling between any two schemes, and proves the asymptotic equivalence results. Sections 4–7 contain the proofs of the main results, and in Section 8 we reflect upon various performance implications. We conclude in Section 9 with some pointers to open problems and future research.

2 Main Results

2.1 Model description and notation

Consider a system with N parallel identical server pools and a single dispatcher where tasks arrive as a Poisson process of rate $\lambda(N)$. Arriving tasks cannot be queued, and must immediately be assigned to one of the server pools to start execution. The execution times are assumed to be exponentially distributed with unit mean, and do not depend on the number of other tasks receiving service. Each server pool is however only able to accommodate a maximum of B simultaneous tasks (possibly $B = \infty$), and when a task is allocated to a server pool that is already handling B active tasks, it gets permanently discarded.

Specifically, when a task arrives, the dispatcher assigns it to the server pool with
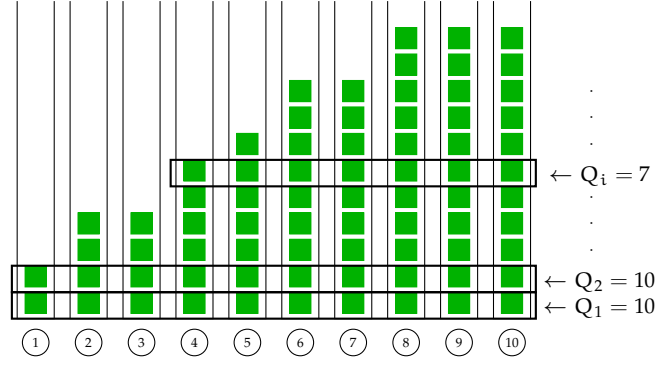
Figure 1: The occupancy state of the system; When the server pools are arranged in nondecreasing order of the number of active tasks, Q_i represents the width of the ith row, as shown above.

the minimum number of active tasks among $d(N)$ randomly selected server pools ($1 \leqslant d(N) \leqslant N$). As mentioned earlier, this assignment strategy is called a JSQ($d(N)$) scheme, as it closely resembles the power-of-d version of the Join-the-Shortest-Queue (JSQ) policy, and will also consisely be referred to as such in the special case $d(N) = N$. We will consider an asymptotic regime where the number of server pools N and the task arrival rate $\lambda(N)$ grow large in proportion, with $\lambda(N)/N \to \lambda \leqslant B$ as $N \to \infty$. For convenience, we denote $K = \lfloor \lambda \rfloor$ and $f = \lambda - K \in [0, 1)$.

For any $d(N)$ ($1 \leqslant d(N) \leqslant N$), let $\mathbf{Q}^{d(N)}(t) = (Q_1^{d(N)}(t), Q_2^{d(N)}(t), \ldots, Q_B^{d(N)}(t))$ be the system occupancy state, where $Q_i^{d(N)}(t)$ is the number of server pools under the JSQ($d(N)$) scheme with i or more active tasks at time t, $i = 1, \ldots, B$. A schematic diagram of the Q_i-values is provided in Figure 1. We occasionally omit the superscript $d(N)$, and replace it by N, to refer to the Nth system, when the value of $d(N)$ is clear from the context. In case of a finite buffer size $B < \infty$, when a task is discarded, we call it an *overflow* event, and we denote by $L^{d(N)}(t)$ the total number of overflow events under the JSQ($d(N)$) policy up to time t.

Throughout we assume that at each arrival epoch the server pools are ordered in nondecreasing order of the number of active tasks (ties can be broken arbitrarily), see Figure 1, and whenever we refer to some ordered server pool, it should be understood with respect to this prior ordering, unless mentioned otherwise.

Boldfaced letters will be used to denote vectors. A sequence of random variables $\{X_N\}_{N \geqslant 1}$ is said to be $O_P(g(N))$, or $o_P(g(N))$, for some function $g : \mathbb{N} \to \mathbb{R}_+$, if the sequence of scaled random variables $\{X_N/g(N)\}_{N \geqslant 1}$ is a tight sequence, or converges to zero in probability, respectively. Whenever we mention 'with high probability', it should be understood as 'with probability tending to 1 as the underlying scaling parameter tends to infinity'. For stochastic boundedness of a process we refer to [19, Definition 5.4]. Also, f will be called 'diverging to infinity' if $g(N) \to \infty$ as $N \to \infty$. For any complete separable metric space E, denote by $D_E[0, \infty)$, the set of all E-valued cádlág (right continuous with left limit exists) processes. By the symbol '$\xrightarrow{\mathcal{L}}$' we denote convergence in distribution for real-valued random variables, and with respect to Skorohod-J_1 topology for *cádlág* processes.

6

2.2 Fluid-limit results

In order to state the fluid-limit results, we first introduce some useful notation. Denote the fluid-scaled system occupancy state by $\mathbf{q}^{d(N)}(t) := \mathbf{Q}^{d(N)}(t)/N$. We will denote by $\tilde{S} = \{\mathbf{Q} \in \mathbb{Z}^B : Q_i \leqslant Q_{i-1} \text{ for all } i = 2, \ldots, B\}$ and $S = \{\mathbf{q} \in [0,1]^B : q_i \leqslant q_{i-1} \text{ for all } i = 2, \ldots, B\}$ the set of all possible unscaled and fluid-scaled occupancy states, respectively. Further define $S^N := S \cap \{i/N : 1 \leqslant i \leqslant N\}^B$ as the space of all fluid-scaled occupancy states of the Nth system. We take the following product norm on S: for $\mathbf{q}_1 = (q_{1,1}, q_{1,2}, \ldots, q_{1,B})$, $\mathbf{q}_2 = (q_{2,1}, q_{2,2}, \ldots, q_{2,B}) \in S$,

$$\rho(\mathbf{q}_1, \mathbf{q}_2) := \sum_{i=1}^{B} \frac{|q_{1,i} - q_{2,i}|}{2^i},$$

and all the convergence results below will be with respect to product topology. We often write $\rho(\mathbf{q}_1, \mathbf{q}_2)$ as $\|\mathbf{q}_1 - \mathbf{q}_2\|$. Let $(E, \hat{\rho})$ be a metric space. We call a function $g : S \to E$ Lipschitz continuous on S, if there exists $L > 0$, such that for all $x, y \in S$,

$$\hat{\rho}(g(x), g(y)) \leqslant L\|x - y\|.$$

For any $\mathbf{q} \in S$, denote by $m(\mathbf{q}) = \min\{i : q_{i+1} < 1\}$ the minimum number of active tasks among all server pools, with the convention that $q_{B+1} = 0$ if $B < \infty$. Now distinguish two cases, depending on whether the normalized arrival rate λ is larger than $m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$ or not. If $\lambda \leqslant m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$, then define

$$p_{m(\mathbf{q})-1}(\mathbf{q}) = 1, \quad \text{and} \quad p_i(\mathbf{q}) = 0 \quad \text{for all} \quad i \neq m(\mathbf{q}) - 1.$$

On the other hand, if $\lambda > m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$, then

$$p_i(\mathbf{q}) = \begin{cases} m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})/\lambda & \text{for} \quad i = m(\mathbf{q}) - 1, \\ 1 - p_{m(\mathbf{q})-1}(\mathbf{q}) & \text{for} \quad i = m(\mathbf{q}), \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Note that the assumption $\lambda \leqslant B$ ensures that the latter case cannot occur when $B < \infty$ and $m(\mathbf{q}) = B$.

Theorem 2.1. (Universality of fluid limit for JSQ($d(N)$) scheme) *Assume* $\mathbf{q}^{d(N)}(0) \xrightarrow{\mathbb{P}} \mathbf{q}^\infty \in S$ *as* $N \to \infty$. *For the JSQ($d(N)$) scheme with* $d(N)$ *diverging to infinity, the sequence of processes* $\{\mathbf{q}^{d(N)}(t)\}_{t \geqslant 0}$ *has a weak limit* $\{\mathbf{q}(t)\}_{t \geqslant 0}$ *that satisfies the system of integral equations*

$$q_i(t) = q_i(0) + \lambda \int_0^t p_{i-1}(\mathbf{q}(s))ds - i \int_0^t (q_i(s) - q_{i+1}(s))ds, \quad i = 1, \ldots, B,$$

where $\mathbf{q}(0) = \mathbf{q}^\infty$ *and the coefficients* $p_i(\cdot)$ *are as defined above.*

The above theorem shows that the fluid-level dynamics do not depend on the specific growth rate of $d(N)$ as long as $d(N) \to \infty$ as $N \to \infty$. In particular, the JSQ($d(N)$) scheme with $d(N) \to \infty$ as $N \to \infty$ exhibits the same behavior as the ordinary JSQ policy, and thus achieves fluid-level optimality. This result can be intuitively interpreted as follows. Since $d(N)$ is growing, for large N, at an arrival epoch, if the fraction of server pools

7

with the minimum number of active tasks becomes positive, then with high probability at least one of the $d(N)$ selected server pools will be from the ones with the minimum number of active tasks. This ensures that as long as $d(N) \to \infty$ as $N \to \infty$, the difference in Q_i-values between the ordinary JSQ policy and the JSQ$(d(N))$ scheme can not become $O(N)$, yielding fluid-level optimality.

The coefficient $p_i(\mathbf{q})$ represents the fraction of incoming tasks assigned to server pools with exactly i active tasks in the fluid-level state $\mathbf{q} \in S$. Assuming $m(\mathbf{q}) < B$, a strictly positive fraction $1 - q_{m(\mathbf{q})+1}$ of the server pools have exactly $m(\mathbf{q})$ active tasks. Since $d(N) \to \infty$ as $N \to \infty$, the fraction of incoming tasks that get assigned to server pools with $m(\mathbf{q}) + 1$ or more active tasks is therefore zero: $p_i(\mathbf{q}) = 0$ for all $i = m(\mathbf{q}) + 1, \ldots, B - 1$. Also, tasks at server pools with exactly i active tasks are completed at (normalized) rate $i(q_i - q_{i+1})$, which is zero for all $i = 1, \ldots, m(\mathbf{q}) - 1$, and hence the fraction of incoming tasks that get assigned to server pools with $m(\mathbf{q}) - 2$ or less active tasks is zero as well: $p_i(\mathbf{q}) = 0$ for all $i = 0, \ldots, m(\mathbf{q}) - 2$. This only leaves the fractions $p_{m(\mathbf{q})-1}(\mathbf{q})$ and $p_{m(\mathbf{q})}(\mathbf{q})$ to be determined. Now observe that the fraction of server pools with exactly $m(\mathbf{q}) - 1$ active tasks is zero. However, since tasks at server pools with exactly $m(\mathbf{q})$ active tasks are completed at (normalized) rate $m(\mathbf{q})(1 - q_{m(\mathbf{q})+1}) > 0$, incoming tasks can be assigned to server pools with exactly $m(\mathbf{q}) - 1$ active tasks at that rate. We thus need to distinguish between two cases, depending on whether the normalized arrival rate λ is larger than $m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$ or not. If $\lambda \leqslant m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$, then all the incoming tasks can be assigned to server pools with exactly $m(\mathbf{q}) - 1$ active tasks, so that $p_{m(\mathbf{q})-1}(\mathbf{q}) = 1$ and $p_{m(\mathbf{q})}(\mathbf{q}) = 0$. On the other hand, if $\lambda > m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})$, then not all incoming tasks can be assigned to server pools with exactly $m(\mathbf{q}) - 1$ active tasks, and a positive fraction will be assigned to server pools with exactly $m(\mathbf{q})$ active tasks: $p_{m(\mathbf{q})-1}(\mathbf{q}) = m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})/\lambda$ and $p_{m(\mathbf{q})}(\mathbf{q}) = 1 - p_{m(\mathbf{q})-1}(\mathbf{q})$.

It is easily verified that the unique fixed point of the differential equation in Theorem 2.1 is given by

$$q_i^\star = \begin{cases} 1 & i = 1, \ldots, K \\ f & i = K + 1 \\ 0 & i = K + 2, \ldots, B, \end{cases} \tag{2.2}$$

and thus $\sum_{i=1}^{B} q_i^\star = \lambda$. This is consistent with the results in Mukhopadhyay *et al.* [17, 18] and Xie *et al.* [32] for fixed d, where taking $d \to \infty$ yields the same fixed point. However, the results in [17, 18, 32] for fixed d cannot be directly used to handle joint scalings, and do not yield the universality of the entire fluid-scaled sample path for arbitrary initial states as established in Theorem 2.1.

Having obtained the fixed point of the fluid limit, we now establish the interchange of the mean-field ($N \to \infty$) and stationary ($t \to \infty$) limits. Let

$$\pi^{d(N)}(\cdot) = \lim_{t \to \infty} \mathbb{P}\left(\mathbf{q}^{d(N)}(t) = \cdot\right)$$

be the stationary measure of the occupancy states of the Nth system.

Proposition 2.2 (Interchange of limits). *The sequence of stationary measures $\left\{\pi^{d(N)}\right\}_{N \geqslant 1}$ with $d(N) \to \infty$ as $N \to \infty$ converges weakly to $\pi^\star$, where $\pi^\star = \delta_{\mathbf{q}^\star}$ with δ_x being the Dirac measure concentrated upon x, and $\mathbf{q}^\star$ defined by (2.2).*

Proof. Observe that $\pi^{d(N)}$ is defined on S, and S is a compact set when endowed with the product topology. Prohorov's theorem implies that the sequence $\{\pi^{d(N)}\}_{N \geqslant 1}$ is relatively compact, and hence, has a convergent subsequence. Let $\{\pi^{d(N_n)}\}_{n \geqslant 1}$ be a convergent subsequence, with $\{N_n\}_{n \geqslant 1} \subset \mathbb{N}$, such that $\pi^{d(N_n)} \xrightarrow{\mathcal{L}} \hat{\pi}$. We show that $\hat{\pi}$ is unique and equals the measure $\pi^\star$.

First of all note that if $\mathbf{q}^{d(N_n)}(0) \sim \pi^{d(N_n)}$, then $\mathbf{q}^{d(N_n)}(t) \sim \pi^{d(N_n)}$. Also, the fact that $\mathbf{q}^{d(N_n)}(t) \xrightarrow{\mathcal{L}} \mathbf{q}(t)$, and $\pi^{d(N_n)} \xrightarrow{\mathcal{L}} \hat{\pi}$, means that $\hat{\pi}$ is a fixed point of the deterministic process $\{\mathbf{q}(t)\}_{t \geqslant 0}$. Since the latter fixed point is unique, $\mathbf{q}^\star$, we can conclude the desired convergence of the stationary measure. $\qquad\square$

2.3 Diffusion-limit results for non-integral λ

As it turns out, the diffusion-limit results may be qualitatively different, depending on whether $f = 0$ or $f > 0$, and we will distinguish between these two cases accordingly. Observe that for any assignment scheme, in the absence of overflow events, the total number of active tasks evolves as the number of jobs in an M/M/∞ system with arrival rate $\lambda(N)$ and unit service rate, for which the diffusion limit is well-known [21]. For the JSQ($d(N)$) scheme with $d(N)/(\sqrt{N}\log(N)) \to \infty$ as $N \to \infty$, we can establish, for suitable initial conditions, that the total number of server pools with $K - 2$ or less and $K + 2$ or more tasks is negligible on the diffusion scale. If $f > 0$, the number of server pools with $K - 1$ tasks is negligible as well, and the dynamics of the number of server pools with K or $K + 1$ tasks can then be derived from the known diffusion limit of the total number of tasks mentioned above. In contrast, if $f = 0$, the number of server pools with $K - 1$ tasks is not negligible on the diffusion scale, and the limiting behavior is qualitatively different, but can still be characterized.

We first consider the case $f > 0$, and define $f(N) := \lambda(N) - KN$. Based on the above observations, we define the following centered and scaled processes:

$$\bar{Q}_i^{d(N)}(t) := \frac{N - Q_i^{d(N)}(t)}{\sqrt{N}} \geqslant 0, \quad i \leqslant K,$$

$$\bar{Q}_{K+1}^{d(N)}(t) := \frac{Q_{K+1}^{d(N)}(t) - f(N)}{\sqrt{N}} \in \mathbb{R}, \tag{2.3}$$

$$\bar{Q}_i^{d(N)}(t) := Q_i^{d(N)}(t) \geqslant 0, \quad \text{for} \quad i \geqslant K + 2.$$

Theorem 2.3. (Universality of diffusion limit for JSQ($d(N)$) scheme, $f > 0$) *If $f > 0$, $\bar{Q}_{K+1}^{d(N)}(0) \xrightarrow{\mathbb{P}} \bar{Q}_{K+1} \in \mathbb{R}$, $\bar{Q}_i^{d(N)}(0) \xrightarrow{\mathbb{P}} 0$ for $i \neq K + 1$, and $d(N)/(\sqrt{N}\log(N)) \to \infty$ as $N \to \infty$, then the following holds as $N \to \infty$:*

(i) *For $i \leqslant K$, $\{\bar{Q}_i^{d(N)}(t)\}_{t \geqslant 0} \xrightarrow{\mathcal{L}} \{\bar{Q}_i(t)\}_{t \geqslant 0}$, where $\bar{Q}_i(t) \equiv 0$.*

(ii) *$\{\bar{Q}_{K+1}^{d(N)}(t)\}_{t \geqslant 0} \xrightarrow{\mathcal{L}} \{\bar{Q}_{K+1}(t)\}_{t \geqslant 0}$, where $\bar{Q}_{K+1}(t)$ is given by the Ornstein-Uhlenbeck process satisfying the following stochastic differential equation:*

$$d\bar{Q}_{K+1}(t) = -\bar{Q}_{K+1}(t)dt + \sqrt{2\lambda}dW(t), \tag{2.4}$$

where $W(t)$ is the standard Brownian motion.

(iii) *For* $i \geqslant K + 2$, $\left\{ \bar{Q}_i^{d(N)}(t) \right\}_{t \geqslant 0} \xrightarrow{\mathcal{L}} \left\{ \bar{Q}_i(t) \right\}_{t \geqslant 0}$, *where* $\bar{Q}_i(t) \equiv 0$.

Loosely speaking, the above theorem says that, if $f > 0$ and $d(N)/(\sqrt{N} \log(N)) \to \infty$ as $N \to \infty$, then over any finite time horizon, there will only be $o_P(\sqrt{N})$ server pools with fewer than K or more than $K + 1$ active tasks, and $fN + O_P(\sqrt{N})$ server pools with precisely $K + 1$ active tasks. Also, as long as $d(N)/(\sqrt{N} \log(N)) \to \infty$ as $N \to \infty$, the JSQ($d(N)$) scheme exhibits the same behavior as the ordinary JSQ policy (i.e., $d(N) = N$), and thus achieves diffusion-level optimality. The result can be heuristically explained as follows. When the number of server pools with the minimum number of active tasks is $O(\sqrt{N})$, the JSQ($d(N)$) scheme should be able to assign the incoming tasks with high probability to one of the server pools with the minimum number of active tasks. To be able to select one of the $O(\sqrt{N})$ server pool out of N server pools, $d(N)$ must grow faster than $\sqrt{N}$. Now further observe that in any finite time interval there are on average $O(N)$ arrivals, and hence it is not enough to assign the incoming task to the appropriate server pool only once. The number of times that the JSQ($d(N)$) scheme fails to assign a task to the 'appropriate' server pool in any finite time interval, should be $o_P(\sqrt{N})$. This gives rise to the additional $\log(N)$ factor in the growth rate of $d(N)$.

2.4 Diffusion-limit results for integral λ

We now turn to the case $f = 0$, and assume that

$$\frac{KN - \lambda(N)}{\sqrt{N}} \to \beta \in \mathbb{R} \quad \text{as} \quad N \to \infty, \tag{2.5}$$

which can be thought of as an analog of the so-called Halfin-Whitt regime [7]. As mentioned above, the limiting behavior in this case is qualitatively different from the case $f > 0$. Hence, we now consider the following scaled quantities:

$$
\begin{aligned}
\hat{Q}_{K-1}^{d(N)}(t) &:= \sum_{i=1}^{K-1} \frac{N - Q_i^{d(N)}(t)}{\sqrt{N}} \geqslant 0, \\
\hat{Q}_K^{d(N)}(t) &:= \frac{N - Q_K^{d(N)}(t)}{\sqrt{N}} \geqslant 0, \\
\hat{Q}_i^{d(N)}(t) &:= \frac{Q_i^{d(N)}(t)}{\sqrt{N}} \geqslant 0, \quad \text{for } i \geqslant K + 1.
\end{aligned}
\tag{2.6}
$$

Theorem 2.4. (Universality of diffusion limit for JSQ($d(N)$) scheme, $f = 0$) *Suppose there exists* $M \geqslant K + 1$, *such that* $Q_{M+1}^{d(N)}(0) \equiv 0$, *and*

$$(\hat{Q}_{K-1}^{d(N)}(0), \hat{Q}_K^{d(N)}(0), \ldots, \hat{Q}_M^{d(N)}(0)) \xrightarrow{\mathcal{L}} (\hat{Q}_{K-1}(0), \hat{Q}_K(0), \ldots, \hat{Q}_M(0))$$

in $\mathbb{R}^{M-K+2}$. *If* $f = 0$, $d(N)/(\sqrt{N} \log(N)) \to \infty$, *Equation* (2.5) *is satisfied, and* $\hat{Q}_{K-1}^{d(N)}(0) \xrightarrow{\mathbb{P}} 0$, *as* $N \to \infty$, *then the process* $\left\{ (\hat{Q}_{K-1}^{d(N)}(t), \hat{Q}_K^{d(N)}(t), \ldots, \hat{Q}_M^{d(N)}(t), \hat{Q}_{M+1}^{d(N)}(t)) \right\}_{t \geqslant 0}$ *converges*

weakly to the process defined as the unique solution to the stochastic integral equation

$$\hat{Q}_K(t) = \hat{Q}_K(0) + \sqrt{2K}W(t) - \int_0^t (\hat{Q}_K(s) + K\hat{Q}_{K+1}(s))ds + \beta t + V_1(t)$$

$$\hat{Q}_{K+1}(t) = \hat{Q}_{K+1}(0) + V_1(t) - (K+1)\int_0^t (\hat{Q}_{K+1}(s) - \hat{Q}_{K+2}(s))ds,$$

$$\hat{Q}_i(t) = \hat{Q}_i(0) - i\int_0^t (\hat{Q}_i(s) - \hat{Q}_{i+1}(s))ds, \quad i = K+2\dots, M-1, \tag{2.7}$$

$$\hat{Q}_M(t) = \hat{Q}_M(0) - M\int_0^t \hat{Q}_M(s)ds,$$

$\hat{Q}_{K-1}(t) \equiv 0$, and $\hat{Q}_{M+1}(t) \equiv 0$, where $W(t)$ is the standard Brownian motion, and $V_1(t)$ is the unique non-decreasing process in $D_{\mathbb{R}_+}[0, \infty)$ satisfying

$$\int_0^t \mathbb{1}_{[\hat{Q}_K(s) \geqslant 0]} dV_1(s) = 0.$$

Unlike the $f > 0$ case, the above theorem says that, if $f = 0$, then over any finite time horizon, there will be $O_P(\sqrt{N})$ server pools with fewer than K or more than K active tasks, and hence most of the server pools have precisely K active tasks.

3 Proof Outline

The proofs of the asymptotic results for the JSQ($d(N)$) scheme in Theorems 2.1, 2.3, and 2.4 involve two main components:

(i) deriving the relevant limiting processes for the ordinary JSQ policy,

(ii) establishing a universality result which shows that the limiting processes for the JSQ($d(N)$) scheme are 'asymptotically equivalent' to those for the ordinary JSQ policy for suitably large $d(N)$.

For Theorems 2.1, 2.3 and 2.4, part (i) will be dealt with in Theorems 5.1, 6.1 and 7.1, respectively. For all three theorems, part (ii) relies on a notion of asymptotic equivalence between different schemes, which is formalized in the next definition.

Definition 1. *Let Π_1 and Π_2 be two schemes parameterized by the number of server pools N. For any positive function $g : \mathbb{N} \to \mathbb{R}_+$, we say that Π_1 and Π_2 are 'g(N)-alike' if there exists a common probability space, such that for any fixed $T \geqslant 0$, for all $i \geqslant 1$,*

$$\sup_{t \in [0,T]} (g(N))^{-1} |Q_i^{\Pi_1}(t) - Q_i^{\Pi_2}(t)| \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad N \to \infty.$$

Intuitively speaking, if two schemes are g(N)-alike, then in some sense, the associated system occupancy states are indistinguishable on g(N)-scale. For brevity, for two schemes Π_1 and Π_2 that are g(N)-alike, we will often say that Π_1 and Π_2 have the same process-level limits on g(N)-scale. The next theorem states a sufficient criterion for the JSQ($d(N)$) scheme and the ordinary JSQ policy to be g(N)-alike, and thus, provides the key vehicle in establishing the universality result in part (ii) mentioned above.

Theorem 3.1. *Let* $g : \mathbb{N} \to \mathbb{R}_+$ *be a function diverging to infinity. Then the JSQ policy and the JSQ(d(N)) scheme are* $g(N)$-*alike, with* $g(N) \leqslant N$, *if*

$$\text{(i)} \quad d(N) \to \infty, \quad \textit{for} \quad g(N) = O(N), \tag{3.1}$$

$$\text{(ii)} \quad d(N) \left(\frac{N}{g(N)} \log \left(\frac{N}{g(N)} \right) \right)^{-1} \to \infty, \quad \textit{for} \quad g(N) = o(N). \tag{3.2}$$

Theorem 3.1 can be intuitively explained as follows. The choice of $d(N)$ should be such that the JSQ($d(N)$) scheme, at each arrival, with high probability selects one of the server pools with the minimum number of tasks, if the total number of server pools with the minimum number of tasks is of order $g(N)$. Moreover, in any finite time interval, the total number of times it fails to do so, should be of order lower than that of $g(N)$. These conditions imply that $d(N)$ must diverge if $g(N) = O(N)$, or grow faster than $(N/g(N)) \log(N/g(N))$, if $g(N) = o(N)$.

In order to obtain the fluid and diffusion limits for various schemes, the two main scales that we consider are $g(N) \sim N$ and $g(N) \sim \sqrt{N}$, respectively. The next two immediate corollaries of the above theorem will imply that it is enough to investigate the ordinary JSQ policy in various regimes.

Corollary 3.2. *If* $d(N) \to \infty$ *as* $N \to \infty$, *then the JSQ(d(N)) scheme and the ordinary JSQ policy are* N-*alike.*

Remark 3.3. The growth condition on $d(N)$ in order for the JSQ($d(N)$) scheme to be N-alike to the ordinary JSQ policy, stated in the above corollary, is not only sufficient, but also necessary. Specifically, if $\liminf_{N \to \infty} d(N) \leqslant d < \infty$, then consider a subsequence along which the limit of $d(N)$ exists and is uniformly bounded by d. Therefore, one can choose a further subsequence, such that $d(N) = d$ for all N along that subsequence. Now, from the fluid-limit result for the JSQ(d) scheme [17, 18], one can see that it differs from that of the JSQ policy stated in (2.1), and hence the JSQ($d(N)$) scheme is not N-alike to the ordinary JSQ policy.

Corollary 3.4. *If* $d(N)/(\sqrt{N} \log(N)) \to \infty$ *as* $N \to \infty$, *then the JSQ(d(N)) scheme and the ordinary JSQ policy are* $\sqrt{N}$-*alike.*

We will prove the universality result in Theorem 3.1 in the next section. The key challenge is that a direct comparison of the JSQ($d(N)$) scheme and the ordinary JSQ policy is not straightforward. Hence, to compare the JSQ($d(N)$) scheme with the JSQ policy, we adopt a two-stage approach based on a novel class of schemes, called CJSQ($n(N)$), as a convenient intermediate scenario. Specifically, for some nonnegative integer-valued sequence $\{n(N)\}_{N \geqslant 1}$, with $n(N) \leqslant N$, we introduce a class of schemes named CJSQ($n(N)$), containing all the schemes that always assign the incoming task to one of the $n(N) + 1$ lowest ordered server pools. Note that when $n(N) = 0$, the class only contains the ordinary JSQ policy.

Just like the JSQ($d(N)$) scheme, the schemes in the class CJSQ($n(N)$) may be thought of as "sloppy" versions of the JSQ policy, in the sense that tasks are not necessarily assigned to a server pool with the minimum number of active tasks but to one of the $n(N) + 1$ lowest ordered server pools, as graphically illustrated in Figure 2a. Below we often will not differentiate among the various schemes in the class CJSQ($n(N)$), and prove

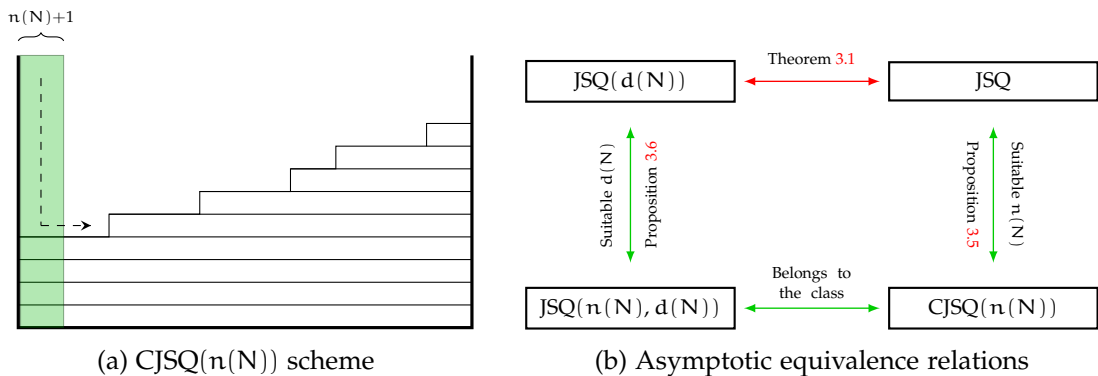(a) CJSQ(n(N)) scheme   (b) Asymptotic equivalence relations

Figure 2: (Left) The class CJSQ(n(N)) is depicted in a high-level view of the system, where as in Figure 1 the server pools are arranged in nondecreasing order of the number of active tasks, and the arrival must be assigned through the left tunnel. (Right) The equivalence structure is depicted for various intermediate load balancing schemes to facilitate the comparison between the JSQ(d(N)) scheme and the ordinary JSQ policy.

a common property possessed by all these schemes. Hence, with minor abuse of notation, we will often denote a typical assignment scheme in this class by CJSQ(n(N)). Note that the JSQ(d(N)) scheme is guaranteed to identify the lowest ordered server pool, but only among a randomly sampled subset of d(N) server pools. In contrast, a scheme in the class in CJSQ(n(N)) only guarantees that one of the $n(N) + 1$ lowest ordered server pools is selected, but across the entire system of N server pools. We will show that for sufficiently small n(N), any scheme from the class CJSQ(n(N)) is still 'close' to the ordinary JSQ policy in terms of g(N)-alikeness as stated in the next proposition.

Proposition 3.5. *For any function* $g : \mathbb{N} \to \mathbb{R}_+$ *diverging to infinity, if* $n(N)/g(N) \to 0$ *as* $N \to \infty$, *then the JSQ policy and the CJSQ(n(N)) schemes are* g(N)-*alike*.

In order to prove this proposition, we introduce in Section 4.1 a novel stochastic coupling called the T-coupling, to construct a common probability space, and establish the property of g(N)-alikeness.

Next we compare the CJSQ(n(N)) schemes with the JSQ(d(N)) scheme. The comparison follows a somewhat similar line of argument as in [16, Section 4], and involves a JSQ(n(N), d(N)) scheme which is an intermediate blend between the CJSQ(n(N)) schemes and the JSQ(d(N)) scheme. Specifically, the JSQ(n(N), d(N)) scheme selects a candidate server pool in the exact same way as the JSQ(d(N)) scheme. However, it only assigns the task to that server pool if it belongs to the $n(N) + 1$ lowest ordered ones, and to a randomly selected server pool among these otherwise. By construction, the JSQ(n(N), d(N)) scheme belongs to the class CJSQ(n(N)).

We now consider two T-coupled systems with a JSQ(d(N)) and a JSQ(n(N), d(N)) scheme. Assume that at some specific arrival epoch, the incoming task is assigned to the kth ordered server pool in the system under the JSQ(d(N)) scheme. If $k \in \{1, 2, \ldots, n(N) + 1\}$, then the scheme JSQ(n(N), d(N)) also assigns the arriving task to the kth ordered

server pool. Otherwise it dispatches the arriving task uniformly at random among the first $n(N) + 1$ ordered server pools.

We will establish a sufficient criterion on $d(N)$ in order for the JSQ($d(N)$) scheme and JSQ($n(N), d(N)$) scheme to be close in terms of $g(N)$-alikeness, as stated in the next proposition.

Proposition 3.6. *Assume,* $n(N)/g(N) \to 0$ *as* $N \to \infty$ *for some function* $g : \mathbb{N} \to \mathbb{R}_+$ *diverging to infinity. The JSQ($n(N), d(N)$) scheme and the JSQ($d(N)$) scheme are $g(N)$-alike if the following condition holds:*

$$\frac{n(N)}{N} d(N) - \log \frac{N}{g(N)} \to \infty, \quad as \quad N \to \infty. \tag{3.3}$$

Finally, Proposition 3.6 in conjunction with Proposition 3.5 yields Theorem 3.1. The overall proof strategy as described above, is schematically represented in Figure 2b.

Remark 3.7. Note that, sampling *without* replacement polls more server pools than *with* replacement, and hence the minimum number of active tasks among the selected server pools is stochastically smaller in the case without replacement. As a result, for sufficient conditions as in Theorem 3.1 it is enough to consider sampling with replacement.

4 Universality Property

In this section we formalize the proof outlined in the previous section. In Subsection 4.1 we first introduce the T-coupling between any two task assignment schemes. This coupling is used to derive stochastic inequalities in Subsection 4.2, stated as Proposition 4.1 and Lemma 4.2, which in turn, are used to prove Propositions 3.5, 3.6 and Theorem 3.1 in Subsection 4.3.

4.1 Stochastic coupling

Throughout this subsection we fix N, and suppress the superscript N in the notation. Let $Q_i^{\Pi_1}(t)$ and $Q_i^{\Pi_2}(t)$ denote the number of server pools with at least i active tasks, at time t, in two systems following schemes Π_1 and Π_2, respectively. With a slight abuse of terminology, we occasionally use Π_1 and Π_2 to refer to systems following schemes Π_1 and Π_2, respectively. To couple the two systems, we synchronize the arrival epochs and maintain a single exponential departure clock with instantaneous rate at time t given by $M(t) := \max \left\{ \sum_{i=1}^{B} Q_i^{\Pi_1}(t), \sum_{i=1}^{B} Q_i^{\Pi_2}(t) \right\}$. We couple the arrivals and departures in the various server pools as follows:

(1) *Arrival:* At each arrival epoch, assign the incoming task in each system to one of the server pools according to the respective schemes.

(2) *Departure:* Define

$$H(t) := \sum_{i=1}^{B} \min \left\{ Q_i^{\Pi_1}(t), Q_i^{\Pi_2}(t) \right\}$$
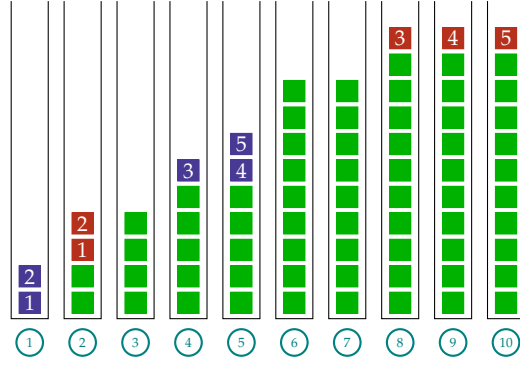
Figure 3: Superposition of the occupancy states at sgreenome particular time instant, of schemes Π_1 and Π_2 when the server pools in both systems are arranged in nondecreasing order of the number of active tasks. The Π_1 system is the union of the green and blue tasks, and the Π_2 system is the union of the green and red tasks.

and

$$p(t) := \begin{cases} \dfrac{H(t)}{M(t)}, & \text{if} \quad M(t) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

At each departure epoch t_k (say), draw a uniform$[0,1]$ random variable $U(t_k)$. The departures occur in a coupled way based upon the value of $U(t_k)$. In either of the systems, assign a task index (i,j), if that task is at the jth position of the ith ordered server pool. Let $\mathcal{A}_1(t)$ and $\mathcal{A}_2(t)$ denote the set of all task-indices present at time t in systems Π_1 and Π_2, respectively. Color the indices (or tasks) in $\mathcal{A}_1 \cap \mathcal{A}_2$, $\mathcal{A}_1 \setminus \mathcal{A}_2$ and $\mathcal{A}_2 \setminus \mathcal{A}_1$, green, blue and red, respectively, and note that $|\mathcal{A}_1 \cap \mathcal{A}_2| = H(t)$. Define a total order on the set of indices as follows: $(i_1, j_1) < (i_2, j_2)$ if $i_1 < i_2$, or $i_1 = i_2$ and $j_1 < j_2$. Now, if $U(t_k) \leqslant p(t_k-)$, then select one green index uniformly at random and remove the corresponding tasks from both systems. Otherwise, if $U(t_k) > p(t_k-)$, then choose one integer m, uniformly at random from all the integers between 1 and $M(t) - H(t) = M(t)(1 - p(t))$, and remove the tasks corresponding to the mth smallest (according to the order defined above) red and blue indices in the corresponding systems. If the number of red (or blue) tasks is less than m, then do nothing.

The above coupling has been schematically represented in Figure 3, and will henceforth be referred to as T-coupling, where T stands for 'task-based'. Now we need to show that, under the T-coupling, the two systems, considered independently, evolve according to their own statistical laws. This can be seen in several steps. Indeed, the T-coupling basically uniformizes the departure rate by the maximum number of tasks present in either of the two systems. Then informally speaking, the green regions signifies the common portion of tasks, and the red and blue region represent the separate contributions. Now observe that

(i) The total departure rate from Π_i is

$$M(t) \left[p(t) + (1 - p(t)) \frac{|\mathcal{A}_i \setminus \mathcal{A}_{3-i}|}{M(t) - H(t)} \right] = |\mathcal{A}_1 \cap \mathcal{A}_2| + |\mathcal{A}_i \setminus \mathcal{A}_{3-i}| = |\mathcal{A}_i|, \quad i = 1, 2.$$

15

(ii) Assuming without loss of generality $|\mathcal{A}_1| \geqslant |\mathcal{A}_2|$, each task in Π_1 is equally likely to depart.

(iii) Each task in Π_2 within $\mathcal{A}_1 \cap \mathcal{A}_2$ and each task within $\mathcal{A}_2 \setminus \mathcal{A}_1$ is equally likely to depart, and the probabilities of departures are proportional to $|\mathcal{A}_1 \cap \mathcal{A}_2|$ and $|\mathcal{A}_2 \setminus \mathcal{A}_1|$, respectively.

4.2 Stochastic inequalities

Now, as in [16] we define a notion of comparison between two T-coupled systems. Two T-coupled systems are said to *differ in decision* at some arrival epoch, if the index of the ordered server pool joined by the arriving task at that epoch, differs in the two systems. Denote by $\Delta_{\Pi_1, \Pi_2}(t)$, the cumulative number of times that the two systems Π_1 and Π_2 differ in decision up to time t.

Proposition 4.1. *For two T-coupled systems under any two schemes Π_1 and Π_2 the following inequality is preserved*

$$\sum_{i=1}^{B} \left| Q_i^{\Pi_1}(t) - Q_i^{\Pi_2}(t) \right| \leqslant 2\Delta_{\Pi_1,\Pi_2}(t) \qquad \forall\, t \geqslant 0, \tag{4.1}$$

provided the two systems start from the same occupancy state at time $t = 0$.

The proof follows a somewhat similar line of argument as in [15, 16], but is provided below since the coupling is different here. For any scheme Π, define $I_{\Pi}(c) := \max\{i : Q_i^{\Pi} \geqslant N - c + 1\}$, $c = 1, \ldots, N$.

Proof of Proposition 4.1. We use forward induction on event times, i.e., time epochs when either an arrival or a departure takes place. Assume the inequality in (4.1) holds at time epoch t_0. We denote by $\tilde{Q}^{\Pi}$ the updated occupancy state after the next event at time epoch t_1, and distinguish between two cases depending on whether t_1 is an arrival epoch or a departure epoch.

If t_1 is an arrival epoch and if the systems differ in decision, then observe that the left side of (4.1) can increase at most by two. In this case, the right side also increases by two, and the ordering is preserved. Therefore, it is enough to prove that the right side of (4.1) remains unchanged if the two systems do not differ in decision. In that case, assume that both Π_1 and Π_2 assign the arriving task to the kth ordered server pool. Then

$$\tilde{Q}_i^{\Pi} = \begin{cases} Q_i^{\Pi} + 1, & \text{for } i = I_{\Pi}(k) + 1, \\ Q_j^{\Pi}, & \text{otherwise,} \end{cases} \tag{4.2}$$

if $I_{\Pi}(k) < B$; otherwise all the Q_i-values remain unchanged. If $I_{\Pi_1}(k) = I_{\Pi_2}(k)$, then the left side of (4.1) clearly remains unchanged. Now, without loss of generality, assume $I_{\Pi_1}(k) < I_{\Pi_2}(k)$. Therefore,

$$Q_{I_{\Pi_1}(k)+1}^{\Pi_1}(t_0) < Q_{I_{\Pi_1}(k)+1}^{\Pi_2}(t_0) \quad \text{and} \quad Q_{I_{\Pi_2}(k)+1}^{\Pi_1}(t_0) < Q_{I_{\Pi_2}(k)+1}^{\Pi_2}(t_0).$$

After an arrival, the $(I_{\Pi_1}(k) + 1)^{\text{th}}$ term in the left side of (4.1) decreases by one, and the $(I_{\Pi_2}(k) + 1)^{\text{th}}$ term increases by one. Thus the inequality is preserved.

16

If t_1 is a departure epoch, then first consider the case when the departure occurs from the green region. In that case, without loss of generality, assume that a potential departure occurs from the kth ordered server pool, for some $k \in \{1, 2, \ldots, N\}$. Also note that a departure in either of the two systems can change at most one of the Q_i-values. Thus

$$\tilde{Q}_i^{\Pi} = \begin{cases} Q_i^{\Pi} - 1, & \text{for } i = I_{\Pi}(k), \\ Q_j^{\Pi}, & \text{otherwise,} \end{cases} \tag{4.3}$$

if $I_{\Pi}(k) \geqslant 1$; otherwise all the Q_i-values remain unchanged.

If at time epoch t_0, $I_{\Pi_1}(k) = I_{\Pi_2}(k) = I$, then both $Q_{I^{\Pi_1}}$ and $Q_{I^{\Pi_2}}$ decrease by one, and hence the left side of (4.1) does not change.

Otherwise, without loss of generality assume $I_{\Pi_1}(k) < I_{\Pi_2}(k)$. Then observe that

$$Q_{I_{\Pi_1}(k)}^{\Pi_1}(t_0) \leqslant Q_{I_{\Pi_1}(k)}^{\Pi_2}(t_0) \quad \text{and} \quad Q_{I_{\Pi_2}(k)}^{\Pi_1}(t_0) < Q_{I_{\Pi_2}(k)}^{\Pi_2}(t_0).$$

Furthermore, after the departure, $Q_{I_{\Pi_1}(k)}^{\Pi_1}$ decreases by one, therefore $|Q_{I_{\Pi_1}(k)}^{\Pi_1} - Q_{I_{\Pi_1}(k)}^{\Pi_2}|$ increases by one, and $Q_{I_{\Pi_2}(k)}^{\Pi_2}$ decreases by one, thus $|Q_{I_{\Pi_2}(k)}^{\Pi_1} - Q_{I_{\Pi_2}(k)}^{\Pi_2}|$ decreases by one. Hence, in total, the left side of (4.1) remains the same. Now if a departure occurs from the blue and/or red region, then for some i_1 and/or i_2, $(Q_{i_1}^{\Pi_1} - Q_{i_1}^{\Pi_2})^+$ or $(Q_{i_2}^{\Pi_2} - Q_{i_2}^{\Pi_1})^+$ (or both) decreases, and the other terms remain unchanged, and hence the left side clearly decreases or remains unchanged. $\square$

In order to compare the JSQ policy with the CJSQ($n(N)$) schemes, denote by $Q_i^{\Pi_1}(t)$ and $Q_i^{\Pi_2}(t)$ the number of server pools with at least i tasks under the JSQ policy and CJSQ($n(N)$) scheme, respectively. Now, in order to prove Proposition 3.5, we will need the following lemma.

Lemma 4.2. *For any* $k \in \{1, 2, \ldots B\}$,

$$\left\{ \sum_{i=1}^{k} Q_i^{\Pi_1}(t) - kn(N) \right\}_{t \geqslant 0} \leqslant_{st} \left\{ \sum_{i=1}^{k} Q_i^{\Pi_2}(t) \right\}_{t \geqslant 0} \leqslant_{st} \left\{ \sum_{i=1}^{k} Q_i^{\Pi_1}(t) \right\}_{t \geqslant 0}, \tag{4.4}$$

provided at $t = 0$ *the two systems start from the same occupancy states.*

In the next two remarks we comment on the contrast of Lemma 4.2 and the underlying T-coupling with stochastic dominance properties for the ordinary JSQ policy in the existing literature and the S-coupling technique in reference [16], respectively

Remark 4.3. The stochastic ordering in Lemma 4.2 is to be contrasted with the weak majorization results in [23, 24, 25, 28, 31] in the context of the ordinary JSQ policy in the single-server queueing scenario, and in [10, 12, 13, 22] in the scenario of state-dependent service rates, non-decreasing with the number of active tasks. In the current infinite-server scenario, the results in [10, 12, 13, 22] imply that for any non-anticipating scheme Π taking assignment decision based on the number of active tasks only, for all $t \geqslant 0$,

$$\sum_{m=1}^{\ell} X_{(m)}^{JSQ}(t) \leqslant_{st} \sum_{m=1}^{\ell} X_{(m)}^{\Pi}(t), \text{ for } \ell = 1, 2, \ldots, N, \tag{4.5}$$

$$\{L^{JSQ}(t)\}_{t \geqslant 0} \leqslant_{st} \{L^{\Pi}(t)\}_{t \geqslant 0}, \tag{4.6}$$

where $X^{\Pi}_{(m)}(t)$ is the number of tasks in the mth ordered server pool at time t in the system following scheme Π and $L^{\Pi}(t)$ is the total number of overflow events under policy Π up to time t. Observe that $X^{\Pi}_{(m)}$ can be visualized as the mth largest (rightmost) vertical bar (or stack) in Figure 1. Thus (4.5) says that the sum of the lengths of the ℓ largest *vertical* stacks in a system following any scheme Π is stochastically larger than or equal to that following the ordinary JSQ policy for any $\ell = 1, 2, \ldots, N$. Mathematically, this ordering can be equivalently written as

$$\sum_{i=1}^{B} \min\left\{\ell, Q_i^{JSQ}(t)\right\} \leqslant_{st} \sum_{i=1}^{B} \min\left\{\ell, Q_i^{\Pi}(t)\right\}, \tag{4.7}$$

for all $\ell = 1, \ldots, N$. In contrast, in order to show asymptotic equivalence on various scales, we need to both upper and lower bound the occupancy states of the CJSQ($n(N)$) schemes in terms of the JSQ policy, and therefore need a much stronger hold on the departure process. The T-coupling provides us just that, and has several useful properties that are crucial for our proof technique. For example, Proposition 4.1 uses the fact that if two systems are T-coupled, then departures cannot increase the sum of the absolute differences of the Q_i-values, which is not true for the coupling considered in the above-mentioned literature. The left stochastic ordering in (4.4) also does not remain valid in those cases. Furthermore, observe that the right inequality in (4.4) (i.e., Q_i's) implies the stochastic inequality is *reversed* in (4.7), which is counter-intuitive in view of the optimality properties of the ordinary JSQ policy studied in the literature, as mentioned above. The fundamental distinction between the two coupling techniques is also reflected by the fact that the T-coupling does not allow for arbitrary nondecreasing state-dependent departure rate functions, unlike the couplings in [10, 12, 13, 22].

Remark 4.4. As briefly mentioned in the introduction, in the current infinite-server scenario, the departures of the ordered server pools cannot be coupled, mainly since the departure rate at the mth ordered server pool, for some $m = 1, 2, \ldots, N$, depends on its number of active tasks. It is worthwhile to mention that the coupling in this paper is stronger than that used in [16]. Observe that due to Lemma 4.2, the absolute difference of the occupancy states of the JSQ policy and any scheme from the CJSQ class at any time point can be bounded deterministically (without any terms involving the cumulative number of lost tasks). It is worth emphasizing that the universality result on some specific scale, stated in Theorem 3.1 does not depend on the behavior of the JSQ policy on that scale, whereas in [16] it does, mainly because the upper and lower bounds in [16, Corollary 3.3] involve tail sums of two different policies. Also, the bound in the current paper does not depend upon t, and hence, applies in the steady state as well. Moreover, the coupling in [16] compares the k *highest* horizontal bars, whereas the present paper compares the k *lowest* horizontal bars. As a result, the bounds on the occupancy states established in [16, Corollary 3.3] involves tail sums of the occupancy states of the ordinary JSQ policy, which necessitates proving the ℓ_1 convergence of the occupancy states of the ordinary JSQ policy. In contrast, the bound we establish in the present paper, involves only a single component (see equations (4.9) and (4.10)), and thus, the convergence with respect to product topology suffices.

Proof of Lemma 4.2. Fix any $k \geqslant 1$. We will use forward induction on the event times, i.e., time epochs when either an arrival or a departure occurs, and assume the two systems

to be T-coupled as described in Section 4.1. We suppose that the two inequalities hold at time epoch t_0, and will prove that they continue to hold at time epoch t_1.

(a) We first prove the left inequality in (4.4). We distinguish between two cases depending on whether the next event time t_1 is an arrival epoch or a departure epoch. We first consider the case of an arrival. Since at each arrival, there can be an increment of size at most one, if $\sum_{i=1}^k Q_i^{\Pi_1}(t_0) - kn(N) < \sum_{i=1}^k Q_i^{\Pi_2}(t_0)$, the inequality holds trivially at time t_1. Therefore, consider the case when $\sum_{i=1}^k Q_i^{\Pi_1}(t_0) - kn(N) = \sum_{i=1}^k Q_i^{\Pi_2}(t_0)$. Now observe that,

$$\sum_{i=1}^k Q_i^{\Pi_2}(t_0) = \sum_{i=1}^k Q_i^{\Pi_1}(t_0) - kn(N) \leqslant kN - kn(N).$$

Hence, $Q_k^{\Pi_2}(t_0) \leqslant N - n(N)$, which in turn implies that at time t_1, $\sum_{i=1}^k Q_i^{\Pi_2}$ increases by 1, and the inequality is preserved. We now assume the case of a departure. Then also if $\sum_{i=1}^k Q_i^{\Pi_1}(t_0) - kn(N) < \sum_{i=1}^k Q_i^{\Pi_2}(t_0)$, the inequality holds trivially at time t_1. Otherwise assume $\sum_{i=1}^k Q_i^{\Pi_1}(t_0) - kn(N) = \sum_{i=1}^k Q_i^{\Pi_2}(t_0)$. In this case if the departure occurs from the green region in Figure 3, then both $\sum_{i=1}^k Q_i^{\Pi_1}$ and $\sum_{i=1}^k Q_i^{\Pi_2}$ change in a similar fashion (i.e., either decrease by one or remain unchanged). Else, if the departure occurs from the red and blue regions, since $\sum_{i=1}^k Q_i^{\Pi_1} \geqslant \sum_{i=1}^k Q_i^{\Pi_2}$, by virtue of the T-coupling, if $\sum_{i=1}^k Q_i^{\Pi_2}$ decreases by one, then so does $\sum_{i=1}^k Q_i^{\Pi_1}$. To see this observe the following:

$$\sum_{i=1}^k Q_i^{\Pi_1} \geqslant \sum_{i=1}^k Q_i^{\Pi_2} \implies \sum_{i=1}^k (Q_i^{\Pi_1} - Q_i^{\Pi_2})^+ \geqslant \sum_{i=1}^k (Q_i^{\Pi_2} - Q_i^{\Pi_1})^+. \qquad (4.8)$$

Therefore, if $m \leqslant \sum_{i=1}^k (Q_i^{\Pi_2} - Q_i^{\Pi_1})^+$, then $m \leqslant \sum_{i=1}^k (Q_i^{\Pi_1} - Q_i^{\Pi_2})^+$. Hence the inequality will be preserved.

(b) We now prove the right inequality in (4.4) and again distinguish between two cases. If t_1 is an arrival epoch, we assume for a similar reason as above, $\sum_{i=1}^k Q_i^{\Pi_2}(t_0) = \sum_{i=1}^k Q_i^{\Pi_1}(t_0)$. In this case when a task arrives, if it gets admitted under the CJSQ(n(N)) scheme and increases $\sum_{i=1}^k Q_i^{\Pi_2}$, then clearly $\sum_{i=1}^k (N - Q_i^{\Pi_1}(t)) > 0$, and hence the incoming task will increase $\sum_{i=1}^k Q_i^{\Pi_1}$, as well, and the inequality will be preserved. If t_1 is a departure epoch with $\sum_{i=1}^k Q_i^{\Pi_2}(t_0) = \sum_{i=1}^k Q_i^{\Pi_1}(t_0)$, then by virtue of the T-coupling again, if $\sum_{i=1}^k Q_i^{\Pi_1}$ decreases by one, then by the argument in (a) above, so does $\sum_{i=1}^k Q_i^{\Pi_2}$, thus preserving the inequality. $\qquad \square$

4.3 Asymptotic equivalence

Proof of Proposition 3.5. Using Lemma 4.2, there exists a common probability space such that for any $k \geqslant 1$ we can write

$$
\begin{aligned}
Q_k^{\Pi_2}(t) &= \sum_{i=1}^k Q_i^{\Pi_2}(t) - \sum_{i=1}^{k-1} Q_i^{\Pi_2}(t) \\
&\leqslant \sum_{i=1}^k Q_i^{\Pi_1}(t) - \sum_{i=1}^{k-1} Q_i^{\Pi_1}(t) + kn(N) \\
&= Q_k^{\Pi_1}(t) + kn(N).
\end{aligned}
\qquad (4.9)
$$

19

Similarly, we can write

$$
\begin{aligned}
Q_k^{\Pi_2}(t) &= \sum_{i=1}^{k} Q_i^{\Pi_2}(t) - \sum_{i=1}^{k-1} Q_i^{\Pi_2}(t) \\
&\geqslant \sum_{i=1}^{k} Q_i^{\Pi_1}(t) - kn(N) - \sum_{i=1}^{k-1} Q_i^{\Pi_1}(t) \\
&= Q_k^{\Pi_1}(t) - kn(N).
\end{aligned}
\tag{4.10}
$$

Therefore, for all $k \geqslant 1$, we have, $\sup_t |Q_k^{\Pi_2}(t) - Q_k^{\Pi_1}(t)| \leqslant kn(N)$. Since $n(N)/g(N) \to \infty$ as $N \to \infty$, the proof is complete. $\qquad\square$

Proof of Proposition 3.6. For any $T \geqslant 0$, let $A^N(T)$ and $\Delta^N(T)$ be the total number of arrivals to the system and the cumulative number of times that the JSQ($d(N)$) scheme and the JSQ($n(N), d(N)$) scheme differ in decision up to time T. Using Proposition 4.1 it suffices to show that for any $T \geqslant 0$, $\Delta^N(T)/g(N) \xrightarrow{\mathbb{P}} 0$ as $N \to \infty$. Observe that at any arrival epoch, the systems under the JSQ($d(N)$) and JSQ($n(N), d(N)$) schemes will differ in decision only if none of the $n(N) + 1$ lowest ordered server pools get selected by the JSQ($d(N)$) scheme.

Now at the time of an arrival, the probability that the JSQ($d(N)$) scheme does not select one of the $n(N) + 1$ lowest ordered server pools, is given by

$$
p(N) = \left(1 - \frac{n(N) + 1}{N}\right)^{d(N)}.
$$

Since at each arrival epoch $d(N)$ server pools are selected independently, given $A^N(T)$, $\Delta^N(T) \sim \text{Bin}(A_N(T), p(N))$.

Note that, for $T \geqslant 0$, Markov's inequality yields

$$
\mathbb{P}\left(\Delta^N(T) \geqslant g(N) \,\middle|\, A_N(T)\right) \leqslant \frac{\mathbb{E}\left(\Delta^N(T)\right)}{g(N)} = \frac{A_N(T)}{g(N)} \left(1 - \frac{n(N) + 1}{N}\right)^{d(N)}.
$$

Since $\left\{A^N(T)/N\right\}_{N \geqslant 1}$ is a tight sequence of random variables, in order to ensure that $\Delta^N(T)/g(N)$ converges to zero in probability, it is enough to have

$$
\begin{aligned}
&\frac{N}{g(N)} \left(1 - \frac{n(N) + 1}{N}\right)^{d(N)} \to 0 \\
&\Longleftarrow \exp\left(\log\left(\frac{N}{g(N)}\right) - d(N)\frac{n(N)}{N}\right) \to 0 \\
&\Longleftrightarrow d(N)\frac{n(N)}{N} - \log\left(\frac{N}{g(N)}\right) \to \infty.
\end{aligned}
\tag{4.11}
$$

$\qquad\square$

We now use Propositions 3.5 and 3.6 to prove Theorem 3.1.

Proof of Theorem 3.1. Fix any $d(N)$ satisfying either (3.1) or (3.2). From Propositions 3.5 and 3.6 observe that it is enough to show that there exists an $n(N)$ with $n(N) \to \infty$ and $n(N)/g(N) \to 0$, as $N \to \infty$, such that

$$\frac{n(N)}{N} d(N) - \log\left(\frac{N}{g(N)}\right) \to \infty.$$

(i) If $g(N) = O(N)$, then observe that $\log(N/g(N))$ is $O(1)$. Since $d(N) \to \infty$, choosing $n(N) = N/\log(d(N))$ satisfies the above criteria, and hence part (i) of the theorem is proved.

(ii) Next we obtain a choice of $n(N)$ if $g(N) = o(N)$. Note that, if

$$h(N) := \frac{d(N)\frac{g(N)}{N}}{\log\left(\frac{N}{g(N)}\right)} \to \infty, \quad \text{as} \quad N \to \infty,$$

then choosing $n(N) = g(N)/\log(h(N))$, it can be seen that as $N \to \infty$, $n(N)/g(N) \to 0$, and

$$\frac{d(N)\frac{n(N)}{N}}{\log\left(\frac{N}{g(N)}\right)} = \frac{h(N)}{\log(h(N))} \to \infty$$

$$\implies \frac{n(N)}{N} d(N) - \log\left(\frac{N}{g(N)}\right) \to \infty. \tag{4.12}$$

$\square$

5   Fluid Limit of JSQ

In this section we establish the fluid limit for the ordinary JSQ policy. In the proof we will leverage the time scale separation technique developed in [8], suitably extended to an infinite-dimensional space. Specifically, note that the rate at which incoming tasks join a server pool with i active tasks is determined only by the process $\mathbf{Z}^N(\cdot) = (Z_1^N(\cdot), \ldots, Z_B^N(\cdot))$, where $Z_i^N(t) = N - Q_i^N(t)$, $i = 1, \ldots, B$, represents the number of server pools with fewer than i tasks at time t. Furthermore, in any time interval $[t, t+\varepsilon]$ of length $\varepsilon > 0$, the $\mathbf{Z}^N(\cdot)$ process experiences $O(\varepsilon N)$ events (arrivals and departures), while the $\mathbf{q}^N(\cdot)$ process can change by only $O(\varepsilon)$ amount. Therefore, the $\mathbf{Z}^N(\cdot)$ process evolves on a much faster time scale than the $\mathbf{q}^N(\cdot)$ process. As a result, in the limit as $N \to \infty$, at each time point t, the $\mathbf{Z}^N(\cdot)$ process achieves stationarity depending on the instantaneous value of the $\mathbf{q}^N(\cdot)$ process, i.e., a separation of time scales takes place.

In order to illuminate the generic nature of the proof construct, we will allow for a more general task assignment probability and departure dynamics than described in Section 2. Denote by $\bar{\mathbb{Z}}_+$ the one-point compactification of the set of nonnegative integers $\mathbb{Z}_+$, i.e., $\bar{\mathbb{Z}}_+ = \bar{\mathbb{Z}}_+ \cup \{\infty\}$. Equip $\bar{\mathbb{Z}}_+$ with the order topology. Denote $G = \bar{\mathbb{Z}}_+^B$ equipped with product-topology, and with the Borel σ-algebra, $\mathcal{G}$. Let us consider the G-valued process $\mathbf{Z}^N(s) := (Z_i^N(s))_{i \geqslant 1}$ as introduced above. Let $\{\mathcal{R}_i\}_{1 \leqslant i \leqslant B}$ be a partition of G such that $\mathcal{R}_i \in \mathcal{G}$. We assume that a task arriving at (say) t_k is assigned to some server pool with i active tasks is given by $p_{i-1}^N(\mathbf{Q}^N(t_k-)) = \mathbb{1}_{[\mathbf{Z}^N(t_k-) \in \mathcal{R}_i]} f_i(\mathbf{q}^N(t_k-)),$

where $\mathbf{f} = (f_1, \ldots, f_B) : [0,1]^B \to [0,1]^B$ is Lipschitz continuous, i.e., there exists $C_{\mathbf{f}}$, such that for any $\mathbf{q}_1, \mathbf{q}_2 \in S$,

$$\|\mathbf{f}(\mathbf{q}_1) - \mathbf{f}(\mathbf{q}_2)\| \leqslant C_{\mathbf{f}} \|\mathbf{q}_1 - \mathbf{q}_2\|.$$

The partition corresponding to the ordinary JSQ policy can be written as

$$\mathcal{R}_i := \big\{ (z_1, z_2, \ldots, z_B) : z_1 = \ldots = z_{i-1} = 0 < z_i \leqslant z_{i+1} \leqslant \ldots \leqslant z_B \big\}, \qquad (5.1)$$

with the convention that Q_B^N is always taken to be zero, if $B < \infty$, and $f_i \equiv 1$ for all $i = 1, 2, \ldots, B$. The fluid-limit results up to Proposition 5.5 (the relative compactness of the fluid-scaled process) hold true for these general assignment probabilities. It is only when proving Theorem 5.1, that we need to assume the specific $\{\mathcal{R}_i\}_{1 \leqslant i \leqslant B}$ in (5.1). For the departure dynamics, when the system occupancy state is $\mathbf{Q}^N = (Q_1^N, Q_2^N, \ldots, Q_B^N)$, define the total rate at which departures occur from a server pool with i active tasks by $\mu_i^N(\mathbf{Q})$, where $\boldsymbol{\mu}^N(\mathbf{Q}) = (\mu_1^N(\mathbf{Q}), \ldots, \mu_B^N(\mathbf{Q}))$ will be referred to as the departure rate function. The departure dynamics described in Section 2 correspond to $\mu_i^N(\mathbf{Q}) = i(Q_i - Q_{i+1})$ and will be referred to as the infinite-server scenario, since all active tasks are executed concurrently. The single-server scenario, where tasks are executed sequentially, corresponds to the case $\mu_i^N(\mathbf{Q}) = Q_i - Q_{i+1}$.

Assumption 1 (Condition on departure rate function). *The departure rate function $\boldsymbol{\mu}^N : \tilde{S} \to [0, \infty)^B$ satisfies the following conditions*

(a) *There exists a function $\boldsymbol{\mu} : S \to [0, \infty)^B$, such that*

$$\lim_{N \to \infty} \sup_{\mathbf{q} \in S^N} \left\| \frac{1}{N} \boldsymbol{\mu}^N(\lfloor N\mathbf{q} \rfloor) - \boldsymbol{\mu}(\mathbf{q}) \right\| = 0.$$

(b) *The function $\boldsymbol{\mu}$ is Lipschitz continuous in S, i.e., there exists a constant $C_{\boldsymbol{\mu}}$, such that for any $\mathbf{q}_1, \mathbf{q}_2 \in S$,*

$$\|\boldsymbol{\mu}(\mathbf{q}_1) - \boldsymbol{\mu}(\mathbf{q}_2)\| \leqslant C_{\boldsymbol{\mu}} \|\mathbf{q}_1 - \mathbf{q}_2\|.$$

(c) *Also, $\boldsymbol{\mu}^N$ satisfies linear growth constraints in each coordinate, i.e., for all $i \geqslant 1$, there exists $C_i > 0$, such that for all $\mathbf{q} \in S$,*

$$\mu_i^N(\lfloor N\mathbf{q} \rfloor) \leqslant NC_i(1 + \|\mathbf{q}\|).$$

We will often omit $\lfloor \cdot \rfloor$ in the argument of $\boldsymbol{\mu}^N$ for notational convenience.

Under these assumptions on the departure rate function, we prove the following fluid-limit result for the ordinary JSQ policy. Recall the definition of $m(\mathbf{q})$ in Subsection 2.2, and define

$$p_i(\mathbf{q}) = \begin{cases} \min\{\mu_{m(\mathbf{q})}(\mathbf{q})/\lambda, 1\} & \text{for} \quad i = m(\mathbf{q}) - 1, \\ 1 - p_{m(\mathbf{q})-1}(\mathbf{q}) & \text{for} \quad i = m(\mathbf{q}), \\ 0 & \text{otherwise.} \end{cases} \qquad (5.2)$$

Note that $p_i(\cdot)$ in (5.2) is consistent with the one defined in Subsection 2.2 for the proper choice of the departure rate function $\mu_i(\mathbf{q}) = i(q_i - q_{i+1})$.

Theorem 5.1 (Fluid limit of JSQ). *Assume* $\mathbf{q}^N(0) \xrightarrow{\mathbb{P}} \mathbf{q}^\infty \in S$ *and* $\lambda(N)/N \to \lambda > 0$ *as* $N \to \infty$. *Further assume that the departure rate function* μ^N *satisfies Assumption 1. Then the sequence of processes* $\{\mathbf{q}^N(t)\}_{t \geqslant 0}$ *for the ordinary JSQ policy has a continuous weak limit that satisfies the system of integral equations*

$$q_i(t) = q_i(0) + \lambda \int_0^t p_{i-1}(\mathbf{q}(s)) ds - \int_0^t \mu_i(\mathbf{q}(s)) ds, \quad i = 1, 2, \ldots, B, \qquad (5.3)$$

where $\mathbf{q}(0) = \mathbf{q}^\infty$ *and the coefficients* $p_i(\cdot)$ *are defined in (5.2), and may be interpreted as the fractions of incoming tasks assigned to server pools with exactly* i *active tasks.*

We will now verify that the departure rate functions corresponding to the infinite-server and single-server scenarios satisfy the conditions in Assumption 1.

Proposition 5.2. *The following departure rate functions denoted by* $\mu = (\mu_1, \mu_2, \ldots, \mu_B)$, *satisfy the conditions in Assumption 1. For* $\mathbf{Q} \in \tilde{S}$, *and* $\mathbf{q} \in S$,

(i) $\mu_i^N(\mathbf{Q}) = Q_i - Q_{i+1}$, *and* $\mu_i(\mathbf{q}) = q_i - q_{i+1}$, $i \geqslant 1$.

(ii) $\mu_i^N(\mathbf{Q}) = i(Q_i - Q_{i+1})$, *and* $\mu_i(\mathbf{q}) = i(q_i - q_{i+1})$, $i \geqslant 1$.

Proof. Observe that if $B < \infty$, then since componentwise μ_i satisfies all the conditions for all $i \geqslant 1$, μ satisfies the conditions in the product space as well. Therefore, let us consider the case when $B = \infty$. In this case observe that, for both (i) and (ii) condition (a) is immediate, since $\mu^N(\lfloor N\mathbf{q} \rfloor)/N = \mu(\mathbf{q})$ for all $\mathbf{q} \in S^N$. Also, the linear growth rate constraint in condition (c) is satisfied in both cases by taking $C_i = 1$ in (i) and $C_i = i$ in (ii).

Now we will show that in both cases μ is Lipschitz continuous in S.

(i) For $\mu_i(\mathbf{q}) = q_i - q_{i+1}$, $i \geqslant 1$, and $\mathbf{q}_1, \mathbf{q}_2 \in S$,

$$\|\mu(\mathbf{q})\| = \sum_{i \geqslant 1} \frac{|q_i - q_{i+1}|}{2^i} \leqslant \sum_{i \geqslant 1} \frac{q_i}{2^i} + \sum_{i \geqslant 1} \frac{q_{i+1}}{2^i} \leqslant 2 \|\mathbf{q}\|.$$

(ii) Now assume $\mu_i(\mathbf{q}) = i(q_i - q_{i+1})$, $i \geqslant 1$. Since μ is a linear operator on the Banach space (complete normed linear space) $\mathbb{R}^B$, to prove Lipschitz continuity of μ, it is enough to show that μ is continuous at zero. Specifically, we will show that for any sequence $\{\mathbf{q}^n\}_{n \geqslant 1}$, in $\mathbb{R}^B$, $\|\mathbf{q}^n\| \to 0$ implies $\|\mu(\mathbf{q}^n)\| \to 0$. This would imply that there exists $\varepsilon > 0$, such that whenever $\|\mathbf{q}^n\| \leqslant \varepsilon$ with $\mathbf{q}^n \in \mathbb{R}^B$, we have $\|\mu(\mathbf{q}^n)\| < 1$. Then due to linearity of μ, for any $\mathbf{q} \in \mathbb{R}^B$,

$$\|\mu(\mathbf{q})\| = \left\| \frac{\|\mathbf{q}\|}{\varepsilon} \mu \left(\varepsilon \frac{\mathbf{q}}{\|\mathbf{q}\|} \right) \right\|$$

$$\leqslant \frac{\|\mathbf{q}\|}{\varepsilon} \left\| \mu \left(\varepsilon \frac{\mathbf{q}}{\|\mathbf{q}\|} \right) \right\|$$

$$\leqslant \frac{1}{\varepsilon} \|\mathbf{q}\|.$$

To show that μ is continuous at $\mathbf{0} \in \mathbb{R}^B$, fix any $\varepsilon > 0$. Also, fix an $M > 0$, depending upon ε, such that $\sum_{i > M} 1/2^i < \varepsilon/2$. Now, choose $\delta < \varepsilon/(4M)$. Then, for any $\mathbf{q}$ such that

$\|\mathbf{q}\| < \delta$, we have

$$\begin{aligned}
\|\boldsymbol{\mu}(\mathbf{q})\| &= \sum_{i=1}^{\infty} \frac{i|q_i - q_{i+1}|}{2^i} \\
&= \sum_{i=1}^{M} \frac{i|q_i - q_{i+1}|}{2^i} + \frac{\varepsilon}{2} \\
&\leqslant M \sum_{i=1}^{M} \frac{|q_i - q_{i+1}|}{2^i} + \frac{\varepsilon}{2} \\
&\leqslant 2M \|\mathbf{q}\| + \frac{\varepsilon}{2} \leqslant \varepsilon.
\end{aligned}$$

Hence, $\boldsymbol{\mu}$ is Lipschitz continuous on $\mathbb{R}^{\infty}$. $\qquad\square$

5.1 Martingale representation

In this subsection we construct the martingale representation of the occupancy state process $\mathbf{Q}^N(\cdot)$. The component $Q_i^N(t)$, satisfies the identity relation

$$Q_i^N(t) = Q_i^N(0) + A_i^N(t) - D_i^N(t), \quad \text{for} \quad i = 1, \dots, B, \tag{5.4}$$

where

$A_i^N(t) = $ number of arrivals during $[0, t]$ to some server pool with $i - 1$ active tasks,

$D_i^N(t) = $ number of departures during $[0, t]$ from some server pool with i active tasks.

We can express $A_i^N(t)$ and $D_i^N(t)$ as

$$A_i^N(t) = \mathcal{N}_{A,i} \left(\lambda(N) \int_0^t p_{i-1}^N(\mathbf{Q}^N(s)) ds \right),$$

$$D_i^N(t) = \mathcal{N}_{D,i} \left(\int_0^t \mu_i^N(\mathbf{Q}^N(s)) ds \right),$$

where $\mathcal{N}_{A,i}$ and $\mathcal{N}_{D,i}$ are mutually independent unit-rate Poisson processes, $i = 1, 2, \dots, B$. Define the following sigma fields.

$$\mathcal{A}_i^N(t) := \sigma \left(A_i^N(s) : 0 \leqslant s \leqslant t \right),$$
$$\mathcal{D}_i^N(t) := \sigma \left(D_i^N(s) : 0 \leqslant s \leqslant t \right), \text{ for } i \geqslant 1,$$

and the filtration $\mathbf{F}^N \equiv \left\{ \mathcal{F}_t^N : t \geqslant 0 \right\}$ with

$$\mathcal{F}_t^N := \bigvee_{i=1}^{\infty} \left[\mathcal{A}_i^N(t) \vee \mathcal{D}_i^N(t) \right] \tag{5.5}$$

augmented by all the null sets. Now we have the following martingale decomposition from the classical result in [2, Proposition 3].

24

Proposition 5.3. *The following are* $\mathbf{F}^N$*-martingales, for* $i \geqslant 1$:

$$M_{A,i}^N(t) := \mathcal{N}_{A,i}\left(\lambda(N)\int_0^t p_{i-1}^N(\mathbf{Q}^N(s))ds\right) - \lambda(N)\int_0^t p_{i-1}^N(\mathbf{Q}^N(s))ds,$$

$$M_{D,i}^N(t) := \mathcal{N}_{D,i}\left(\int_0^t \mu_i^N(\mathbf{Q}^N(s))ds\right) - \int_0^t \mu_i^N(\mathbf{Q}^N(s))ds,$$

(5.6)

with respective compensator and predictable quadratic variation processes given by

$$\langle M_{A,i}^N\rangle(t) := \lambda(N)\int_0^t p_{i-1}^N(\mathbf{Q}^N(s-))ds,$$

$$\langle M_{D,i}^N\rangle(t) := \int_0^t \mu_i^N(\mathbf{Q}^N(s))ds.$$

Therefore, finally we have the following martingale representation of the Nth process:

$$Q_i^N(t) = Q_i^N(0) + \lambda(N)\int_0^t p_{i-1}^N(\mathbf{Q}^N(s))ds$$

$$- \int_0^t \mu_i^N(\mathbf{Q}^N(s))ds + (M_{A,i}^N(t) - M_{D,i}^N(t)), \quad t \geqslant 0, \quad i = 1,\dots,B.$$

(5.7)

In the proposition below, we prove that the martingale part vanishes when scaled by N. Since convergence in probability in each component implies convergence in probability with respect to the product topology, it is enough to show convergence in each component.

Proposition 5.4. *For all* $i \geqslant 1$,

$$\left\{\frac{1}{N}(M_{A,i}^N(t) - M_{D,i}^N(t))\right\}_{t\geqslant 0} \xrightarrow{\mathcal{L}} \{m(t)\}_{t\geqslant 0} \equiv 0.$$

Proof. Fix any $T \geqslant 0$, and $i \geqslant 1$. From Doob's inequality [11, Theorem 1.9.1.3], we have

$$\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{N}M_{A,i}^N(t) \geqslant \epsilon\right) = \mathbb{P}\left(\sup_{t\in[0,T]} M_{A,i}^N(t) \geqslant N\epsilon\right)$$

$$\leqslant \frac{1}{N^2\epsilon^2}\mathbb{E}\left(\langle M_{A,i}^N\rangle(T)\right)$$

$$\leqslant \frac{1}{N\epsilon^2}\int_0^T p_{i-1}(\mathbf{Q}^N(s-))\lambda N ds$$

$$\leqslant \frac{\lambda T}{N\epsilon^2} \to 0, \text{ as } N \to \infty.$$

Similarly, for $M_{D,i}^N$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\frac{1}{N}M_{D,i}^N(t) \geqslant \epsilon\right) = \mathbb{P}\left(\sup_{t\in[0,T]} M_{D,i}^N(t) \geqslant N\epsilon\right)$$

$$\leqslant \frac{1}{N^2\epsilon^2}\mathbb{E}\left(\langle M_{D,i}^N\rangle(T)\right)$$

$$\leqslant \frac{1}{N^2\epsilon^2}\int_0^T \mu_i^N(\mathbf{Q}^N(s))ds$$

$$\leqslant \frac{2L'}{N\epsilon^2} \to 0, \text{ as } N \to \infty,$$

where the last inequality follows from the linear growth constraint in Assumption 1 (c). Therefore we have uniform convergence over compact sets, and hence with respect to the Skorohod-J_1 topology. $\qquad\square$

5.2 Relative compactness and uniqueness

Now we will first prove the relative compactness of the sequence of fluid-scaled processes. Recall that we denote all the fluid-scaled quantities by their respective small letters, e.g. $\mathbf{q}^N(t) := \mathbf{Q}^N(t)/N$, componentwise, i.e., $q_i^N(t) := Q_i^N(t)/N$ for $i \geqslant 1$. Therefore the martingale representation in (5.7), can be written as

$$
\begin{aligned}
q_i^N(t) = q_i^N(0) &+ \frac{\lambda(N)}{N} \int_0^t p_{i-1}^N(\mathbf{Q}^N(s))ds \\
&- \int_0^t \frac{1}{N}\mu_i^N(\mathbf{Q}^N(s))ds + \frac{1}{N}(M_{A,i}^N(t) - M_{D,i}^N(t)), \quad i = 1,2,\ldots,B,
\end{aligned}
\tag{5.8}
$$

or equivalently,

$$
\begin{aligned}
q_i^N(t) = q_i^N(0) &+ \frac{\lambda(N)}{N} \int_0^t f_i(\mathbf{q}^N(s))\mathbb{1}_{[\mathbf{Z}^N(s)\in\mathcal{R}_i]}ds \\
&- \int_0^t \frac{1}{N}\mu_i^N(\mathbf{Q}^N(s))ds + \frac{1}{N}(M_{A,i}^N(t) - M_{D,i}^N(t)), \quad i = 1,2,\ldots,B.
\end{aligned}
\tag{5.9}
$$

Now, we consider the Markov process $(\mathbf{q}^N, \mathbf{Z}^N)(\cdot)$ defined on $S \times G$. Define a random measure α^N on the measurable space $([0,\infty) \times G, \mathcal{C} \otimes \mathcal{G})$, when $[0,\infty)$ is endowed with Borel sigma algebra $\mathcal{C}$, by

$$
\alpha^N(A_1 \times A_2) := \int_{A_1} \mathbb{1}_{[\mathbf{Z}^N(s)\in A_2]}ds,
\tag{5.10}
$$

for $A_1 \in \mathcal{C}$ and $A_2 \in \mathcal{G}$. Then the representation in (5.9) can be written in terms of the random measure as,

$$
\begin{aligned}
q_i^N(t) = q_i^N(0) &+ \lambda \int_{[0,t]\times\mathcal{R}_i} f_i(\mathbf{q}^N(s))d\alpha^N \\
&- \int_0^t \frac{1}{N}\mu_i^N(\mathbf{Q}^N(s))ds + \frac{1}{N}(M_{A,i}^N(t) - M_{D,i}^N(t)), \quad i = 1,2,\ldots,B.
\end{aligned}
\tag{5.11}
$$

Let $\mathfrak{L}$ denote the space of all measures on $[0,\infty) \times G$ satisfying $\gamma([0,t], G) = t$, endowed with the topology corresponding to weak convergence of measures restricted to $[0,t] \times G$ for each t.

Proposition 5.5. *Assume* $\mathbf{q}^N(0) \xrightarrow{\mathcal{L}} \mathbf{q}(0)$ *as* $N \to \infty$, *then* $\{(\mathbf{q}^N(\cdot), \alpha^N)\}$ *is a relatively compact sequence in* $D_S[0,\infty) \times \mathfrak{L}$ *and the limit* $\{(\mathbf{q}(\cdot), \alpha)\}$ *of any convergent subsequence satisfies*

$$
q_i(t) = q_i(0) + \lambda \int_{[0,t]\times\mathcal{R}_i} f_i(\mathbf{q}(s))d\alpha - \int_0^t \mu_i(\mathbf{q}(s))ds, \quad i = 1,2,\ldots,B.
\tag{5.12}
$$

Remark 5.6. Proposition 5.5 is true even when the function $\mathbf{f}$ in the assignment probability depends on N. In that case the proof will go through by assuming that $\mathbf{f}^N$ converges uniformly to some Lipschitz continuous function $\mathbf{f}$ in the sense of Assumption 1.(a).

Remark 5.7. The relative compactness result in the above proposition holds for an even more general class of assignment probabilities than those considered above. Since the proof will follow a nearly identical line of arguments, we briefly mention them here. Consider a scheme for which the assignment probabilities can be written as

$$p_i^N(\mathbf{Q}^N) = \eta_1 \mathbb{1}_{[\mathbf{Z}^N \in \mathcal{R}_i]} + \eta_2 g_i(\mathbf{q}^N), \quad i = 1, \dots, B,$$

for some fixed $\eta_1, \eta_2 \in [0,1]$, and some Lipschitz continuous function $\mathbf{g} = (g_1, g_2, \dots, g_B) :$ $S \to [0, \infty)^B$. The above scheme assigns a fixed fraction η_1 of incoming tasks according to the ordinary JSQ policy, and a fraction η_2 as some suitable function of the fluid-scaled occupancy states $\mathbf{g}(\mathbf{q})$, for $\mathbf{q} \in S$. In practice, the above scheme can handle (two or more) priorities among the incoming tasks, by assigning the high-priority tasks in accordance with the ordinary JSQ policy, and others governed by the JSQ(d) scheme, say. In that case, the fluid limit in (5.12) will become

$$q_i(t) = q_i(0) + \lambda \eta_1 \alpha([0,t] \times \mathcal{R}_i) + \eta_2 \int_0^t g_i(\mathbf{q}(s))ds - \int_0^t \mu_i(\mathbf{q}(s))ds, \quad i = 1, 2, \dots, B. \quad (5.13)$$

To prove Proposition 5.5, we will verify the conditions of relative compactness from [5]. Let (E, r) be a complete and separable metric space. For any $x \in D_E[0, \infty)$, $\delta > 0$ and $T > 0$, define

$$w'(x, \delta, T) = \inf_{\{t_i\}} \max_i \sup_{s,t \in [t_{i-1}, t_i)} r(x(s), x(t)), \quad (5.14)$$

where $\{t_i\}$ ranges over all partitions of the form $0 = t_0 < t_1 < \dots < t_{n-1} < T \leqslant t_n$ with $\min_{1 \leqslant i \leqslant n}(t_i - t_{i-1}) > \delta$ and $n \geqslant 1$. Below we state the conditions for the sake of completeness.

Theorem 5.8. [5, Corollary 3.7.4] *Let (E, r) be complete and separable, and let $\{X_n\}_{n \geqslant 1}$ be a family of processes with sample paths in $D_E[0, \infty)$. Then $\{X_n\}_{n \geqslant 1}$ is relatively compact if and only if the following two conditions hold:*

(a) *For every $\eta > 0$ and rational $t \geqslant 0$, there exists a compact set $\Gamma_{\eta,t} \subset E$ such that*

$$\varliminf_{n \to \infty} \mathbb{P}(X_n(t) \in \Gamma_{\eta,t}) \geqslant 1 - \eta.$$

(b) *For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that*

$$\varlimsup_{n \to \infty} \mathbb{P}(w'(X_n, \delta, T) \geqslant \eta) \leqslant \eta.$$

Proof of Proposition 5.5. The proof goes in two steps. We first prove the relative compactness, and then show that the limit satisfies (5.12).

Observe from [5, Proposition 3.2.4] that, to prove the relative compactness of the process $\{(\mathbf{q}^N(\cdot), \alpha^N)\}$, it is enough to prove relative compactness of the individual components. Note that, from Prohorov's theorem [5, Theorem 3.2.2], $\mathfrak{L}$ is compact, since G is compact. Now, relative compactness of α^N follows from the compactness of $\mathfrak{L}$ under the topology of weak convergence of measures and Prohorov's theorem.

To claim the relative compactness of $\{\mathbf{q}^N(\cdot)\}$, first observe that $[0,1]^B$ is compact with respect to product topology, and S is a closed subset of $[0,1]^B$, and hence S is also

compact with respect to product topology. So, the compact containment condition (a) of Theorem 5.8 is satisfied by taking $\Gamma_{\eta,t} \equiv S$.

For condition (b), we will show for each coordinate i, that for any $\eta > 0$, there exists $\delta > 0$, such that for any $t_1, t_2 > 0$ with $|t_1 - t_2| < \delta$,

$$\overline{\lim_{n \to \infty}} \, \mathbb{P} \left(|q_i^n(t_1) - q_i^n(t_2)| \geqslant \eta \right) = 0.$$

With respect to product topology, this will imply that for any $\eta > 0$, there exists $\delta > 0$, such that for any $t_1, t_2 > 0$ with $|t_1 - t_2| < \delta$,

$$\overline{\lim_{n \to \infty}} \, \mathbb{P} \left(\|q^n(t_1) - q^n(t_2)\| \geqslant \eta \right) = 0,$$

which in turn will imply condition (b) in Theorem 5.8. To see this, observe that for any fixed $\eta > 0$ and $T > 0$, we can choose $\delta' > 0$ small enough, so that for any fine enough finite partition $0 = t_0 < t_1 < \ldots < t_{n-1} < T \leqslant t_n$ of $[0, T]$ with $\min_{1 \leqslant i \leqslant n}(t_i - t_{i-1}) > \delta'$ and $\max_{1 \leqslant i \leqslant n}(t_i - t_{i-1}) < \delta$, $\overline{\lim}_{n \to \infty} \mathbb{P} \left(\|q^n(t_i) - q^n(t_{i+1})\| \geqslant \eta \right) = 0$ for all $1 \leqslant i \leqslant n$.

Now fix any $0 \leqslant t_1 < t_2 < \infty$, and $1 \leqslant i \leqslant B$.

$$|q_i^N(t_1) - q_i^N(t_2)|$$

$$\leqslant \lambda \alpha^N([t_1, t_2] \times \mathcal{R}_i) + \int_{t_1}^{t_2} \frac{1}{N} \mu_i^N(\mathbf{Q}^N(s)) ds$$

$$+ \frac{1}{N} |M_{A,i}^N(t_1) - M_{D,i}^N(t_1) - M_{A,i}^N(t_2) + M_{D,i}^N(t_2)|$$

$$\leqslant \lambda'(t_2 - t_1) + \frac{1}{N} |M_{A,i}^N(t_1) - M_{D,i}^N(t_1) - M_{A,i}^N(t_2) + M_{D,i}^N(t_2)|,$$

for some $\lambda' \in \mathbb{R}$, using the linear growth constraint of μ^N due to Assumption 1(c). Now, from Proposition 5.4, we get, for any $T \geqslant 0$,

$$\sup_{t \in [0,T]} \frac{1}{N} |M_{A,i}^N(t_1) - M_{D,i}^N(t_1) - M_{A,i}^N(t_2) + M_{D,i}^N(t_2)| \xrightarrow{\mathbb{P}} 0.$$

To prove that the limit $\{(\mathbf{q}(\cdot), \alpha)\}$ of any convergent subsequence satisfies (5.12), we will use the continuous-mapping theorem [30, Section 3.4]. Specifically, we will show that the right side of (5.11) is a continuous map of suitable arguments. Let $\{\mathbf{q}(t)\}_{t \geqslant 0}$ and $\{\mathbf{y}(t)\}_{t \geqslant 0}$ be an S-valued and an $\mathbb{R}^B$-valued cádlág function, respectively. Also, let α be a measure on the measurable space $([0, \infty) \times G, \mathcal{C} \otimes \mathcal{G})$. Then for $\mathbf{q}^0 \in S$, define for $i \geqslant 1$,

$$F_i(\mathbf{q}, \alpha, \mathbf{q}^0, \mathbf{y})(t) := q_i^0 + y_i(t) + \lambda \int_{[0,t] \times \mathcal{R}_i} f_i(\mathbf{q}(s)) d\alpha - \int_0^t \mu_i(\mathbf{q}(s)) ds.$$

Observe that it is enough to show $\mathbf{F} = (F_1, \ldots, F_B)$ is a continuous operator. Indeed, in that case the right side of (5.11) can be written as $\mathbf{F}(\mathbf{q}^N, \alpha^N, \mathbf{q}^N(0), \mathbf{y}^N)$, where $\mathbf{y}^N = (y_1^N, \ldots, y_B^N)$ with $y_i^N = (M_{A,i}^N - M_{D,i}^N)/N$, and since each argument converges we will get the convergence to the right side of (5.12). Therefore, we now prove the continuity of $\mathbf{F}$ below. In particular assume that the sequence of processes $\{(\mathbf{q}^N, \mathbf{y}^N)\}_{N \geqslant 1}$ converges to $\{(\mathbf{q}, \mathbf{y})\}$, for any fixed $t \geqslant 0$, the measure $\alpha^N([0, t], \cdot)$ on G converges weakly to $\alpha([0, t], \cdot)$, and the sequence of S-valued random variables $\mathbf{q}^N(0)$ converges weakly to $\mathbf{q}(0)$. Fix any $T \geqslant 0$ and $\varepsilon > 0$.

28

(i) Choose $N_1 \in \mathbb{N}$, such that $\sup_{t \in [0,T]} \left\| \mathbf{q}^N(t) - \mathbf{q}(t) \right\| < \varepsilon/(4TC_\mu)$. In that case, observe that

$$\sup_{t \in [0,T]} \int_0^t \left\| \boldsymbol{\mu}(\mathbf{q}^N(s)) - \boldsymbol{\mu}(\mathbf{q}(s)) \right\| ds \leqslant T \sup_{t \in [0,T]} \left\| \boldsymbol{\mu}(\mathbf{q}^N(t)) - \boldsymbol{\mu}(\mathbf{q}(t)) \right\|$$

$$\leqslant TC_\mu \sup_{t \in [0,T]} \left\| \mathbf{q}^N(t) - \mathbf{q}(t) \right\| < \frac{\varepsilon}{4},$$

where we have used the Lipschitz continuity of $\boldsymbol{\mu}$ due to Assumption 1(b).

(ii) Choose $N_2 \in \mathbb{N}$, such that $\sup_{t \in [0,T]} \left\| \mathbf{y}^N(t) - \mathbf{y}(t) \right\| < \varepsilon/4$,

(iii) Choose $N_3 \in \mathbb{N}$, such that

$$\sum_{i \geqslant 1} \frac{\lambda}{2^i} \left| \int_{[0,T] \times \mathcal{R}_i} f_i(\mathbf{q}^N(s)) d\alpha^N - \int_{[0,T] \times \mathcal{R}_i} f_i(\mathbf{q}(s)) d\alpha \right| < \frac{\varepsilon}{4}.$$

This can be done as follows: choose $M \in \mathbb{N}$ large enough so that $\sum_{i > M} 2^{-i} < \varepsilon/8$. Now for $i \leqslant M$, since $\alpha^N([0,T], \cdot)$ converges weakly to $\alpha([0,T], \cdot)$, and M is finite, we can choose $N_3 \in \mathbb{N}$ such that

$$\sum_{i=1}^M \frac{\lambda}{2^i} \left| \int_{[0,T] \times \mathcal{R}_i} f_i(\mathbf{q}^N(s)) d\alpha^N - \int_{[0,T] \times \mathcal{R}_i} f_i(\mathbf{q}(s)) d\alpha \right|$$

$$\leqslant \sum_{i=1}^M \frac{\lambda}{2^i} \int_{[0,T] \times \mathcal{R}_i} |f_i(\mathbf{q}^N(s)) - f_i(\mathbf{q}(s))| d\alpha^N + \sum_{i=1}^M \frac{\lambda}{2^i} |\alpha^N([0,T] \times \mathcal{R}_i) - \alpha([0,T] \times \mathcal{R}_i)|$$

$$\leqslant \sum_{i=1}^M \frac{\lambda}{2^i} TC_f \sup_{s \in [0,T]} \left\| \mathbf{q}^N(s) - \mathbf{q}(s) \right\| + \sum_{i=1}^M \frac{\lambda}{2^i} |\alpha^N([0,T] \times \mathcal{R}_i) - \alpha([0,T] \times \mathcal{R}_i)| < \frac{\varepsilon}{4}.$$

(iv) Choose $N_4 \in \mathbb{N}$, such that $\left\| \mathbf{q}^N(0) - \mathbf{q}(0) \right\| < \varepsilon/4$.

Let $\hat{N} = \max \left\{ N_1, N_2, N_3, N_4 \right\}$, then for $N \geqslant \hat{N}$,

$$\sup_{t \in [0,T]} \left\| \mathbf{F}(\mathbf{q}^N, \alpha^N, \mathbf{q}^N(0), \mathbf{y}^N) - \mathbf{F}(\mathbf{q}, \alpha, \mathbf{q}(0), \mathbf{y}) \right\| (t) < \varepsilon.$$

Thus the proof of continuity of $\mathbf{F}$ is complete. $\qquad\square$

To characterize the limit in (5.12), for any $\mathbf{q} \in S$, define the Markov process $\mathbf{Z_q}$ on G as

$$\mathbf{Z_q} \to \begin{cases} \mathbf{Z_q} + e_i & \text{at rate} \quad \mu_i(\mathbf{q}) \\ \mathbf{Z_q} - e_i & \text{at rate} \quad \lambda \mathbb{1}_{[\mathbf{Z_q} \in \mathcal{R}_i]}, \end{cases} \tag{5.15}$$

where e_i is the ith unit vector, $i = 1, \ldots, B$.

Proof of Theorem 5.1. Having proved the relative compactness in Proposition 5.5, it follows from analogous arguments as used in the proof of [8, Theorem 3], that the limit of any convergent subsequence of the sequence of processes $\left\{ \mathbf{q}^N(t) \right\}_{t \geqslant 0}$ satisfies

$$q_i(t) = q_i(0) + \lambda \int_0^t \pi_{\mathbf{q}(s)}(\mathcal{R}_i) ds - \int_0^t \mu_i(\mathbf{q}(s)) ds, \quad i = 1, 2, \ldots, B, \tag{5.16}$$

for *some* stationary measure $\pi_{\mathbf{q}(t)}$ of the Markov process $\mathbf{Z}_{\mathbf{q}(t)}$ described in (5.15) satisfying $\pi_{\mathbf{q}}\{\mathbf{Z} : Z_i = \infty\} = 1$ if $q_i < 1$.

Now it remains to show that $\mathbf{q}(t)$ *uniquely* determines $\pi_{\mathbf{q}(t)}$, and that $\pi_{\mathbf{q}(s)}(\mathcal{R}_i) = p_{i-1}(\mathbf{q}(s))$ described in (5.2). As mentioned earlier, in this proof we will now assume the specific assignment probabilities in (5.1), corresponding to the ordinary JSQ policy. To see this, fix any $\mathbf{q} = (q_1, \ldots, q_B) \in S$, and assume that there exists $m \geqslant 0$, such that $q_{m+1} < 1$ and $q_1 = \ldots = q_m = 1$, with the convention that $q_0 \equiv 1$ and $q_{B+1} \equiv 0$ if $B < \infty$. In that case,

$$\pi_{\mathbf{q}}\big(\{Z_{m+1} = \infty, Z_{m+2} = \infty, \ldots, Z_B = \infty\}\big) = 1.$$

Also, note that $q_i = 1$ forces $dq_i/dt \leqslant 0$, i.e., $\lambda\pi_{\mathbf{q}}(\mathcal{R}_i) \leqslant \mu_i(\mathbf{q})$ for all $i = 1, \ldots, m$, and in particular $\pi_{\mathbf{q}}(\mathcal{R}_i) = 0$ for all $i = 1, \ldots, m-1$. Thus,

$$\pi_{\mathbf{q}}\big(\{Z_1 = 0, Z_2 = 0, \ldots, Z_{m-1} = 0\}\big) = 1.$$

Therefore, $\pi_{\mathbf{q}}$ is determined only by the stationary distribution of the mth component, which can be described as a birth-death process

$$Z \to \begin{cases} Z+1 & \text{at rate} & \mu_m(\mathbf{q}) \\ Z-1 & \text{at rate} & \lambda\mathbb{1}_{[Z>0]} \end{cases} \tag{5.17}$$

and let $\pi^{(m)}$ be its stationary distribution. Now it is enough to show that $\pi^{(m)}$ is uniquely determined by $\mu_m(\mathbf{q})$. First observe that the process on $\bar{\mathbb{Z}}$ described in (5.17) is reducible, and can be decomposed into two irreducible classes given by $\mathbb{Z}$ and $\{\infty\}$, respectively. Therefore, if $\pi^{(m)}(Z = \infty) = 0$ or 1, then it is unique. Indeed, if $\pi^{(m)}(Z = \infty) = 0$, then Z is birth-death process on $\mathbb{Z}$ only, and hence it has a unique stationary distribution. Otherwise, if $\pi^{(m)}(Z = \infty) = 1$, then it is trivially unique. Now we distinguish between two cases depending upon whether $\mu_m(\mathbf{q}) \geqslant \lambda$ or not.

Note that if $\mu_m(\mathbf{q}) \geqslant \lambda$, then $\pi^{(m)}(Z \geqslant k) = 1$ for all $k \geqslant 0$. On $\bar{\mathbb{Z}}$ this shows that $\pi^{(m)}(Z = \infty) = 1$. Furthermore, if $\mu_m(\mathbf{q}) < \lambda$, we will show that $\pi^{(m)}(Z = \infty) = 0$. On the contrary, assume $\pi^{(m)}(Z = \infty) = \varepsilon \in (0, 1]$. Also, let $\hat{\pi}^{(m)}$ be the unique stationary distribution of the birth-death process in (5.17) restricted to $\mathbb{Z}$. Therefore,

$$\pi^{(m)}(Z > 0) = \hat{\pi}^{(m)}(Z > 0) + \varepsilon > \hat{\pi}^{(m)}(Z > 0) = \frac{\mu_m(\mathbf{q})}{\lambda},$$

and $\pi_{\mathbf{q}}(\mathcal{R}_m) = \pi^{(m)}(Z > 0) > \mu_m(\mathbf{q})/\lambda$. Putting this value in the fluid-limit equation (5.3), we obtain that $dq_m(t)/dt > 0$. Since $q_m(t) = 1$, this leads to a contradiction, and hence it must be the case that $\pi^{(m)}(Z = \infty) = 0$.

Therefore, for all $\mathbf{q} \in S$, $\pi_{\mathbf{q}}$ is uniquely determined by $\mathbf{q}$. Furthermore, we can identify the expression for $\pi_{\mathbf{q}}(\mathcal{R}_i)$ as

$$\pi_{\mathbf{q}}(\mathcal{R}_i) = \begin{cases} \min\{\mu_i(\mathbf{q})/\lambda, 1\} & \text{for} \quad i = m, \\ 1 - \min\{\mu_i(\mathbf{q})/\lambda, 1\} & \text{for} \quad i = m+1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.18}$$

and hence $\pi_{\mathbf{q}(s)}(\mathcal{R}_i) = p_{i-1}(\mathbf{q}(s))$ as claimed. $\qquad\square$

6 Diffusion Limit of JSQ: Non-integral λ

In this section we establish the diffusion-scale behavior of the ordinary JSQ policy in the case when λ is not an integer, i.e., $f > 0$. Recall that $f(N) = \lambda(N) - KN$. In this regime, let us define the following centered and scaled processes:

$$
\begin{aligned}
\bar{Q}_i^N(t) &= N - Q_i^N(t) \geqslant 0 \quad \text{for} \quad i \leqslant K-1 \\
\bar{Q}_K^N(t) &:= \frac{N - Q_K^N(t)}{\log(N)} \geqslant 0 \\
\bar{Q}_{K+1}^N(t) &:= \frac{Q_{K+1}^N(t) - f(N)}{\sqrt{N}} \in \mathbb{R} \\
\bar{Q}_i^N(t) &:= Q_i^N(t) \geqslant 0 \quad \text{for} \quad i \geqslant K+2.
\end{aligned}
\tag{6.1}
$$

Theorem 6.1. [Diffusion limit for JSQ policy; $f > 0$] *Assume $\bar{Q}_i^N(0) \xrightarrow{\mathcal{L}} \bar{Q}_i(0)$ in $\mathbb{R}$, $i \geqslant 1$, and $\lambda(N)/N \to \lambda > 0$ as $N \to \infty$, with $f = \lambda - \lfloor \lambda \rfloor > 0$, then*

(i) $\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \bar{Q}_{K-1}^N(t) \leqslant 1 \right) = 1$, *and* $\{\bar{Q}_i^N(t)\}_{t\geqslant 0} \xrightarrow{\mathcal{L}} \{\bar{Q}_i(t)\}_{t\geqslant 0}$, *where* $\bar{Q}_i(t) \equiv 0$, *provided* $\lim_{N\to\infty} \mathbb{P}\left(\bar{Q}_{K-1}^N(0) \leqslant 1 \right) = 1$, *and* $\bar{Q}_i^N(0) \xrightarrow{\mathbb{P}} 0$ *for* $i \leqslant K-2$.

(ii) $\{\bar{Q}_K^N(t)\}_{t\geqslant 0}$ *is a stochastically bounded sequence of processes in* $D_{\mathbb{R}}[0,\infty)$.

(iii) $\{\bar{Q}_{K+1}^N(t)\}_{t\geqslant 0} \xrightarrow{\mathcal{L}} \{\bar{Q}_{K+1}(t)\}_{t\geqslant 0}$, *where* $\bar{Q}_{K+1}(t)$ *is given by the Ornstein-Uhlenbeck process satisfying the following stochastic differential equation:*

$$
d\bar{Q}_{K+1}(t) = -\bar{Q}_{K+1}(t)dt + \sqrt{2\lambda}dW(t),
$$

where $W(t)$ is the standard Brownian motion, provided $\bar{Q}_{K+1}^N(0) \xrightarrow{\mathcal{L}} \bar{Q}_{K+1}(0)$ in $\mathbb{R}$.

(iv) *For $i \geqslant K+2$, $\{\bar{Q}_i^N(t)\}_{t\geqslant 0} \xrightarrow{\mathcal{L}} \{\bar{Q}_i(t)\}_{t\geqslant 0}$, where $\bar{Q}_i(t) \equiv 0$, provided $\bar{Q}_i^N(0) \xrightarrow{\mathbb{P}} 0$.*

Note that statements (i) and (ii) in Theorem 6.1 imply statement (i) in Theorem 2.3, for the JSQ policy, while (iii) and (iv) in Theorem 6.1 are equivalent with statements (ii) and (iii) in Theorem 2.3. In view of the universality result in Corollary 3.4, it thus suffices to prove Theorem 6.1.

The rest of this section is devoted to the proof of Theorem 6.1. From a high level, the idea of the proof is the following. Introduce

$$
Y^N(t) := \sum_{i=1}^{B} Q_i^N(t), \qquad D_+^N(t) := \sum_{i=1}^{K} (N - Q_i^N(t)), \qquad D_-^N(t) := \sum_{i=K+2}^{B} Q_i^N(t). \tag{6.2}
$$

and observe that

$$
\begin{aligned}
Q_{K+1}^N(t) + KN &= \sum_{i=1}^{B} Q_i^N(t) + \sum_{i=1}^{K} (N - Q_i^N(t)) - \sum_{i=K+2}^{B} Q_i^N(t) \\
&= Y^N(t) + D_+^N(t) - D_-^N(t).
\end{aligned}
$$

We show in Proposition 6.4 that the sequence of processes $\{D_+^N(t)\}_{t\geqslant 0}$ is $O_\mathbb{P}(\log(N))$, which implies that the number of server pools with fewer than K active tasks is negligible

on $\sqrt{N}$-scale. Furthermore, in Proposition 6.3 we prove that since $\lambda < B$ the number of tasks that are assigned to server pools with at least $K + 1$ tasks converges to zero in probability and hence, for a suitable starting state, $\{D_-^N(t)\}_{t \geqslant 0}$ converges to the zero process. As we will show, this also means that $Y^N(t)$ behaves with high probability as the total number of tasks in an $M/M/\infty$ system. Therefore with the help of the following diffusion limit result for the $M/M/\infty$ system in [21, Theorem 6.14], we conclude the proof of statement (iii) of Theorem 6.1.

Theorem 6.2 ([21, Theorem 6.14]). *Let* $\{Y_\infty^N(t)\}_{t \geqslant 0}$ *be the total number of tasks in an $M/M/\infty$ system with arrival rate $\lambda(N)$ and unit-mean service time. If $(Y_\infty^N(0) - \lambda(N))/\sqrt{N} \to \nu \in \mathbb{R}$, then the process $\{\bar{Y}_\infty^N(t)\}_{t \geqslant 0}$, with*

$$\bar{Y}_\infty^N(t) = \frac{Y_\infty^N(t) - \lambda(N)}{\sqrt{N}},$$

converges weakly to an Ornstein-Uhlenbeck process $\{X(t)\}_{t \geqslant 0}$ described by the stochastic differential equation

$$X(0) = \nu, \qquad dX(t) = -X(t)dt + \sqrt{2\lambda}dW(t).$$

The next two propositions state asymptotic properties of $\{D_+^N(t)\}_{t \geqslant 0}$ and $\{D_-^N(t)\}_{t \geqslant 0}$ mentioned before, which play a crucial role in the proof of Theorem 6.1. Let $B_{K+1}^N(t)$ be the cumulative number of tasks up to time t that are assigned to some server pool having at least $K + 1$ active tasks if $B > K + 1$, and that are lost if $B = K + 1$.

Proposition 6.3. *Under the assumptions of Theorem 6.1, for any $T \geqslant 0$, $B_{K+1}^N(T) \xrightarrow{\mathbb{P}} 0$, and consequently, $\sup_{t \in [0,T]} D_-^N(t) \xrightarrow{\mathbb{P}} 0$ as $N \to \infty$, provided $D_-^N(0) \xrightarrow{\mathbb{P}} 0$.*

Informally speaking, the above proposition implies that for large N, there will be almost no server pool with $K + 2$ or more tasks in any finite time horizon, if the system starts with no server pools with more than $K + 1$ tasks. The next proposition shows that the number of server pools having fewer than K tasks is of order $\log(N)$ in any finite time horizon.

Proposition 6.4. *Under the assumptions of Theorem 6.1, the sequence $\{D_+^N(t)/\log(N)\}_{t \geqslant 0}$ is stochastically bounded in $D_\mathbb{R}[0, \infty)$, provided $\{D_+^N(0)/\log(N)\}_{N \geqslant 1}$ is a tight sequence of random variables.*

Before providing the proofs of the above two propositions, we first prove Theorem 6.1 using Propositions 6.3 and 6.4.

Proof of Theorem 6.1. First observe that (iv) and (ii) immediately follows from Propositions 6.3 and 6.4, respectively.

To prove (i), fix any $T \geqslant 0$. We will show that

$$\lim_{N \to \infty} \mathbb{P}\left(\sup_{t \in [0,T]} \sum_{i=1}^{K-1} \bar{Q}_i^N(t) \leqslant 1\right) = 1. \tag{6.3}$$

Since $\bar{Q}_i^N \leqslant 1$ implies that $\bar{Q}_{i-1}^N \leqslant 1$ for $i = 2, \ldots, K$, this then completes the proof of (i). Note that the process $\sum_{i=1}^{K-1} \bar{Q}_i^N(\cdot)$ increases by one when there is a departure from some

server pool with at most $K-1$ active tasks, and if positive, decreases by one whenever there is an arrival. Therefore it can be thought of as a birth-death process with state-dependent instantaneous birth rate $\sum_{i=1}^{K-1} i(Q_i^N(t) - Q_{i+1}^N(t))$, and constant instantaneous death rate $\lambda(N)$. Since

$$\sum_{i=1}^{K-1} i(Q_i^N(t) - Q_{i+1}^N(t)) = \sum_{i=1}^{K-1} Q_i^N(t) - (K-1)Q_K^N(t) \leqslant (K-1)(N - Q_K^N(t)),$$

the process $\left\{ \sum_{i=1}^{K-1} \bar{Q}_i^N(t) \right\}_{t \geqslant 0}$ is stochastically upper bounded by a birth-and-death process $\left\{ Z^N(t) \right\}_{t \geqslant 0}$ with birth rate $(K-1)(N - Q_K^N(t))$ and constant death rate $\lambda(N)$. Due to (ii), we can claim that for *any* nonnegative sequence $\ell(N)$ diverging to infinity,

$$\lim_{N \to \infty} \mathbb{P} \left(\sup_{t \in [0,T]} (N - Q_K^N(t)) \leqslant \ell(N) \log(N) \right) = 1.$$

Let $\left\{ \eta^N(n) \right\}_{n \geqslant 1}$ denote the discrete uniformized chain of the upper bounding birth-death process. Also, let $K_N(t)$ denote the number of jumps taken up to time t by $\left\{ \eta^N(n) \right\}_{n \geqslant 1}$. Since the jump rate of the process is $O(N)$, we have for *any* nonnegative sequence $\ell^0(N)$ diverging to infinity, and for any $T \geqslant 0$,

$$\lim_{N \to \infty} \mathbb{P} \left(K_N(T) \leqslant N\ell^0(N) \right) = 1.$$

Given Q_K^N, considering the $\eta^N(\cdot)$ Markov chain, the probability of one birth is bounded from above by

$$p_{Q_K^N} = \frac{(K-1)(N - Q_K^N)}{N + (K-1)(N - Q_K^N)}.$$

Now, $Z^N(\cdot)$ will exceed 1 if and only if there are at least two successive births. Hence,

$$\begin{aligned}
\mathbb{P} \left(\sup_{t \in [0,T]} Z^N(t) \leqslant 1 \right) &= \mathbb{P} \left(\sup_{n \leqslant K_N(T)} \eta^N(n) \leqslant 1 \right) \\
&\geqslant \mathbb{P} \left(\sup_{n \leqslant N\ell^0(N)} \eta^N(n) \leqslant 1 \right) \mathbb{P} \left(K_N(T) \leqslant N\ell^0(N) \right) + \mathbb{P} \left(K_N(T) > N\ell^0(N) \right).
\end{aligned} \tag{6.4}$$

Again we can write the first term of the last inequality above as

$$\begin{aligned}
&\mathbb{P} \left(\sup_{n \leqslant N\ell^0(N)} \eta^N(n) \leqslant 1 \right) \\
&\geqslant \mathbb{P} \left(\sup_{n \leqslant N\ell^0(N)} \eta^N(n) \leqslant 1 \,\middle|\, \sup_{t \in [0,T]} (N - Q_K^N(t)) \leqslant \ell(N) \log(N) \right) \\
&\qquad\qquad\qquad\qquad \times \mathbb{P} \left(\sup_{t \in [0,T]} (N - Q_K^N(t)) \leqslant \ell(N) \log(N) \right) \\
&\geqslant \left(1 - \left(\frac{(K-1)\ell(N) \log(N)}{N + (K-1)\ell(N) \log(N)} \right)^2 \right)^{N\ell^0(N)} \times \mathbb{P} \left(\sup_{t \in [0,T]} (N - Q_K^N(t)) \leqslant \ell(N) \log(N) \right).
\end{aligned}$$

33

If we choose $\ell(N)$ and $\ell^0(N)$ such that $\ell(N)^2 \ell^0(N) \log(N)/N \to 0$ as $N \to \infty$, then the expression on the right of (6.4) converges to 1 (one can see that this choice is always feasible). Hence the proof of (i) is complete.

For (iii), recall that $Y_\infty^N(t)$ denotes the total number of tasks in an $M/M/\infty$ system with arrival rate $\lambda(N)$ and exponential service time distribution with unit mean. Also, Proposition 6.3 implies that under the assumptions of the theorem, in any finite time horizon, with high probability there will be no arrival to a server pool with $K+1$ or more active tasks. Now observe that since $B \geqslant K+1$, for any $T \geqslant 0$,

$$\mathbb{P}\left(\exists\, t \in [0, T] : Y^N(t) \neq Y_\infty^N(t)\right) \leqslant \mathbb{P}\left(\exists\, t \in [0, T] : B_{K+1}^N(t) \geqslant 1\right) \to 0, \quad \text{as} \quad N \to \infty.$$

Propositions 6.3 and 6.4 then yield

$$\sup_{t \in [0,T]} \frac{1}{\sqrt{N}} \left| Q_{K+1}^N(t) - f(N) - (Y_\infty^N(t) - \lambda(N)) \right|$$

$$= \sup_{t \in [0,T]} \frac{1}{\sqrt{N}} \left| \sum_{i=1}^{B} Q_i^N(t) + \sum_{i=1}^{K} (N - Q_i^N(t)) - \sum_{i=K+2}^{B} Q_i^N(t) - KN - f(N) - (Y_\infty^N(t) - \lambda(N)) \right|$$

$$= \sup_{t \in [0,T]} \frac{1}{\sqrt{N}} \left[Y^N(t) - Y_\infty^N(t) + D_N^+(t) - D_-^N(t) \right] \to 0,$$

as $N \to \infty$, which in conjunction with [21, Theorem 6.14], as mentioned earlier, gives the desired diffusion limit. □

Proof of Proposition 6.3. Couple the $M/M/\infty$ system and a system under the ordinary JSQ policy in the natural way, until an overflow event occurs in the latter system. Observe that for any fixed $M > 0$, the event $\left[\sup_{t \in [0,T]} B_{K+1}^N(t) \geqslant M\right]$ will occur only if for some $t' \leqslant T$, some arriving task is assigned to a server pool with more than K active tasks, and in that case, there exists $t'' \leqslant t'$, such that $Y^N(t'') > (\lambda + \varepsilon)N$, for some $\varepsilon > 0$ with $\lambda + \varepsilon < 1$. Since, for any $t \in [0, t'']$, $Y^N(t) = Y_\infty^N(t)$, we have

$$\sup_{t \in [0,T]} B_{K+1}^N(t) \geqslant M$$

$$\implies \sup_{t'' \in [0,t']} Y^N(t'') \geqslant (\lambda + \varepsilon)N$$

$$\implies \sup_{t'' \in [0,t']} Y_\infty^N(t'') \geqslant (\lambda + \varepsilon)N \tag{6.5}$$

$$\implies \sup_{t \in [0,T]} (Y_\infty^N(t) - \lambda(N)) > \varepsilon N + o(N)$$

$$\implies \sup_{t \in [0,T]} \frac{1}{\sqrt{N}} (Y_\infty^N(t) - \lambda(N)) > \varepsilon \sqrt{N} + o(\sqrt{N}).$$

From Theorem 6.14 of [21], we know that the process $\left\{(Y^N(t) - \lambda(N))/\sqrt{N}\right\}_{t \geqslant 0}$ is stochastically bounded. Hence, Equation (6.5) yields that for any $T \geqslant 0$, $\sup_{t \in [0,T]} B_{K+1}^N(t)$ converges to zero in probability as $N \to \infty$. Consequently, from the assumption of Theorem 6.1 that $D_-^N(0) \xrightarrow{\mathbb{P}} 0$, the conclusion $\sup_{t \in [0,T]} D_-^N(t) \xrightarrow{\mathbb{P}} 0$, is immediate. □

Proof of Proposition 6.4. Observe that $\sum_{i=1}^{K}(N - Q_i^N(\cdot))$ increases by one when there is a departure from some server pool with at most K active tasks, and if positive, decreases by one whenever there is an arrival. Therefore the process $\{D_+^N(t)\}_{t \geqslant 0}$ increases by one at rate $\sum_{i=1}^{K} i(Q_i(t) - Q_{i+1}(t)) = \sum_{i=1}^{K}(Q_i(t) - Q_{K+1}(t))$, and while positive, decreases by one at constant rate $\lambda(N)$. Now, to prove stochastic boundedness of the sequence of processes $\{D_+^N(t)/\log(N)\}_{t \geqslant 0}$, we will show that for any fixed $T \geqslant 0$ and any function $\ell(N)$ diverging to infinity (i.e., such that $\ell(N) \to \infty$ as $N \to \infty$),

$$\mathbb{P}\left(\sup_{t \in [0,T]} D_+^N(t) > \ell(N)\log(N)\right) \to 0. \tag{6.6}$$

Let $\{X^N(n)\}_{n \geqslant 0}$ be the discrete jump chain, and $K_N(t)$ be the number of jumps before time t, of the process $\{D_+^N(t)\}_{t \geqslant 0}$. Hence, for any fixed $T \geqslant 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{t \in [0,T]} D_+^N(t) > \ell(N)\log(N)\right) \\
&= \mathbb{P}\left(\sup_{n \leqslant K_N(T)} X^N(n) > \ell(N)\log(N)\right) \\
&\leqslant \mathbb{P}\left(\sup_{n \leqslant N\ell_0(N)} X^N(n) > \ell(N)\log(N)\right) \mathbb{P}\left(K_N(T) \leqslant N\ell_0(N)\right) \\
&\quad + \mathbb{P}\left(K_N(T) > N\ell_0(N)\right),
\end{aligned}
\tag{6.7}
$$

for some function $\ell_0(N) : \mathbb{N} \to \mathbb{N}$, to be chosen according to Lemma 6.5 below. Now, observe that $K_N(T)$ is upper bounded by a Poisson random variable with parameter $\lambda(N)T + \int_0^T \sum_{i=1}^{K}(Q_i(s) - Q_{K+1}(s))ds$, and $\sum_{i=1}^{K}(Q_i(s) - Q_{K+1}(s)) \leqslant KN$. Hence for any function $\ell_0(N)$ diverging to infinity, we have

$$\mathbb{P}\left(K_N(T) > N\ell_0(N)\right) \to 0.$$

To control the first term, it is enough to note that $\sum_{i=1}^{K}(Q_i(t) - Q_{K+1}(t)) \leqslant KN < \lambda N$. Hence the process $\{X^N(n)\}_{n \geqslant 1}$ can be stochastically upper bounded by the process $\{\hat{X}^N(n)\}_{n \geqslant 1}$, defined as follows:

$$\hat{X}^N(n+1) = \begin{cases} \hat{X}^N(n) + 1 & \text{with prob. } K/(K+\lambda) \\ (\hat{X}^N(n) - 1) \vee 0 & \text{with prob. } \lambda/(K+\lambda) \end{cases} \tag{6.8}$$

Therefore combining Lemma 6.5 below for the above Markov process $\{\hat{X}^N(n)\}_{n \geqslant 0}$ with Equation (6.7) we obtain Equation (6.6). Hence the proof is complete. $\qquad \square$

Lemma 6.5. *For any function $\ell(N) : \mathbb{N} \to \mathbb{N}$, diverging to infinity, there exists another function $\ell_0(N) : \mathbb{N} \to \mathbb{N}$, diverging to infinity, such that*

$$\mathbb{P}\left(\sup_{n \leqslant N\ell_0(N)} \hat{X}^N(n) > \ell(N)\log(N)\right) \to 0.$$

35

Proof. We will use a regenerative approach to prove the lemma. Let $p := K/(K + \lambda)$. Note that then $p < q := 1 - p$. Define the ith regeneration time ρ_i of the Markov chain as follows: $\rho_0 = 0$, and $\rho_i := \min \{ k > \rho_{i-1} : \hat{X}_k = 0 \}$, for $i \geqslant 1$. Also define, $m_i := \max \{ \hat{X}_k : \rho_{i-1} \leqslant k < \rho_i \}$, for $i \geqslant 1$, and $\xi(n) := \min \{ i : \rho_i \geqslant n \}$, for $n \geqslant 1$. Now observe that [6, XIV.2],

$$\mathbb{P}\left(m_i \geqslant M \right) = p \times \frac{\frac{q}{p} - 1}{\left(\frac{q}{p} \right)^M - 1} \leqslant a^{-M}, \tag{6.9}$$

for some $a > 1$, since $q/p > 1$. Thus the tail of the distribution of the maximum attained in one regeneration period decays exponentially. Recall that, in n steps the Markov chain exhibits $\xi(n)$ regenerations. Hence, for any $\ell_0(N)$ and $\ell(N)$,

$$\mathbb{P}\left(\sup_{n \leqslant N\ell_0(N)} \hat{X}^N(n) > \ell(N) \log(N) \right) = \mathbb{P}\left(\sup_{i \leqslant \xi(N\ell_0(N))} m_i > \ell(N) \log(N) \right)$$

$$\leqslant 1 - \left(1 - a^{-\ell(N)\log(N)} \right)^{\xi(N\ell_0(N))} \leqslant 1 - \left(1 - a^{-\ell(N)\log(N)} \right)^{N\ell_0(N)}. \tag{6.10}$$

Now, for given $\ell(N)$, choose $\ell_0(N)$ diverging to infinity, such that

$$N\ell_0(N) a^{-\ell(N)\log(N)} \to 0 \quad \text{as} \quad N \to \infty.$$

Since the condition is equivalent to

$$\log(N) + \log(\ell_0(N)) - \ell(N) \log(a) \log(N) \to -\infty,$$

it is evident that such a choice of $\ell_0(N)$ is always possible. Hence, for such a choice of $\ell_0(N)$ the probability in Equation (6.10) converges to zero and the proof is complete. $\square$

7 Diffusion Limit of JSQ: Integral λ

In this section we analyze the diffusion-scale behavior of the ordinary JSQ policy when λ is an integer, i.e., $f = 0$, and

$$\frac{KN - \lambda(N)}{\sqrt{N}} \to \beta, \quad \text{as} \quad N \to \infty,$$

with $\beta \in \mathbb{R}$ being a fixed real number. Throughout this section we assume $B = K + 1$. Thus, tasks that arrive when all the server pools have $K + 1$ active tasks, are permanently discarded. For brevity in notation, define, $Z_1^N(t) = \sum_{i=1}^{K}(N - Q_i^N(t))$ and $Z_2^N(t) := Q_{K+1}^N(t)$. Note that $Z_1^N(t)$ corresponds to $D_+^N(t)$ in the previous section. Also recall (2.6), and define

$$\zeta_1^N(t) := \frac{Z_1^N(t)}{\sqrt{N}} = \hat{Q}_{K-1}^N(t) + \hat{Q}_K^N(t)$$

$$\zeta_2^N(t) := \frac{Z_2^N(t)}{\sqrt{N}} = \hat{Q}_{K+1}^N(t), \tag{7.1}$$

with $\hat{Q}_{K-1}^N(t)$, $\hat{Q}_K^N(t)$, and $\hat{Q}_{K+1}^N(t)$ as in (2.6).

Theorem 7.1. *Assume that* $(\zeta_1^N(0), \zeta_2^N(0)) \xrightarrow{\mathcal{L}} (\zeta_1(0), \zeta_2(0))$ *in* $\mathbb{R}^2$ *as* $N \to \infty$. *Then the two-dimensional process* $\{(\zeta_1^N(t), \zeta_2^N(t))\}_{t \geqslant 0}$ *converges weakly to the process* $\{(\zeta_1(t), \zeta_2(t))\}_{t \geqslant 0}$ *in* $D_{\mathbb{R}^2}[0, \infty)$ *governed by the following stochastic recursion equation:*

$$\zeta_1(t) = \zeta_1(0) + \sqrt{2K}W(t) - \int_0^t (\zeta_1(s) + K\zeta_2(s))ds + \beta t + V_1(t)$$

$$\zeta_2(t) = \zeta_2(0) + V_1(t) - (K+1)\int_0^t \zeta_2(s)ds,$$

where W *is the standard Brownian motion, and* $V_1(t)$ *is the unique non-decreasing process in* $D_{\mathbb{R}_+}[0, \infty)$ *satisfying*

$$\int_0^t \mathbb{1}_{[\zeta_1(s) \geqslant 0]} dV_1(s) = 0.$$

Remark 7.2. Note that $Y^N(t) - KN = Z_2^N(t) - Z_1^N(t)$. Thus, under the assumption in (2.5), the diffusion limit in Theorem 7.1 implies that

$$\frac{Y^N(\cdot) - \lambda(N)}{\sqrt{N}} = \frac{Y^N(\cdot) - KN}{\sqrt{N}} + \frac{KN - \lambda(N)}{\sqrt{N}} \xrightarrow{\mathcal{L}} \zeta_2(\cdot) - \zeta_1(\cdot) + \beta.$$

Writing $X(t) = \zeta_2(t) - \zeta_1(t) - \beta$, from Theorem 7.1, one can note that the process $\{X(t)\}_{t \geqslant 0}$ satisfies

$$dX(t) = -X(t)dt - \sqrt{2K}dW(t),$$

which is consistent with the diffusion-level behavior of $Y^N(\cdot)$ stated in Theorem 6.2.

Next, using the arguments in the proof of Proposition 6.4 one can see that the process

$$\sum_{i=1}^{K-1} \frac{N - Q_i^N(\cdot)}{\sqrt{N}} = \hat{Q}_{K-1}^N(\cdot) \xrightarrow{\mathbb{P}} 0,$$

provided $\hat{Q}_{K-1}^N(0) \xrightarrow{\mathbb{P}} 0$. Thus, Theorem 7.1 yields the diffusion limit for the ordinary JSQ policy in the case $B = K + 1$. The proof for $B > K + 1$ then follows from exactly the same arguments as provided in [4, Section 5.2]. The idea is that since the process $Q_{K+1}^N(\cdot)$, when scaled by $\sqrt{N}$, is stochastically bounded, the probability that on any finite time interval, it will take value N (or equivalently, all server pools will have at least $K + 1$ active tasks) vanishes as N grows large. Therefore, the dynamics of the limit of $(\hat{Q}_{K+2}^N(\cdot), \ldots, \hat{Q}_M(\cdot))$ becomes deterministic, and the limit of $\hat{Q}_{K+1}^N(\cdot)$ for $B > K + 1$ becomes a transformation of the limit of $\hat{Q}_{K+1}^N(\cdot)$ for $B = K + 1$, as described in Theorem 2.4. Hence, note that the diffusion limit in Theorem 7.1 is equivalent to the one in Theorem 2.4. In view of the universality result in Corollary 3.4, it thus suffices to prove Theorem 7.1.

We will use the reflection argument developed in [4] to prove Theorem 7.1. Observe that the evolution of $\{(Z_1^N(t), Z_2^N(t))\}_{t \geqslant 0}$ can be described by the following stochastic recursion which is explained in detail below.

$$Z_1^N(t) = Z_1^N(0) + A_1 \left(\int_0^t (KN - Z_1^N(s) - KZ_2^N(s))ds \right) - D_1(\lambda(N)t) + U_1^N(t)$$

$$Z_2^N(t) = Z_2^N(0) + U_1^N(t) - D_2 \left(\int_0^t (K+1)Z_2^N(t)ds \right) - U_2^N(t),$$

(7.2)

where A_1, D_1 and D_2 are unit-rate Poisson processes, and

$$U_1^N(t) = \int_0^t \mathbb{1}_{[Z_1^N(s)=0]} dD_1(\lambda(N)s)$$

$$U_2^N(t) = \int_0^t \mathbb{1}_{[Z_2^N(s)=C\sqrt{N}]} dD_1(\lambda(N)s).$$

(7.3)

The components of Equation (7.2) can be explained as follows. The process $Z_1(t)$ increases by one when a departure occurs from a server pool with at most K active tasks, and it decreases by one when an arriving task is assigned to a server pool with at most K active tasks. Hence the instantaneous rate of increase at time s is given by

$$\sum_{i=1}^{K} i(Q_i^N(t) - Q_{i+1}^N(t)) = \sum_{i=1}^{K} Q_i^N(t) - KQ_{K+1}^N(t)$$

$$= KN - \sum_{i=1}^{K}(N - Q_i^N(t)) - KQ_{K+1}^N(t)$$

$$= KN - Z_1^N(t) - KZ_2^N(t),$$

and the instantaneous rate of decrease is given by the arrival rate $\lambda(N)$. But Z_1^N cannot be negative, and hence the arrivals when Z_1^N is zero, add to Z_2^N, and the rate of increase of the Z_2^N process is given by the overflow process U_1^N. Since $B = K+1$, the rate of decrease of Z_2^N equals the total number of tasks at server pools with exactly $K+1$ tasks, which is given by $(K+1)Z_2^N$. This explains the rate in the Poisson process $D_2(\cdot)$. Finally, since Z_2^N is upper bounded by N, U_2^N is the overflow of the Z_2^N process with $C = \sqrt{N}$, i.e., the number of arrivals to the system when $Z_2^N = N$. The existence and uniqueness of the above stochastic recursion can be proved following the arguments in [19, Section 2].

Martingale representation We now introduce the martingale representation for (7.2), and following similar arguments as in [4, Subsection 4.3], we obtain the following scaled, square integrable martingales with appropriate filtration:

$$M_{1,1}^N(t) = \frac{1}{\sqrt{N}}A_1\left(\int_0^t (KN - Z_1^N(s) - KZ_2^N(s))ds\right) - \frac{1}{\sqrt{N}}\int_0^t (KN - Z_1^N(s) - KZ_2^N(s))ds$$

$$M_{1,2}^N(t) = \frac{1}{\sqrt{N}}(D_1(\lambda(N)t) - \lambda(N)t)$$

$$M_{2,1}^N(t) = \frac{1}{\sqrt{N}}D_2\left(\int_0^t (K+1)Z_2^N(t)ds\right) - \frac{K+1}{\sqrt{N}}\int_0^t Z_2^N(s)ds,$$

(7.4)

with $V_1^N(t) := U_1^N(t)/\sqrt{N}$ and $V_2^N(t) := U_2^N(t)/\sqrt{N}$, and the predictable quadratic variation processes given by

$$\langle M_{1,1}^N\rangle(t) = \frac{1}{N}\int_0^t (KN - Z_1^N(s) - KZ_2^N(s))ds$$

$$\langle M_{1,2}^N\rangle(t) = \frac{\lambda(N)t}{N}$$

(7.5)

$$\langle M_{2,1}^N\rangle(t) = \frac{K+1}{N}\int_0^t Z_2^N(s)ds.$$

Therefore, we have the following martingale representation for (7.2):

$$\zeta_1^N(t) = \zeta_1^N(0) + M_{1,1}^N(t) - M_{1,2}^N(t) - \int_0^t (\zeta_1^N(s) + K\zeta_2^N(s))ds + \frac{t(KN - \lambda(N))}{\sqrt{N}} + V_1^N(t)$$

$$\zeta_2^N(t) = \zeta_2^N(0) + V_1^N(t) - M_{2,1}^N(t) - (K+1)\int_0^t \zeta_2^N(s)ds - V_2^N(t)$$

(7.6)

Convergence of independent martingales We now show the convergence of the martingales defined in (7.4) using the functional central limit theorem.

Lemma 7.3. *As* $N \to \infty$,

$$\left\{ \left(M_{1,1}^N(t), M_{1,2}^N(t), M_{2,1}^N(t) \right) \right\}_{t \geqslant 0} \xrightarrow{\mathcal{L}} \left\{ \left(\sqrt{K}W_1(t), \sqrt{K}W_2(t), 0 \right) \right\}_{t \geqslant 0}$$

in $D_{\mathbb{R}^3}[0,\infty)$, *where* W_1, W_2 *are independent standard Brownian motions.*

Proof. From Theorem 2.1 we know that for any fixed $T \geqslant 0$,

$$\sup_{t \in [0,T]} Z_1^N(t)/N \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \sup_{t \in [0,T]} Z_2^N(t)/N \xrightarrow{\mathbb{P}} 0.$$

This yields the following convergence results:

$$\langle M_{1,1}^N \rangle(T) \xrightarrow{\mathbb{P}} KT$$
$$\langle M_{1,2}^N \rangle(T) \xrightarrow{\mathbb{P}} \lambda T = KT$$
$$\langle M_{2,1}^N \rangle(T) \xrightarrow{\mathbb{P}} 0.$$

(7.7)

Then, using a random time change, the continuous-mapping theorem and functional central limit theorem [19, Theorem 4.2], [4, Lemma 6], we get the convergence of the martingales. $\qquad\square$

Now we use the continuous-mapping theorem to prove the convergence of the processes described in (7.6). To proceed in that direction, we need the following proposition, which is analogous to [4, Lemma 1].

Proposition 7.4. *Let* $B \in \bar{\mathbb{R}}_+$, $b \in \mathbb{R}^2$, $(y_1, y_2) \in D^2[0,\infty)$, *and* $(x_1, x_2) \in D^2[0,\infty)$ *be defined by the following recursion: for* $t \geqslant 0$,

$$x_1(t) = b_1 + y_1(t) + \int_0^t (-x_1(s) - Kx_2(s))ds + u_1(t)$$

$$x_2(t) = b_2 + y_2(t) + (K+1)\int_0^t (-x_2(s))ds + u_1(t) - u_2(t),$$

(7.8)

where u_1 *and* u_2 *are unique non-decreasing functions in* D, *such that*

$$\int_0^\infty \mathbb{1}_{[x_1(s)>0]} du_1(t) = 0$$

$$\int_0^\infty \mathbb{1}_{[x_2(s)<B]} du_2(t) = 0.$$

(7.9)

Then, (x, u) *is the unique solution to the above system. Furthermore, there exist functions* (f, g) : $(\bar{\mathbb{R}}, \mathbb{R}^2, D_{\mathbb{R}}^2[0, \infty)) \to (D_{\mathbb{R}}^2[0, \infty), D_{\mathbb{R}}^2[0, \infty))$ *with* $x = f(B, b, y)$ *and* $u = g(B, b, y)$, *which are continuous when* $\bar{\mathbb{R}}_+$ *is equipped with order topology,* $D_{\mathbb{R}}[0, \infty)$ *is equipped with topology of uniform convergence over compact sets, and* $(\bar{\mathbb{R}}, \mathbb{R}^2, D_{\mathbb{R}}^2[0, \infty))$ *and* $(D_{\mathbb{R}}^2[0, \infty), D_{\mathbb{R}}^2[0, \infty))$ *are equipped with product topology.*

The proof of the above proposition follows from similar arguments as described in the proof of [4, Lemma 1], and hence is omitted.

Proof of Theorem 7.1. Observe that the stochastic recursion equations described by (7.6) fit in the framework of the recursion described by (7.8), by taking $b_i = \zeta_i^N(0)$, $i = 1, 2$, $C = \sqrt{N}$, $y_1(t) = M_{1,1}^N(t) - M_{1,2}^N(t) + t(KN - \lambda(N))/\sqrt{N}$, and $y_2(t) = -M_{2,1}^N(t)$ for the Nth process.

By the assumptions of the theorem we have $\zeta_i^N(0) \xrightarrow{\mathcal{L}} \zeta_i(0)$, for $i = 1, 2$. Also, by Lemma 7.3, $\{(M_{1,1}^N(t), M_{1,2}^N(t), M_{2,1}^N(t))\}_{t \geq 0} \xrightarrow{\mathcal{L}} \{(\sqrt{K}W_1(t), \sqrt{K}W_2(t), 0)\}_{t \geq 0}$. Hence, for the limiting process, $y_1(t) = \sqrt{K}W_1(t) - \sqrt{K}W_2(t) + \beta t \equiv \sqrt{2K}W(t) + \beta t$ and $y_2(t) \equiv 0$. Finally, using the continuous-mapping theorem we get the desired convergence as in the proof of [4, Theorem 2]. □

8 Performance Implications

8.1 Evolution of number of tasks at tagged server pool

We now provide some insights into the steady-state dynamics of the number of tasks at a particular server pool in the regime $d(N) \to \infty$ as $N \to \infty$. Due to exchangeability of the server pools, asymptotically, the dynamics at a particular server pool depends on the system only through the mean-field limit, or the global system state averages. Based on the fixed point (2.2), we claim (without proof) that the steady-state dynamics can be described as follows:

(i) If a server pool contains $\lceil \lambda \rceil$ active tasks, then with high probability no further task will be assigned to it.

(ii) Similarly, if a departure occurs from a server pool having $K = \lfloor \lambda \rfloor$ active tasks, a task will immediately be assigned to it.

(iii) Since the total flow of arrivals that join server pools with exactly K active tasks, are distributed uniformly among all such server pools, each server pool with exactly K active tasks will observe an arrival rate $\lambda p_K(\mathbf{q}^\star)/(q_K^\star - q_{K+1}^\star) = (K+1)f/(1-f)$.

(iv) Finally, the rate of departure from a server pool with $K + 1$ active tasks is given by $K + 1$.

Let $S_k^{d(N)}(t)$ denote the number of tasks at server pool k at time t in the Nth system under the JSQ($d(N)$) scheme. Combining all the above, provided $d(N) \to \infty$ as $N \to \infty$, the process $\{S_k^{d(N)}(t)\}_{t \geq 0}$ converges in distribution to the process $\{S(t)\}_{t \geq 0}$, described as follows:

(i) If $f > 0$, then $\{S(t)\}_{t \geqslant 0}$ is a two-state process, taking values K and $K + 1$, with transition rate from K to $K + 1$ given by $(K + 1)f/(1 - f)$, and from $K + 1$ to K given by $K + 1$. So the steady-state distribution is $\mathbb{P}(S = K) = 1 - f$, and $\mathbb{P}(S = K + 1) = f$, i.e., for $i \geqslant 1$, $\mathbb{P}(S = i) = q_i^\star - q_{i+1}^\star$, which agrees with the fixed point (2.2) of the fluid limit.

(ii) If $f = 0$, then $\{S(t)\}_{t \geqslant 0}$ is a constant process, taking value $\lambda = K$.

8.2 Evolution of number of tasks observed by tagged task

To analyze the performance perceived by a particular tagged task with execution time T, observe that in steady state the probability that it will join a server pool with i active tasks is given by $p_i(\mathbf{q}^\star) = K(1 - f)/\lambda$ for $i = K - 1$, $(K + 1)f/\lambda$ for $i = K$, and 0 otherwise. In the time interval $[0, T]$, the number of active tasks in the server pool it joins, is again a birth-death process $\{\hat{S}(t)\}_{0 \leqslant t \leqslant T}$, whose dynamics is the same as of $\{S(t)\}_{t \geqslant 0}$ process conditioned on having one permanent task (i.e., its departure is not allowed). Therefore, $\{\hat{S}(t)\}_{0 \leqslant t \leqslant T}$ can be described as follows:

(i) If $f > 0$, then $\{\hat{S}(t)\}_{0 \leqslant t \leqslant T}$ is a two-state process, taking values K and $K + 1$, with transition rate from K to $K + 1$ given by $(K + 1)f/(1 - f)$, and from $K + 1$ to K given by K. The steady-state distribution of the process is then given by $\mathbb{P}(\hat{S} = K) = K(1 - f)/\lambda$, and $\mathbb{P}(\hat{S} = K + 1) = (K + 1)f/\lambda$.

(ii) If $f = 0$, then $\{\hat{S}(t)\}_{0 \leqslant t \leqslant T}$ is a constant process, taking value $\lambda = K$.

Observe in both of the above two cases that the initial distribution of $\hat{S}(t)$ coincides with its stationary distribution. Now, if the performance perceived by the tagged task is measured as a function $h : \mathbb{N} \to \mathbb{R}$ of the number of concurrent tasks, then the relevant performance measure is given by

$$\mathbb{E}\left(\frac{1}{T}\int_0^T h(\hat{S}(t))dt\right) = \frac{1}{\lambda}((1 - f)Kh(K) + f(K + 1)h(K + 1)), \tag{8.1}$$

independent of the execution time T. Notice that if $h(x) = 1/(x + 1)$, then the above performance measure becomes the constant $(K(K + 2) - f)/((K + 1)(K + 2))$.

8.3 Loss probabilities

We now examine the asymptotic behavior of the loss probability when the buffer capacity at each server is $B < \infty$ and the arrival rate $\lambda(N)$ satisfies (2.5) with $K = B$. We will establish lower and upper bounds, and prove that these asymptotically coincide. When the buffer capacity B is finite, to characterize the asymptotic steady-state loss probability of the JSQ($d(N)$) scheme, we bound it from below and above by that of an ordinary and a modified Erlang loss system, respectively. The lower and upper bounds rely on a stochastic comparison.

Suppose $Y_1(t)$ and $Y_2(t)$ are two non-explosive, continuous-time Markov processes taking values in a complete separable metric space E. Let $X_1(t)$ and $X_2(t)$ be two birth-death processes defined on the same probability space, with finite state spaces $\{0, 1, \ldots, n_1\}$

and $\{0, 1, \ldots, n_2\}$, whose birth rates are $f_1(X_1(t), Y_1(t))$ and $f_2(X_2(t), Y_2(t))$, and death rates are $g_1(X_1(t), Y_1(t))$ and $g_2(X_2(t), Y_2(t))$, respectively.

Lemma 8.1. *If $n_1 \leqslant n_2$, and for all $x \in \{0, 1, \ldots, n_1\}$, $f_1(x, y_1) \leqslant f_2(x, y_2)$ and $g_1(x, y_1) \geqslant g_2(x, y_2)$, for all $y_1, y_2 \in E$, then $\{X_1(t)\}_{t \geqslant 0} \leqslant_{st} \{X_2(t)\}_{t \geqslant 0}$, provided $X_1(0) \leqslant_{st} X_2(0)$.*

Proof. The proof is fairly straightforward, but we present it briefly for the sake of completeness. First we suitably couple the two processes, and then as before, using the forward induction on event times, we show that the inequality holds throughout the sample path. Define the processes $(X_1(\cdot), X_2(\cdot), Y_1(\cdot), Y_2(\cdot))$ on the same probability space. Due to the assumptions in the theorem, we do not need any condition on the evolution of Y_1 and Y_2, provided that they are defined on the same probability space. Maintain two exponential clocks of rate $M_B := \max\{f_1(x_1, y_1), f_2(x_2, y_2)\}$ (birth-clock) and $M_D := \max\{g_1(x_1, y_1), g_2(x_2, y_2)\}$ (death-clock), respectively. When the birth-clock rings, draw a single uniform$[0, 1]$ random variable u say, and a birth occurs in the X_1 process and X_2 process if $u \leqslant f_1(x_1, y_1)/M_B$ and $u \leqslant f_2(x_2, y_2)/M_B$, respectively. Couple the deaths also, in a similar fashion. Note that the processes thus constructed satisfy the relevant statistical laws in terms of the transition rates $f_1(x_1, y_1)$ and $f_2(x_2, y_2)$.

Now under the above coupling we prove the inequality. Assume that the inequality holds at event time t_0, and $X_1(t_0) = x_1$ and $X_2(t_0) = x_2$. Note that if $x_1 < x_2$, then trivially the inequality holds at the next event time t_1. Therefore, without loss of generality, assume $x_1 = x_2 = x \leqslant n_1$. We will distinguish between two cases depending on whether the birth-clock or death-clock rings at time epoch t_1. In the former case, observe that since $f_1(x, y_1) \leqslant f_2(x, y_2)$ for all $y_1, y_2 \in E$, whenever there is a birth in the X_1 process, there will be a birth in the X_2 process as well. Thus the inequality is preserved. Alternatively, if the death-clock rings at time epoch t_1, then observe that since $g_1(x, y_1) \geqslant g_2(x, y_2)$ for all $y_1, y_2 \in E$, whenever there is a death in the X_2 process, there will be a death in the X_1 process as well, and the inequality is preserved. This completes the proof. $\square$

Denote by $\mathrm{Er}(C, \lambda)$ an Erlang loss system with capacity C, load λ, and exponential service times with unit mean. We further introduce a modified Erlang loss system $\hat{\mathrm{Er}}(n, d)$ with capacity $B(N - n)$, and arrival rate λ, with unit-exponential service times, where a fraction

$$p(n, d) := \left(1 - \frac{n+1}{N}\right)^d,$$

of tasks is rejected upfront, independent of any other processes. Note that the number of active tasks in the $\hat{\mathrm{Er}}(n, d)$ system evolves like an $\mathrm{Er}(B(N - n), \lambda p(n, d))$ system.

Define $C(N) := BN$, $\hat{C}(N) := B(N - n(N))$, and $\hat{\lambda}(N) := \lambda(N)p(n(N), d(N))$. Denote the total number of active tasks at time t in the Nth system following the JSQ($d(N)$) scheme, an $\mathrm{Er}(C(N), \lambda(N))$ system, and an $\hat{\mathrm{Er}}(n(N), d(N))$ system by $Y^{d(N)}(t)$, $Y^N_{\mathrm{Er}}(t)$, and $Y^N_{\hat{\mathrm{Er}}}(t)$, respectively. Denote the associated steady-state loss probabilities by $L^{d(N)}$, $L(C, \lambda)$ and $\hat{L}(n, d)$, respectively.

Lemma 8.2. *For all $N \geqslant 1$, $d(N) \geqslant 1$, and $n(N) < N$,*

(a) $\{Y^N_{\hat{\mathrm{Er}}}(t)\}_{t \geqslant 0} \leqslant_{st} \{Y^{d(N)}(t)\}_{t \geqslant 0} \leqslant_{st} \{Y^N_{\mathrm{Er}}(t)\}_{t \geqslant 0}$,

(b) $L(C(N), \lambda(N)) \leqslant L^{d(N)} \leqslant \hat{L}(n(N), d(N))$.

Proof. (a) For the lower bound, observe that the rate of increase of the process $Y^{d(N)}(\cdot)$ is at most that of the process $Y^N_{Er}(\cdot)$, and the rate of decrease at any state is the same in both processes. Thus, Lemma 8.1 implies that if both systems start from the same occupancy states, then $\{Y^{d(N)}(t)\}_{t\geqslant 0} \leqslant_{st} \{Y^N_{Er}(t)\}_{t\geqslant 0}$. Consequently, in the steady state, $Y^{d(N)}(\infty) \leqslant_{st} Y^N_{Er}(\infty)$, and invoking Little's law yields $L(C(N), \lambda(N)) \leqslant L^{d(N)}$.

For the upper bound, first observe that at any arrival, as long as one of the $n(N)$ lowest-ordered server pools is sampled, which occurs with probability $1 - p(n(N), d(N))$, a task can only get lost when the total number of active tasks is at least $B(N - n(N))$. Thus when the total number of active tasks $Y^{d(N)}(\cdot)$ in the system under the JSQ($d(N)$) scheme is y, the rate of increase of $Y^{d(N)}(t)$ is at least $\lambda(N)(1 - p(n(N), d(N)))$ if $y \leqslant B(N - n(N))$, and the rate of decrease is given by y. Comparing with the modified Erlang loss system $\hat{Er}(n(N), d(N))$ and using Lemma 8.1, we obtain that if $Y^{d(N)}(0) \geqslant_{st} Y^N_{\hat{Er}}(0)$, then

$$\left\{Y^{d(N)}(t)\right\}_{t\geqslant 0} \geqslant_{st} \left\{Y^N_{\hat{Er}}(t)\right\}_{t\geqslant 0}.$$

The proof of the upper bound $L^{d(N)} \leqslant \hat{L}(n(N), d(N))$ is then completed by again invoking Little's law.

(b) Little's law implies

$$L^{d(N)} = 1 - \frac{1}{\lambda(N)} \lim_{T\to\infty} \int_0^T Y^{d(N)}(t)dt,$$

and similarly for the $Er(C(N), \lambda(N))$ and $\hat{Er}(n(N), d(N))$ systems. Statement (b) then follows from statement (a). $\qquad\square$

The proposition below states that the limiting loss probability for the JSQ($d(N)$) scheme vanishes as long as $d(N) \to \infty$.

Proposition 8.3. *For any $\lambda \leqslant B$, if $d(N) \to \infty$ as $N \to \infty$, then $L^{d(N)} \to 0$, as $N \to \infty$.*

Proof. From (4.11) and (4.12), we know if $d(N) \to \infty$, then there exists $n(N)$ such that as $N \to \infty$, $n(N)/N \to 0$ and $p(n(N), d(N)) \to 0$. For such a choice of $n(N)$, $\lambda(N)/C(N) \to \lambda/B \leqslant 1$, and $\hat{\lambda}(N)/\hat{C}(N) \to \lambda/B \leqslant 1$ as $N \to \infty$. Therefore, using Lemma 8.2 and the standard results of the Erlang loss function [9], we complete the proof of the proposition. $\qquad\square$

Remark 8.4. Note that in view of the results in [17, 18] for the JSQ(d) schemes with fixed d, following the arguments as in Remark 3.3, the growth condition $d(N) \to \infty$ as $N \to \infty$ is also necessary to achieve an asymptotically zero probability of loss.

We now further show that the steady-state loss probability multiplied by $\sqrt{N}$ converges to a non-degenerate limit, which is the same as in an $Er(C(N), \lambda(N))$ system. The next theorem also establishes that if (2.5) is satisfied, and $d(N)/(\sqrt{N}\log(N)) \to 0$ as $N \to \infty$, then the steady-state loss probability is of higher order than $1/\sqrt{N}$. This indicates that the growth rate $\sqrt{N}\log(N)$ is not only sufficient but also nearly necessary.

Theorem 8.5 (Scaled loss probability). *Assume that $d(N)/(\sqrt{N}\log(N)) \to \infty$, as $N \to \infty$, and $\lambda(N)$ satisfies (2.5) with $K = B$. Then,*

$$\lim_{N\to\infty} \sqrt{N}\, L^{d(N)} = \frac{\phi(\beta)}{\sqrt{B}\Phi(\beta)}, \tag{8.2}$$

43

where $\phi(\cdot)$ and $\phi(\cdot)$ are the density and distribution function of the standard Normal distribution, respectively.

Since the right side of (8.2) corresponds to the asymptotic steady-state loss probability in an $Er(C(N), \lambda(N))$ system [1, 9, 29], we thus conclude that (8.2) is optimal on $\sqrt{N}$-scale in terms of loss probability.

Proof of Theorem 8.5. The idea again is to suitably bound the steady-state loss probability of the JSQ($d(N)$) scheme. Using Lemma 8.2 and [1, Chapter 7, Theorem 15 (2)], [29], we obtain the lower bound as

$$L^{d(N)} \geqslant L(C(N), \lambda(N))$$
$$\implies \varliminf_{N \to \infty} \sqrt{N} L^{d(N)} \geqslant \varliminf_{N \to \infty} \sqrt{N} L(C(N), \lambda(N)) = \frac{\phi(\beta)}{\sqrt{B} \Phi(\beta)}. \tag{8.3}$$

For the upper bound, from (4.11) and (4.12), we know if $d(N)/(\sqrt{N} \log(N)) \to \infty$ as $N \to \infty$, then there exists $n(N)$ with $n(N)/\sqrt{N} \to 0$ and

$$\sqrt{N} p(n(N), d(N)) \to 0, \quad \text{as} \quad N \to \infty. \tag{8.4}$$

Take such an $n(N)$. Again using [1, Chapter 7, Theorem 15 (2)], we know that since as $N \to \infty$, $\hat{\lambda}(N)/\hat{C}(N)$ converges to one and $\hat{C}(N)/N$ converges to B,

$$\lim_{N \to \infty} \sqrt{N} L(\hat{C}(N), \hat{\lambda}(N)) = \frac{\phi(\beta)}{\sqrt{B} \Phi(\beta)}. \tag{8.5}$$

Therefore, Lemma 8.2, and Equations (8.4), (8.5) yield

$$\varlimsup_{N \to \infty} \sqrt{N} L^{d(N)} \leqslant \varlimsup_{N \to \infty} \sqrt{N} L(C(N), \lambda(N)) + \varlimsup_{N \to \infty} \sqrt{N} p(n(N), d(N)) = \frac{\phi(\beta)}{\sqrt{B} \Phi(\beta)}. \tag{8.6}$$

Combination of the lower bound in (8.3) and the above upper bound completes the proof. $\qquad\square$

Remark 8.6 (Almost necessary condition for growth rate)**.** It is worthwhile to mention that when $\lambda = K > 0$ and $\lambda(N)$ satisfies (2.5), the growth condition $d(N)/(\sqrt{N} \log(N)) \to \infty$, as $N \to \infty$, is nearly necessary in order for the JSQ($d(N)$) scheme to have the same diffusion limit as the ordinary JSQ policy. More precisely, if $d(N)/(\sqrt{N} \log(N)) \to 0$ as $N \to \infty$, then the diffusion limit of the JSQ($d(N)$) scheme differs from the ordinary JSQ policy. In this remark we briefly sketch the outline of the proof. We will assume that the $d(N)$ server pools are chosen with replacement, to avoid cumbersome notation. But the proof technique and the result holds if the server pools are chosen without replacement. Assume on the contrary that as in the ordinary JSQ policy, if $N^{-1/2}(KN - \sum_{i=1}^{K} Q_i^{d(N)}(0))$ is tight, then $N^{-1/2}(KN - \sum_{i=1}^{K} Q_i^{d(N)}(t))$ is a stochastically bounded process. We argue that in this case, for any finite time t, the cumulative number of tasks joining a server with K active tasks (or the cumulative number of lost tasks in case $K = B$) $L^{d(N)}(t)$ does not scale with $\sqrt{N}$, and arrive at a contradiction. Indeed, $\{L^{d(N)}(t)\}_{t \geqslant 0}$ admits the following martingale decomposition:

$$L^{d(N)}(t) = M_L^N(t) + \langle M_L^N \rangle(t), \tag{8.7}$$

where $\{M_L^N(t)\}_{t \geqslant 0}$ is a martingale with compensator and predictable quadratic variation process given by

$$\langle M_L^N \rangle(t) = \lambda(N) \int_0^t \left(Q_K^{d(N)}(s-)/N \right)^{d(N)} ds.$$

Since $\langle M_L^N \rangle(t)/N \leqslant \lambda t$, $\{M_L^N(t)/\sqrt{N}\}_{t \geqslant 0}$ is stochastically bounded. We will show that $\langle M_L^N \rangle(t)$ is stochastically unbounded on $\sqrt{N}$-scale. From (8.7), this will imply that the process $\{L^{d(N)}(t)/\sqrt{N}\}_{t \geqslant 0}$ is stochastically unbounded, which will complete the proof. Note that

$$Q_K^{d(N)}(s) = N - (N - Q_K^{d(N)}(s)) \geqslant N - \sum_{i=1}^K (N - Q_i^{d(N)}(s)),$$

and hence,

$$\langle M_L^N \rangle(t) \geqslant \lambda(N) \int_0^t \left(1 - \frac{1}{N} \sum_{i=1}^K (N - Q_i^{d(N)}(s)) \right)^{d(N)} ds$$

$$\geqslant \lambda(N) t \left(1 - \frac{1}{N} \sup_{s \in [0,t]} \sum_{i=1}^K (N - Q_i^{d(N)}(s)) \right)^{d(N)}.$$

For any $T \geqslant 0$, since $\sup_{t \in [0,T]} \left(KN - \sum_{i=1}^K Q_i^{d(N)}(t) \right)$ is $O_P(\sqrt{N})$, for any function $c(N)$ growing to infinity (to be chosen later), we have with probability tending to 1,

$$\frac{\lambda(N)T}{\sqrt{N}} \left(1 - \frac{1}{N} \sup_{t \in [0,T]} \left(KN - \sum_{i=1}^K Q_i^{d(N)}(t) \right) \right)^{d(N)} \geqslant \frac{\lambda(N)T}{\sqrt{N}} \left(1 - \frac{\sqrt{N}c(N)}{N} \right)^{d(N)}$$

$$\geqslant \frac{\lambda(N)T}{\sqrt{N}} \left(1 - \frac{c(N)}{\sqrt{N}} \right)^{d(N)}.$$

Now since $d(N)/\sqrt{N} \log(N) \to 0$ as $N \to \infty$, define $\omega(N) := \sqrt{N} \log(N)/d(N)$, which goes to infinity as N grows large. Choose $c(N)$ such that $c(N)/\omega(N) \to 0$, as $N \to \infty$. In that case,

$$\frac{\lambda(N)T}{\sqrt{N}} \left(1 - \frac{c(N)}{\sqrt{N}} \right)^{d(N)} = T \exp \left[\log(\sqrt{N} - \beta) + \frac{\sqrt{N} \log(N)}{\omega(N)} \log \left(1 - \frac{c(N)}{\sqrt{N}} \right) \right]$$

$$= T \exp \left[\log(\sqrt{N} - \beta) - \frac{\sqrt{N} \log(N)}{\omega(N)} \frac{c(N)}{\sqrt{N}} \right]$$

$$\to \infty \quad \text{as } N \to \infty.$$

9 Conclusion

In the present paper we have investigated asymptotic optimality properties for JSQ(d) load balancing schemes in large-scale systems. Specifically, we considered a system of N parallel identical server pools and a single dispatcher which assigns arriving tasks to the server with the minimum number of tasks among $d(N)$ randomly selected server pools.

We showed that the fluid limit in a regime where the total arrival rate and number of server pools grow large in proportion coincides with that for the ordinary JSQ policy ($d(N) = N$) as long as $d(N) \to \infty$ as $N \to \infty$, however slowly. We also proved that the diffusion limit in the Halfin-Whitt regime corresponds to that for the ordinary JSQ policy as long as $d(N)$ grows faster than $\sqrt{N} \log(N)$, and that the latter growth rate is in fact nearly necessary. These results indicate that the optimality of the JSQ policy can be preserved at the fluid-level and diffusion-level while reducing the communication overhead by nearly a factor $O(N)$ and $O(\sqrt{N}/\log(N))$, respectively. In future work we plan to further establish convergence rates and extend the results to non-exponential service requirement distributions.

The proofs of the asymptotic optimality properties rely on a novel stochastic coupling construction to bound the difference in the system occupancy processes between the JSQ policy and a JSQ(d) scheme with an arbitrary value of d. It is worth observing that the coupling construction is two-dimensional in nature, and fundamentally different from the classical coupling approach used for deriving stochastic dominance properties for the ordinary JSQ policy and for establishing universality in the single-server case [16]. As it turns out, a direct comparison between the JSQ policy and a JSQ(d) scheme is a significant challenge. Hence, we adopted a two-stage approach based on a novel class of schemes which always assign the incoming task to one of the server pools with the $n(N) + 1$ smallest number of tasks. Just like the JSQ($d(N)$) scheme, these schemes may be thought of as 'sloppy' versions of the JSQ policy. Indeed, the JSQ($d(N)$) scheme is guaranteed to identify the server pool with the minimum number of tasks, but only among a randomly sampled subset of $d(N)$ server pools. In contrast, the schemes in the above class only guarantee that one of the $n(N) + 1$ server pools with the smallest number of tasks is selected, but across the entire system of N server pools. We showed that the system occupancy processes for an intermediate blend of these schemes are simultaneously close on a $g(N)$ scale (e.g. $g(N) = N$ or $g(N) = \sqrt{N}$) to both the JSQ policy and the JSQ($d(N)$) scheme for suitably chosen values of $d(N)$ and $n(N)$ as function of $g(N)$. Based on the latter asymptotic universality, it then sufficed to establish the fluid and diffusion limits for the ordinary JSQ policy.

Acknowledgment

References

[1] Borovkov, A. A. (1976). *Stochastic Processes in Queueing Theory*. Springer New York, New York, NY.

[2] Davis, M. H. A. (1976). The representation of martingales of jump processes. *SIAM Journal on Control and Optimization*, 14(4):623–638.

[3] Ephremides, A., Varaiya, P., and Walrand, J. (1980). A simple dynamic routing problem. *IEEE Transactions on Automatic Control*, 25(4):690–693.

[4] Eschenfeldt, P. and Gamarnik, D. (2015). Join the shortest queue with many servers. The heavy traffic asymptotics. *arXiv:1502.00999*.

[5] Ethier, S. N. and Kurtz, T. G. (2009). *Markov Processes: Characterization and Convergence*. John Wiley & Sons.

[6] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*. Wiley.

[7] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588.

[8] Hunt, P. and Kurtz, T. (1994). Large loss networks. *Stochastic Processes and their Applications*, 53(2):363–378.

[9] Jagerman, D. (1974). Some properties of the Erlang loss function. *The Bell System Technical Journal*, 53(3):525–551.

[10] Johri, P. K. (1989). Optimality of the shortest line discipline with state-dependent service rates. *European Journal of Operational Research*, 41(2):157–161.

[11] Liptser, R. and Shiryaev, A. (1989). *Theory of Martingales*. Springer.

[12] Menich, R. (1987). Optimality of shortest queue routing for dependent service stations. In *Proc. 26th IEEE Conference on Decision and Control*, pages 1069–1072.

[13] Menich, R. and Serfozo, R. F. (1991). Optimality of routing and servicing in dependent parallel processing systems. *Queueing Systems*, 9(4):403–418.

[14] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104.

[15] Mukherjee, D., Borst, S. C., van Leeuwaarden, J. S. H., and Whiting, P. A. (2016a). Universality of load balancing schemes on the diffusion scale. *J. Appl. Probab.*, 53(4).

[16] Mukherjee, D., Borst, S. C., van Leeuwaarden, J. S. H., and Whiting, P. A. (2016b). Universality of power-of-d load balancing in many-server systems.

[17] Mukhopadhyay, A., Karthik, A., Mazumdar, R. R., and Guillemin, F. (2015a). Mean field and propagation of chaos in multi-class heterogeneous loss models. *Performance Evaluation*, 91:117–131.

[18] Mukhopadhyay, A., Mazumdar, R. R., and Guillemin, F. (2015b). The power of randomized routing in heterogeneous loss systems. In *27th International Teletraffic Congress*, pages 125–133.

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

[20] Patel, P., Bansal, D., Yuan, L., Murthy, A., Greenberg, A., Maltz, D. A., Kern, R., Kumar, H., Zikos, M., Wu, H., Kim, C., and Karri, N. (2013). Ananta: cloud scale load balancing. *ACM SIGCOMM Computer Communication Review*, 43(4):207–218.

[21] Robert, P. (2003). *Stochastic Networks and Queues*. Springer Berlin Heidelberg.

[22] Sparaggis, P. D., Towsley, D., and Cassandras, C. G. (1993). Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. *Journal of Applied Probability*, 30(1):223–236.

[23] Sparaggis, P. D., Towsley, D., and Cassandras, C. G. (1994). Sample path criteria for weak majorization. *Advances in Applied Probability*, 26(1):155–171.

[24] Towsley, D. (1995). Application of majorization to control problems in queueing systems. In Chrétienne, P., Coffman, E. G., Lenstra, J. K., and Liu, Z., editors, *Scheduling Theory and its Applications*, chapter 14. John Wiley & Sons, Chichester.

[25] Towsley, D., Sparaggis, P., and Cassandras, C. (1992). Optimal routing and buffer allocation for a class of finite capacity queueing systems. *IEEE Transactions on Automatic Control*, 37(9):1446–1451.

[26] Turner, S. R. (1998). The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences*, 12(01):109.

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

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

[29] Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal*, 63(5):689–708.

[30] Whitt, W. (2002). *Stochastic-Process Limits*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York.

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

[32] Xie, Q., Dong, X., Lu, Y., and Srikant, R. (2015). Power of d choices for large-scale bin packing. In *Proceedings of ACM SIGMETRICS '15*, volume 43, pages 321–334, USA. ACM Press.