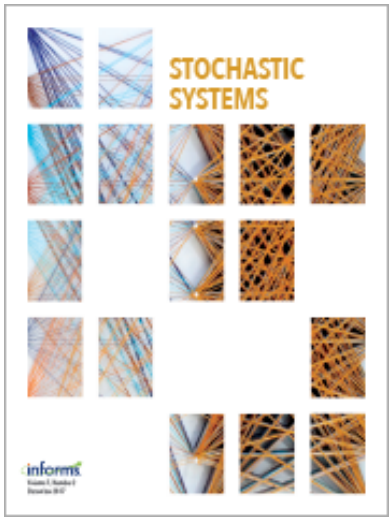


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Utility Maximizing Load Balancing Policies

Diego Goldsztajn, Sem C. Borst, Johan S.H. van Leeuwaarden

To cite this article:
Diego Goldsztajn, Sem C. Borst, Johan S.H. van Leeuwaarden (2023) Utility Maximizing Load Balancing Policies. Stochastic Systems 13(2):211-246. https://doi.org/10.1287/stsy.2022.0103

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 13, No. 2, June 2023, pp. 211–246
ISSN 1946-5238 (online)

Utility Maximizing Load Balancing Policies

Diego Goldsztajn,[a,*] Sem C. Borst,[a] Johan S.H. van Leeuwaarden[b]

[a] Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands; [b] Department of Econometrics and Operations Research, Tilburg University, 5037 AB Tilburg, Netherlands

*Corresponding author

Contact: d.e.goldsztajn@tue.nl, https://orcid.org/0000-0001-7212-0309 (DG); s.c.borst@tue.nl, https://orcid.org/0000-0003-3306-6447 (SCB); j.s.h.vanleeuwaarden@uvt.nl, https://orcid.org/0000-0001-9752-0018 (JSHvL)

Received: December 16, 2021
Revised: June 21, 2022
Accepted: September 19, 2022
Published Online in Articles in Advance: November 21, 2022

https://doi.org/10.1287/stsy.2022.0103

Abstract. Consider a service system where incoming tasks are instantaneously dispatched to one out of many heterogeneous server pools. Associated with each server pool is a concave utility function that depends on the class of the server pool and its current occupancy. We derive an upper bound for the mean normalized aggregate utility in stationarity and introduce two load balancing policies that achieve this upper bound in a large-scale regime. Furthermore, the transient and stationary behavior of these asymptotically optimal load balancing policies is characterized on the scale of the number of server pools in the same large-scale regime.

Funding: This work was supported by the Netherlands Organization for Scientific Research (NWO) through [Gravitation Grant NETWORKS-024.002.003] and [Gravitation Grant Vici 202.068].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsy.2022.0103.

Keywords: load balancing • utility maximization • large-scale asymptotics

1. Introduction

We consider a service system where incoming tasks are instantaneously assigned to one out of many heterogeneous server pools. All the tasks sharing a server pool are executed in parallel and the execution times do not depend on the class of the server pool or the number of tasks currently contending for service. Nevertheless, associated with each server pool is a not necessarily increasing concave utility function that does depend on the class of the server pool and the number of tasks currently sharing it. These features are characteristic of streaming and online gaming services, where the duration of tasks is mainly determined by the application, but still congestion can have a strong impact on the experienced performance (e.g., video resolution and fluency).

The goal is to assign tasks to maximize the overall utility of the system, defined as the aggregate utility of all the server pools normalized by the number of server pools. We derive an upper bound for its stationary mean through an optimization problem where the optimization variable is a sequence that describes the distribution of a fractional number of tasks across the server pools; the objective of the problem is the overall utility function, and the main constraint is that the total number of tasks must be equal to the offered load of the system. We construct an optimal (fractional) task assignment that solves this problem and has a particularly insightful structure, and we formulate the upper bound for the mean stationary overall utility in terms of this solution.

Armed with the previous insight, we propose and analyze two assignment policies that maintain the occupancy state of the system aligned with an optimal task assignment. Specifically, we examine a policy that assigns every new task to a server pool with the largest marginal utility; this policy is dubbed join the largest marginal utility (JLMU). We also introduce a multithreshold policy that follows the same greedy principle but only approximately and uses significantly less state information. The optimal threshold values depend on the typically unknown offered load of the system and are adjusted over time through an inbuilt learning scheme; thus, we name this policy self-learning threshold assignment (SLTA). Assuming exponential service times, we characterize the asymptotic transient and stationary behavior of both policies on the scale of the number of server pools, and we prove that both policies achieve the upper bound for the mean stationary overall utility as the number of server pools grows large.

A fundamental difference between JLMU and SLTA is that the former is naturally agnostic to the offered load, whereas for the latter, the optimal thresholds depend on the offered load. However, we show that the online learning scheme of SLTA is capable of finding the optimal threshold values without any prior knowledge of the offered load, which makes it possible to deploy SLTA if the offered load is not known in advance.

1.1. Main Contributions

The main contribution of this paper is an upper bound for the mean stationary overall utility that is asymptotically tight for exponentially distributed service times and thereby serves as a crucial performance benchmark. The asymptotic tightness of the upper bound is proved by studying the stationary behavior of JLMU and SLTA in the regime where the number of server pools grows large and by establishing that both assignment policies achieve the upper bound in the latter regime.

The analysis of JLMU is based on a fluid limit given by an infinite system of differential equations with a discontinuous right-hand side. We prove that the associated initial value problem always has a unique solution by making a connection with a system of integral equations, expressed in terms of Skorokhod one-dimensional reflection mappings and using a uniqueness result for certain Kolmogorov backward equations. Moreover, we show that the fluid limit holds with respect to an ℓ_1 norm and that the system of differential equations is globally asymptotically stable with respect to this norm. These results are used to prove that the stationary distribution of the process that describes the occupancy state of the system converges in ℓ_1 to an optimal task assignment for the offered load of the system. The asymptotic optimality of JLMU is then established by proving that the stationary overall utilities form a convergent and uniformly integrable sequence of random variables; the proof of the latter properties exploits a representation of the overall utility as a linear functional on ℓ_1 and our convergence results with respect to the ℓ_1 norm.

Although SLTA is simple to implement, its analysis is inherently challenging due to the complex interdependence between two components of the policy. Namely, the dispatching rule, which depends on the multiple thresholds, and the online learning scheme, which adjusts the thresholds over time. Furthermore, an additional technical difficulty is that the learning scheme is triggered by excursions of the occupancy state of the system that asymptotically vanish on the scale of the number of server pools.

To analyze the large-scale transient behavior of SLTA, we use a methodology of Goldsztajn et al. (2021a) that allows overcoming of the aforementioned challenges by means of a nontraditional fluid limit analysis. In this paper, we extend the latter methodology to also prove weak convergence of the stationary distribution of the occupancy process and thresholds. Here our contributions are proofs of ergodicity and tightness of stationary distributions through a careful drift analysis, as well as a suitably adapted interchange of limits argument designed to leverage the large-scale transient result obtained with the methodology of Goldsztajn et al. (2021a). Equipped with the convergence results for the stationary distributions, we prove the asymptotic optimality of SLTA in a similar way as for JLMU by showing that all our limit theorems hold with respect to the ℓ_1 norm and exploiting the linear representation of the overall utility function.

1.2. Related Work

Load balancing and task assignment in parallel-server systems has received immense attention in the past decades; some relevant papers are Winston (1977), Vvedenskaya et al. (1996), Mitzenmacher (2001), Lu et al. (2011), Stolyar (2015), and Eschenfeldt and Gamarnik (2018). Traditionally the focus used to be on performance, but more recently the implementation overhead has emerged as an equally important issue. In large-scale deployments, this overhead has two main sources: the communication burden of messaging between the dispatcher and the servers and the operational cost of storing and managing state information at the dispatcher (Gamarnik et al. 2018, 2020). We refer to Van der Boor et al. (2022) for an extensive survey on scalable load balancing.

Whereas the load balancing literature has been predominantly concerned with systems of parallel single-server queues, the present paper considers an infinite-server setting where the service times of tasks do not depend on the number of competing tasks. As noted earlier, this feature is characteristic of streaming applications, where the level of congestion does not significantly affect the duration of tasks. The level of congestion has, however, a strong impact on the amount of resources received by individual streaming sessions, and thus on the experienced quality-of-service, which can be modeled through utility functions. Infinite-server dynamics have been commonly adopted as a natural paradigm for modeling streaming sessions on flow-level (Benameur et al. 2002, Key et al. 2004), and the problem of managing large data centers serving streaming sessions has been recently addressed in Mukherjee et al. (2020). Systems with infinite-server dynamics have also been analyzed in Mukhopadhyay et al. (2015a, b), Karthik et al. (2017), and Xie et al. (2015), which concern loss models that are different in nature from the setting considered in the present paper.

When the server pools are homogeneous, the overall utility is a Schur-concave function of the vector describing the number of tasks at each server pool. In this case, maximizing the aggregate utility of the system boils down to equalizing the number of tasks across the various server pools. Join the shortest queue (JSQ) maximizes the mean stationary overall utility of the system for exponential service times and, in fact, has stronger stochastic

optimality properties (Menich and Serfozo 1991, Sparaggis et al. 1993). In the homogeneous setting, JLMU reduces to JSQ and is thus optimal for exponential service times. In addition, SLTA reduces to the policy considered in Goldsztajn et al. (2021a, b), which asymptotically matches the performance of JSQ on the fluid and diffusion scales for exponentially distributed service times. Although the policy considered in Zhou et al. (2017, 2018) and Horváth et al. (2019) is similar to SLTA in name, this policy does not equalize the queue lengths.

The problem of maximizing the overall utility of the system is more challenging if the server pools are heterogeneous as in this paper. Heterogeneity is the norm in data centers, where servers from different generations coexist because old machines are only gradually replaced by more powerful versions; as shown in Figure 1, this feature has been recently addressed in the load balancing literature for single-server models (Gardner and Stephens 2019, Jaleel et al. 2020, Gardner et al. 2021) but not in the infinite-server context. When the server pools are heterogeneous, it is no longer optimal to maintain an evenly balanced distribution of the load; in fact, it is not even obvious at all how tasks should be distributed to maximize the overall utility function, and the optimal distribution of tasks across the server pools depends on this function. Another striking difference with the homogeneous setting is that JLMU is generally not optimal in the prelimit for exponentially distributed service times; we establish that, in general, the optimality is only achieved asymptotically in the heterogeneous case.

From a theoretical perspective, one of the most interesting features of SLTA is its capacity to learn the offered load of the system. The problem of adaptation to unknown demands was previously addressed in Mukherjee et al. (2017) and Goldsztajn et al. (2018, 2022) in the context of single-server models by assuming that the number of servers can be right-sized on the fly to match the load of the system. However, in the latter papers the dispatching rule remains the same at all times because the right-sizing mechanism alone is sufficient to maintain small queues, by adjusting the number of servers. Different from these right-sizing mechanisms, the learning scheme of SLTA modifies the parameters of the dispatching rule over time to maximize the overall utility of the system.

1.3. Outline of the Paper

In Section 2, we introduce some of the notation used throughout the paper and we prove the upper bound for the mean stationary overall utility. In Section 3, we specify the JLMU and SLTA policies, and we state their asymptotic optimality with respect to the mean stationary overall utility. In Section 4, we present several results that pertain to the asymptotic transient behavior of these two policies and that are used to establish their asymptotic optimality. To characterize the asymptotic behavior of JLMU and SLTA, we construct systems of different sizes on a common probability space in Section 5, where we also prove relative compactness results. Limit theorems for the transient regimes of JLMU and SLTA are proved in Sections 6 and 7, respectively, and the asymptotic optimality of these policies is established in Section 8. Some proofs are deferred to Online Appendices A, B and C.

2. Performance Upper Bound

In this section, we define some of the notation used throughout the paper and we prove the upper bound for the mean stationary overall utility. In Section 2.1, we introduce two descriptors for specifying the state of the system, and we define the overall utility function. In Section 2.2, we present the optimization problem used to derive the upper bound for the mean stationary overall utility. In Section 2.3, we construct a solution of this problem explicitly, and in Section 2.4, we use the constructed solution to formulate the upper bound for the mean stationary overall utility, which we prove in Section 2.5.

Figure 1. Schematic View of Some of the Related Work

	Homogenenous setting	**Heterogeneous setting**
Single-server dynamics	References in Van der Boor et al. (2018)	Jaleel et al. (2020) Gardner et al. (2021)
Infinite-server dynamics	Mukherjee et al. (2020) Goldsztajn et al. (2021a,b)	Present paper

Notes. Most of the load balancing literature concerns systems of parallel and homogeneous single-server queues; this vast literature is surveyed in Van der Boor et al. (2022). Some recent papers study single-server dynamics in heterogeneous settings or infinite-server dynamics in homogeneous settings, whereas the present paper considers a heterogeneous system with infinite-server dynamics.

2.1. Basic Notation

Consider a system with m classes of server pools. All the tasks sharing a server pool are executed in parallel and the execution times do not depend on the class of the server pool or the number of tasks currently contending for service. Nevertheless, associated with each server pool is a concave utility function which does depend on the class of the server pool and the number of tasks sharing it. For example, these functions can be used to model the overall quality-of-service provided to streaming tasks sharing an underlying resource with a fixed capacity. The objective is to assign the incoming tasks to the various server pools to maximize the aggregate utility of all the server pools in stationarity.

The number of server pools is denoted by n, and the number and fraction of server pools of class i are denoted by $A_n(i)$ and $\alpha_n(i) = A_n(i)/n$, respectively. We assume that tasks arrive as a Poisson process of intensity $n\lambda$ with independent and identically distributed service times of mean $1/\mu$, and we define $\boldsymbol{X}_n(i,k)$ as the number of tasks in server pool k of class i; bold symbols are used in the paper to indicate time dependence. Server pools of the same class that have the same number of tasks are exchangeable; thus, we usually consider a different state descriptor. Specifically, we let

$$\boldsymbol{q}_n(i,j) := \frac{1}{n}\sum_{k=1}^{A_n(i)} \mathbb{1}_{\{\boldsymbol{X}_n(i,k)\geq j\}}$$

denote the fraction of server pools which are of class i and have at least j tasks. The values of $\boldsymbol{X}_n$ and $\boldsymbol{q}_n$ at a given time are referred as the occupancy state or task assignment.

The concave utility function associated with server pools of class i is denoted by u_i, and the overall utility of the system is defined as the aggregate utility of all the server pools normalized by the number of server pools. More precisely, we let

$$u_n(\boldsymbol{X}_n) := \frac{1}{n}\sum_{i=1}^{m}\sum_{k=1}^{A_n(i)} u_i(\boldsymbol{X}_n(i,k)).$$

Note that $\boldsymbol{q}_n(i,j) - \boldsymbol{q}_n(i,j+1)$ is the fraction of server pools of class i with j tasks. Thus, the overall utility may equivalently be expressed as

$$u(\boldsymbol{q}_n) := \sum_{i=1}^{m}\sum_{j=0}^{\infty} u_i(j)\left[\boldsymbol{q}_n(i,j) - \boldsymbol{q}_n(i,j+1)\right].$$

Although the overall utility function is generally not linear as a function of $\boldsymbol{X}_n$, it is always linear as a function of $\boldsymbol{q}_n$, as shown by the previous expression.

The total number of tasks in the system, normalized by the number of server pools, can be expressed in terms of the occupancy state $\boldsymbol{q}_n$ as follows:

$$\boldsymbol{s}_n := \sum_{i=1}^{m}\sum_{j=1}^{\infty} \boldsymbol{q}_n(i,j) = \sum_{i=1}^{m}\sum_{j=1}^{\infty} j\left[\boldsymbol{q}_n(i,j) - \boldsymbol{q}_n(i,j+1)\right]. \tag{1}$$

The quantity $j\left[\boldsymbol{q}_n(i,j) - \boldsymbol{q}_n(i,j+1)\right]$ represents the number of tasks in server pools of class i with exactly j tasks, normalized by the total number of server pools. Hence, $\boldsymbol{s}_n$ indeed corresponds to the normalized total number of tasks.

Throughout the paper, we write P and E to denote the probability and expectation with respect to a given probability measure. If $\rho := \lambda/\mu$ denotes the normalized offered load, then the stationary distribution of the total number of tasks is Poisson with mean $n\rho$ due to the infinite-server dynamics of the system. Thus, $E[\boldsymbol{s}_n] = \rho$ in stationarity, for any task assignment policy.

2.2. Optimization Problem

Based on the previous statements, we now formulate an optimization problem that yields an upper bound for the mean stationary overall utility:

$$\begin{aligned} \underset{q}{\text{maximize}} \quad & u(q) \\ \text{subject to} \quad & \sum_{i=1}^{m}\sum_{j=1}^{\infty} q(i,j) = \rho, \\ & 0 \leq q(i,j+1) \leq q(i,j) \leq q(i,0) = \alpha_n(i) \quad \text{for all} \quad i,j. \end{aligned} \tag{2}$$

To see that the optimum of (2) yields an upper bound for the mean stationary overall utility, consider any policy such that $\boldsymbol{q}_n$ has a stationary distribution. We assume that the policy is such that the evolution of the system over time can be described by a Markov process, with a possibly uncountable state space, that has a stationary distribution. Let q_n be a random variable with the stationary distribution of $\boldsymbol{q}_n$ and define $E[q_n]$ as the sequence whose (i, j) element is $E[q_n(i,j)]$. Observe that

$$E[u(q_n)] = u(E[q_n]). \tag{3}$$

Indeed, the utility functions are concave, so for each i there exists $j_i \in \mathbb{N}$ such that $u_i(j)$ and $u_i(k)$ have the same sign if $j,k > j_i$. Therefore, (3) follows from Tonelli's theorem. In addition, $E[q_n]$ satisfies the constraints of (2) because the total number of tasks in stationarity has mean $n\rho$. Thus, $E[u(q_n)]$ is upper bounded by the optimum of (2).

2.3. Structure of an Optimal Solution

For brevity, we refer to an optimizer of (2) as an optimal task assignment; the term optimal fractional task assignment would be more appropriate since the offered load $n\rho$ may not be integral. In this section we define a ranking of the server pools that can be used to construct an optimal task assignment. For this purpose, consider the sets

$$\mathcal{I} := \{1,\dots,m\} \times \mathbb{N} \quad \text{and} \quad \mathcal{I}_+ := \{(i,j) \in \mathcal{I} : j \geq 1\}.$$

A server pool has coordinates $(i,j) \in \mathcal{I}_+$ if its class is i and it has precisely $j-1$ tasks; for example, in Figure 2, server pool A of class 1 has coordinates (1, 4), and both server pools of class 3 have coordinates (3, 1). Because server pools with the same coordinates are statistically identical, we may focus on ranking coordinates rather than server pools.

Formally, we define a total order on $\mathcal{I}_+$ that gives precedence to coordinates associated with larger marginal utilities. The marginal utility of a server pool of class i with j tasks is denoted by $\Delta(i,j) := u_i(j+1) - u_i(j)$ and represents the change in the utility function of such a server pool if it receives an additional task. The marginal utility of the coordinates (i, j) is just $\Delta(i, j-1)$, the marginal utility of server pools of class i with $j-1$ tasks.

Consider the dictionary order $\prec$ on $\mathcal{I}$, defined by

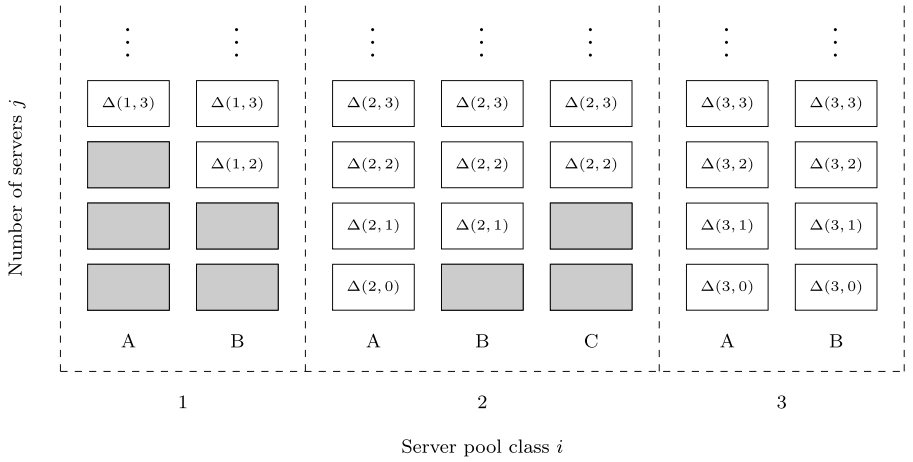
$$(i_1,j_1) \prec (i_2,j_2) \quad \text{if and only if} \quad i_1 < i_2 \quad \text{or} \quad i_1 = i_2 \quad \text{and} \quad j_1 < j_2.$$

We obtain a total order $\lhd$ on $\mathcal{I}_+$ by writing $(i_1,j_1) \lhd (i_2,j_2)$ if and only if one of the following conditions holds:

$$\Delta(i_1,j_1-1) < \Delta(i_2,j_2-1), \text{ or}$$

$$\Delta(i_1,j_1-1) = \Delta(i_2,j_2-1) \text{ and } (i_1,j_1-1) \succ (i_2,j_2-1).$$

Figure 2. Schematic Representation of the Marginal Utilities



Notes. White rectangular slots and gray rectangles represent idle and busy servers, respectively. Each of the columns labeled with letters represents a server pool and the dashed lines enclose server pools of the same class. If the tasks sent to a given server pool are always placed in the first idle server from bottom to top, then the marginal utilities written on top of the idle servers indicate the increase in the aggregate utility of the system when the server receives a task.

In particular, the marginal utility of server pools with coordinates (i_1, j_1) is smaller than or equal to that of server pools with coordinates $(i_2, j_2) \triangleright (i_1, j_1)$. The dictionary order is used to break the tie when both coordinates are associated with the same marginal utility, but a different tie breaking rule could be used instead.

Consider the task assignment q_n^* defined by

$$q_n^*(i,j) := \begin{cases} 0 & \text{if} \quad (i,j) \triangleleft \sigma_n^*, \\ \alpha_n(i) & \text{if} \quad (i,j) \triangleright \sigma_n^*, \\ \rho - \sum_{(r,s) \triangleright \sigma_n^*} \alpha_n(r) & \text{if} \quad (i,j) = \sigma_n^*, \end{cases} \qquad \text{for all} \quad (i,j) \in \mathcal{I}_+. \tag{4}$$

In Section 2.5, we prove that q_n^* constitutes an optimal task assignment if σ_n^* is defined as the unique element of $\mathcal{I}_+$ that satisfies

$$\sum_{(i,j) \triangleright \sigma_n^*} \alpha_n(i) \le \rho < \sum_{(i,j) \trianglerighteq \sigma_n^*} \alpha_n(i). \tag{5}$$

For the uniqueness of σ_n^*, the number of terms in the summations on both sides of (4) increases as the ranking of σ_n^* becomes worse and that the summation on the right has exactly one more term than the summation on the left.

2.3.1. Numerical Examples. Figure 3 depicts the optimal task assignments obtained through (4) for two sets of utility functions and different values of ρ. If n is such that $n\alpha_n(i)$ and $n\rho$ are integers for all i, then the plots can be interpreted as sets of n adjacent columns, where each column represents a server pool and the colored portion of a column indicates the number of tasks sharing the server pool, as in the diagram of Figure 2. The thick vertical lines separate the server pool classes and the quantities $q_n^*(i,j)$ can be read off by rotating the plots.

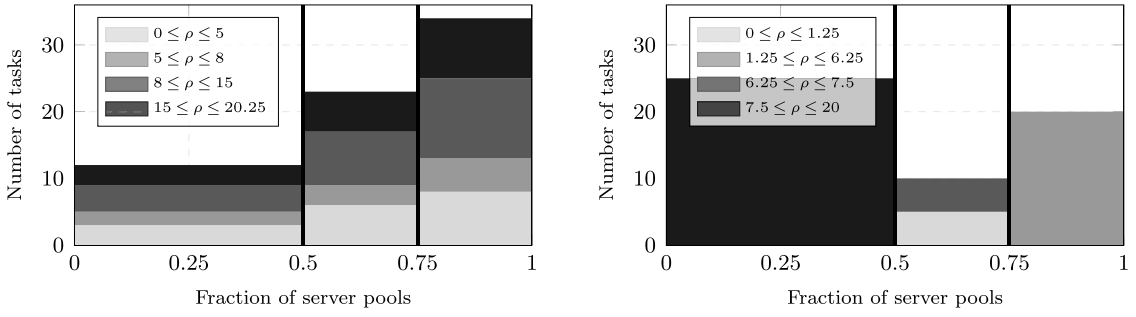
The left plot corresponds to utility functions of the form $u_i(x) = xg(r(i)/x)$, with g a concave and increasing function. These utility functions can be used to model the aggregate quality-of-service provided to streaming tasks sharing a single server pool. The quantity $r(i)$ represents the total amount of resources in a server pool of class i and $g(r(i)/x)$ models the quality-of-service provided to a single task when the server pool is shared by x tasks and each task gets a fraction $r(i)/x$ of the total resources. For a given ρ, the left plot of Figure 3 depicts server pools with roughly $r(i)/c(r,\rho)$ tasks, with $c(r,\rho)$ a normalizing constant that does not depend on g; for some values of ρ, all server pools have exactly $r(i)/c(r,\rho)$ tasks, but in other cases some of these numbers are rounded. This behavior is explained by noting that the derivative of $u_i(x)$ can be expressed as a function of $r(i)/x$, thus the occupancy levels $r(i)/c(r,\rho)$ equalize the marginal utilities.

In the left plot of Figure 3, the occupancy levels of the various server pools maintain approximately fixed ratios as ρ increases. The right plot shows a completely different behavior: as ρ increases from 0 to 1.25, only the occupancy of server pools of class 2 grows, but from 1.25 to 6.25, only the occupancy of server pools of class 3 grows, and eventually exceeds the occupancy of server pools of class 2. Furthermore, as ρ increases beyond 7.5, only the occupancy of server pools of class 1 increases.

2.4. Performance Upper Bound

We now state the upper bound for the mean stationary overall utility. The proof uses the remarks of Sections 2.2 and 2.3 and is given in Section 2.5.

Figure 3. Distribution of the Offered Load Across the Various Server Pools Under the Optimal Task Assignment of (4)



Notes. (Left) $u_i(x) = x \log(r(i)/x)$ with $r = (5, 10, 15)$. (Right) $u_1(x) = x$, $u_2(x) = 2x - x^2/20$, $u_3(x) = 3x/2$ if $x < 20$ and $u_3(x) = 30$ if $x \ge 20$. In both cases $m = 3$ and $\alpha_n = (1/2, 1/4, 1/4)$.

Theorem 1. *Consider any task assignment policy such that the occupancy process $\boldsymbol{q}_n$ has a stationary distribution and let q_n be a random variable distributed as this stationary distribution. Then*

$$E\big[u(q_n)\big] \le u\big(q_n^*\big).$$

In Section 3.4, we establish that the upper bound is asymptotically achievable when service times are exponentially distributed. In particular, we will see that JLMU achieves the upper bound of Theorem 1 as the number of server pools grows large; recall that JLMU is generally not optimal in the prelimit, not even for exponential service times. Moreover, we will establish that SLTA also achieves the upper bound asymptotically while relying on considerably less state information.

2.5. Proof of the Upper Bound

Recall from Section 2.2 that the optimum of (2) is an upper bound for the mean stationary overall utility. Therefore, we only need to prove that the task assignment q_n^* defined in (4) is an optimizer of (2). For this purpose, we introduce the following definition. We say that a sequence $q \in \mathbb{R}^{\mathcal{I}}$ is eventually zero if there exists $k > 0$ such that $q(i,j) = 0$ for all i and $j > k$. The following lemma implies that q_n^* is an optimizer of (2) if we impose the additional constraint that the solution must be eventually zero.

Lemma 1. *If q satisfies the constraints of (2) and is eventually zero, then $u(q) \le u(q_n^*)$.*

Proof. Because q is eventually zero, it is possible to write

$$\begin{aligned} u(q) &= \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} u_i(j)\big[q(i,j) - q(i,j+1)\big] \\ &= \sum_{i=1}^{m} u_i(0)q(i,0) + \sum_{i=1}^{m}\sum_{j=1}^{\infty} \Delta(i,j-1)q(i,j) \\ &= \sum_{i=1}^{m} u_i(0)\alpha_n(i) + \sum_{(i,j)\in\mathcal{I}_+} \Delta(i,j-1)q(i,j). \end{aligned} \tag{6}$$

In the last expression, the terms of the summation are ordered with respect to $\triangleleft$ and in particular in nonincreasing order of the marginal utilities $\Delta(i,j-1)$. The task assignment q_n^* is obtained by choosing the coefficients $q(i,j)$ so that the first coefficients are maximal while all the coefficients add up to ρ. Thus, q_n^* maximizes the right-hand side of (6). □

We now provide a solution of (2), without imposing any additional constraints.

Proposition 1. *The task assignment q_n^* defined in (4) is an optimizer of (2).*

Proof. By Lemma 1, it suffices to prove that $u(q) \le u(q_n^*)$ for each q that is not eventually zero and satisfies the constraints of (2). Next, we fix one such sequence q and we construct an eventually zero sequence z such that $u(q) \le u(z)$ and z satisfies the constraints of (2). Then $u(q) \le u(z) \le u(q_n^*)$ by Lemma 1, as desired.

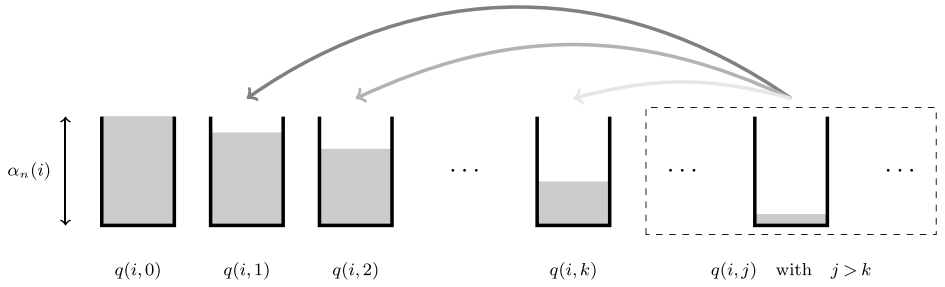
Choose $k \in \mathbb{N}$ such that $\alpha_n(i)k > \rho$ for all i. For each i, we define $z(i,j)$ iteratively, by

$$z(i,j) := \begin{cases} \min\left\{ q(i,j) + \sum_{l=k+1}^{\infty} q(i,j) - \sum_{l=0}^{j-1}\big[z(i,l) - q(i,l)\big], \alpha_n(i) \right\} & \text{if} \quad j \le k, \\ 0 & \text{if} \quad j > k. \end{cases}$$

Informally, each coefficient $q(i,j)$ can be regarded as a container with capacity $\alpha_n(i)$, as shown in Figure 4. For each i, the sequence $q(i,\cdot)$ is transformed into $z(i,\cdot)$ in two steps: first we remove all the mass from the coefficients $q(i,j)$ with $j > k$, and then we place this mass on the coefficients $q(i,j)$ with $j \le k$. In the latter step, we start with the first coefficient, placing as much mass as possible without exceeding the capacity $\alpha_n(i)$. The remainder of mass is placed in the following coefficients in the same fashion, and in increasing order of j. Observe that all the mass will have been placed right after the coefficient $q(i,k)$ is done because q satisfies the constraints of (2) and $\alpha_n(i)k > \rho$. Furthermore, the following property holds:

$$\sum_{j=0}^{k}\big[z(i,j) - q(i,j)\big] = \sum_{j=k+1}^{\infty} q(i,j) \quad \text{for all} \quad i. \tag{7}$$

Figure 4. Schematic View of the Construction of $z(i,\cdot)$ from $q(i,\cdot)$ for Some Fixed i



The overall utility of the task assignment q satisfies

$$\begin{aligned}
u(q) &= \sum_{i=1}^{m}\sum_{j=0}^{\infty} u_i(j)\big[q(i,j)-q(i,j+1)\big] \\
&= \sum_{i=1}^{m}\sum_{j=0}^{k} u_i(j)\big[q(i,j)-q(i,j+1)\big] + \sum_{i=1}^{m}\sum_{j=k+1}^{\infty} u_i(j)\big[q(i,j)-q(i,j+1)\big] \\
&\le \sum_{i=1}^{m}\sum_{j=0}^{k} u_i(j)\big[q(i,j)-q(i,j+1)\big] \\
&+ \sum_{i=1}^{m}\sum_{j=k+1}^{\infty} \big[u_i(k)+(j-k)\Delta(i,k)\big]\big[q(i,j)-q(i,j+1)\big] \\
&= \sum_{i=1}^{m}\sum_{j=0}^{k} u_i(j)\big[q(i,j)-q(i,j+1)\big] + \sum_{i=1}^{m} u_i(k)q(i,k+1) + \sum_{i=1}^{m}\Delta(i,k)\sum_{j=k+1}^{\infty} q(i,j) \\
&= \sum_{i=1}^{m} u_i(0)\alpha_n(i) + \sum_{i=1}^{m}\sum_{j=0}^{k}\Delta(i,j-1)q(i,j) + \sum_{i=1}^{m}\Delta(i,k)\sum_{j=k+1}^{\infty} q(i,j) \\
&= u(z) + \sum_{i=1}^{m}\sum_{j=0}^{k}\Delta(i,j-1)\big[q(i,j)-z(i,j)\big] + \sum_{i=1}^{m}\Delta(i,k)\sum_{j=k+1}^{\infty} q(i,j).
\end{aligned}$$

For the last step, recall that z is eventually zero, so $u(z)$ can be computed as in (6). Note that $q(i,j)\le z(i,j)$ and $\Delta(i,j-1)\ge\Delta(i,k)$ if $j\le k$. Therefore,

$$u(q)\le u(z)+\sum_{i=1}^{m}\Delta(i,k)\sum_{j=1}^{k}\big[q(i,j)-z(i,j)\big]+\sum_{i=1}^{m}\Delta(i,k)\sum_{j=k+1}^{\infty} q(i,j)=u(z)\le u(q_n^*).$$

The middle equality and last inequality follow from (7) and Lemma 1, respectively. □

The proof of Theorem 1 follows easily from Proposition 1.

Proof of Theorem 1. As indicated at the end of Section 2.2, the optimum of (2) upper bounds the mean stationary overall utility $E\big[u(q_n)\big]$. Thus, it follows from Proposition 1 that $u\big(q_n^*\big)$ upper bounds the mean stationary overall utility. □

3. Load Balancing Policies

In this section we describe the load balancing policies considered in the paper, and we state their asymptotic optimality with respect to the mean stationary overall utility when service times are exponential. In Sections 3.1 and 3.2, we specify JLMU and SLTA, respectively. In Section 3.3, we define stochastic models, based on continuous-time Markov chains, for the analysis of both policies. In Section 3.4, we state the asymptotic optimality result.

Before proceeding, it is illustrative to draw an analogy between the setting considered in this paper and the load balancing literature for systems of parallel single-server queues, where the primary objective is to minimize queueing delay. The natural policy for the setting considered in this paper is JLMU, while the natural policy for

minimizing queueing delay in systems of parallel single-server queues is JSQ. The deployment of these policies involves a considerable communication overhead, or storing and managing a significant amount of state information. In the setting considered in this paper, SLTA provides a asymptotically optimal performance for exponential service times and uses substantially less state information than JLMU. From this perspective, SLTA is the counterpart of JIQ in the load balancing literature for systems of parallel single-server queues (Lu et al. 2011, Stolyar 2015).

3.1. Join the Largest Marginal Utility (JLMU)

JLMU assigns every new task to a server pool that currently has the best ranked coordinates, thus also the largest marginal utility. Formally, define

$$\sigma(q) = \big(\sigma_i(q), \sigma_j(q)\big) := \max\big\{(i,j) \in \mathcal{I}_+ : q(i,j-1) > q(i,j)\big\} \tag{8}$$

for each occupancy state q. The maximum is taken with respect to $\lhd$, and the condition $q(i,j-1) > q(i,j)$ implies that some server pool of class i has precisely $j-1$ tasks. If q_n is the occupancy state right before a task arrives, then JLMU assigns the task to a server pool of class $\sigma_i(q_n)$ with exactly $\sigma_j(q_n) - 1$ tasks.

The coordinates obtained through (8) correspond to server pools with the largest marginal utility by definition of $\lhd$. In addition, observe that the dictionary order is used to break ties between coordinates associated with the same marginal utility. If two server pools have the same coordinates, then it does not matter which of them is assigned the new task since they are statistically identical. For definiteness, we postulate that the tie is broken uniformly at random.

If all the server pools have the same utility function, then JLMU reduces to JSQ, and the overall utility is a Schur-concave function of $\boldsymbol{X}_n$. If in addition the service times are exponential, then the stochastic optimality properties proved in Sparaggis et al. (1993) and Menich and Serfozo (1991) for JSQ imply that JLMU maximizes the mean stationary overall utility in this homogeneous setting. It might be natural to expect that the optimality with respect to the mean stationary overall utility extends to the heterogeneous setting. We refute this, however, in Section 8.3, where we construct a heterogeneous system for which JLMU is strictly suboptimal. The constructed example also hints at the underlying reasons for the suboptimality in the heterogeneous case. Essentially, instead of always assigning incoming tasks greedily, such that the increase in the overall utility is maximal, it is sometimes advantageous to dispatch the new tasks conservatively, to hedge against pronounced drops of the overall utility that may be caused by a quick succession of departures. The right balance between greedy and conservative actions depends intricately on the utility functions, but we prove that JLMU is always asymptotically optimal for exponential service times, regardless of the specific set of utility functions; this result is stated formally in Section 3.4.

3.2. Self-Learning Threshold Assignment (SLTA)

JLMU relies on complete information about the number of tasks per server pool, which could be impractical in large-scale deployments. In contrast, SLTA only requires to store at most two bits per server pool, which is considerably less state information. To specify this policy, we need to describe its two components. Namely, the dispatching rule, for assigning the incoming tasks to the server pools, and the learning scheme, for dynamically adjusting a set of thresholds that the dispatching rule uses.

Consider the unique enumeration $\{(i_k, j_k) : k \geq 1\}$ of $\mathcal{I}_+$ such that $(i_k, j_k) \triangleright (i_{k+1}, j_{k+1})$ for all k. Given $r \geq 1$, we define a set of thresholds by

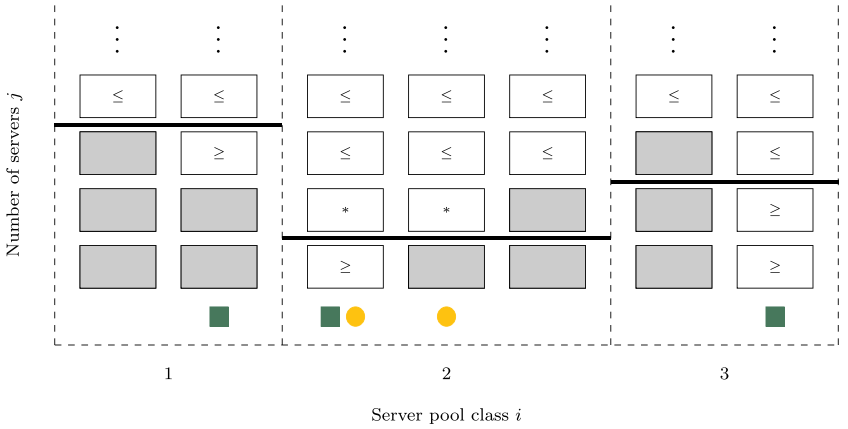
$$\ell_i(r) := \max\big\{j \geq 0 : (i,j) \triangleright (i_r, j_r) \text{ or } j = 0\big\} \quad \text{for all} \quad i.$$

Recall the optimal task assignment defined at the end of Section 2.3. The learning scheme keeps an estimate (i_{r_n}, j_{r_n}) of the coordinates σ^*_n, which depend on the typically unknown offered load of the system. The index $\boldsymbol{r}_n$ determines this estimate and is used to compute thresholds from the previous expression, which are in turn used to assign tokens to the server pools. Specifically, a server pool of class i with exactly $j-1$ tasks has

- A green token if $j - 1 < \ell_i(\boldsymbol{r}_n)$,
- A yellow token if $i = i_{\boldsymbol{r}_n}$ and $j - 1 \leq \ell_i(\boldsymbol{r}_n)$.

The first condition is equivalent to $\boldsymbol{r}_n > 1$ and $(i,j) \triangleright (i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$, and the second condition is equivalent to $i = i_{\boldsymbol{r}_n}$ and $(i,j) \trianglerighteq (i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$. Also, a server pool can have both a green and a yellow token at the same time. As indicated in Figure 5, a larger increase in the overall utility is obtained by dispatching tasks to server pools with green tokens first, then to server pools with yellow tokens and only afterward to server pools without tokens.

Figure 5. Schematic Representation of the Thresholds and Tokens Used by SLTA for $(i_{r_n}, j_{r_n}) = (2,2)$



Notes. The thresholds are indicated by thick horizontal lines that cross the server pools, and the tokens are represented using squares and circles underneath the server pools; a square corresponds to a green token and a circle corresponds to a yellow token. Assuming that the rectangular slots within a server pool are always filled from bottom to top, the slots marked with an $*$ provide a marginal utility of $\Delta(i_{r_n}, j_{r_n} - 1)$. The symbols $\leq$ and $\geq$ indicate how the marginal utility of the other slots compares to the latter value.

The tokens are used by the dispatching rule. Specifically, when a task arrives, it is assigned to a server pool according to the following criteria.

- In the presence of green tokens of class $i \neq i_{r_n - 1}$, the dispatcher picks one of these green tokens uniformly at random, and if only green tokens of class $i = i_{r_n - 1}$ remain, then one of these is picked. Then the task is sent to the corresponding server pool.
- In the presence of only yellow tokens, the dispatcher picks a yellow token uniformly at random and sends the task to the corresponding server pool.
- Otherwise, the task is sent to a server pool chosen uniformly at random.

If (i_{r_n}, j_{r_n}) are the coordinates σ_n^* defined in (5), then this dispatching rule drives the occupancy state of the system toward the optimal task assignment q_n^* specified in (4).

The learning scheme aims at finding the coordinates σ_n^*, which depend on the typically unknown offered load. The learning scheme is parameterized by $\beta_n > 0$ and adjusts the value of $\boldsymbol{r}_n$ at certain arrival epochs, in steps of one unit. Specifically, when a task arrives, the learning scheme acts only under the following circumstances.

- If the system has at least $n\beta_n$ green tokens and at least one belongs to a server pool of class $i_{r_n - 1}$, then $\boldsymbol{r}_n$ is decremented by one after the task is dispatched.
- If the number of yellow tokens is smaller than or equal to one and there are no other tokens, then $\boldsymbol{r}_n$ is incremented by one after the task is dispatched.

Observe that exactly one of the thresholds changes when the value of $\boldsymbol{r}_n$ is modified and that this threshold changes by one unit. Also, note that

$$n - \sum_{i=1}^{m} n\boldsymbol{q}_n(i, \ell_i(\boldsymbol{r}_n)) \quad \text{and} \quad n\alpha_n(i_{r_n}) - n\boldsymbol{q}_n(i_{r_n}, j_{r_n})$$

are the number of green and yellow tokens, respectively.

3.2.1. Comparison with the Homogeneous Case. When all the server pools are of the same class, SLTA reduces to the load balancing policy studied in Goldsztajn et al. (2021a). In this case, there is a single threshold whose optimal value is simply $\lfloor \rho \rfloor$. When the threshold has this value, the number of green tokens and the total number of tokens are typically small and positive, respectively. On the other hand, when the threshold is below optimal, the total number of tokens tends to be zero, and when the threshold is larger than optimal, the number of green tokens tends to be relatively large. These properties are used to adjust the threshold in an online manner when the offered load is unknown. In few words, the threshold is increased in the absence of tokens, and it is decreased if the number of green tokens is large enough.

In the general case, there are as many thresholds as the number of server pool classes, and the optimal threshold values depend intricately on the utility functions and the offered load. However, the ranking $\triangleleft$ introduced in Section 2.3 makes it possible to express all the thresholds as a function of the coordinates σ_n^*. Hence, the optimal

thresholds can still be found through a one-directional search, but now in the totally ordered space $\mathcal{I}_+$. Moreover, because this space is countable, the search can be carried out by adjusting the integral parameter $\boldsymbol{r}_n$ until $(i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$ reaches the optimal value σ_n^*.

The learning scheme of SLTA operates so that all the thresholds are at their optimal values if and only if $\boldsymbol{r}_n$ is at its optimal value, and when this happens, the number of green tokens and the total number of tokens are typically small and positive, respectively. When $\boldsymbol{r}_n$ is below optimal, all the thresholds are smaller than or equal to their optimal values, and at least one of the thresholds is strictly smaller than optimal; in this case, the total number of tokens tends to be zero. Similarly, when $\boldsymbol{r}_n$ is above optimal, all the thresholds are larger than or equal to their optimal values and at least one of the thresholds is above optimal; as a result, the number of green tokens tends to be relatively large. As in the homogeneous case, these observations are used to adjust $\boldsymbol{r}_n$ over time. Loosely speaking, in the absence of tokens, $\boldsymbol{r}_n$ is increased by one unit, which implies that one of the thresholds is increased by one unit, and when the number of green tokens is large enough, $\boldsymbol{r}_n$ is decreased by one unit, and thus one of the thresholds is decreased by one unit.

The dispatching rule and the online learning scheme of SLTA make distinctions between green tokens of class $i_{\boldsymbol{r}_n-1}$ and green tokens of any other class. The rationale is that the marginal utility of server pools of class i with $\ell_i(\boldsymbol{r}_n) - 1$ tasks is the lowest when $i = i_{\boldsymbol{r}_n-1}$; thus, it makes sense to give green tokens of this class the lowest priority for receiving new tasks. Although this may slightly improve performance, it is not crucial. Nevertheless, the differential treatment of class $i_{\boldsymbol{r}_n-1}$ simplifies the mathematical analysis of SLTA. In particular, the distinction made in the description of the learning scheme ensures that if the system had yellow tokens, then it will continue to have yellow tokens after $\boldsymbol{r}_n$ is decreased, which is used in Remark 1. In addition, the differential treatment by the dispatching rule simplifies the proof of Proposition 6.

3.3. Stochastic Models

If service times are exponentially distributed, then $\boldsymbol{q}_n$ and $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ are continuous-time Markov chains when the load balancing policies are JLMU and SLTA, respectively. In either case, the process $\boldsymbol{s}_n$ that describes the normalized total number of tasks is defined by (1). Due to the infinite-server dynamics of the system, $n\boldsymbol{s}_n$ has the law of an $M/M/\infty$ queue with arrival rate $n\lambda$ and service rate μ.

Let ℓ_1 be the space of absolutely summable sequences in $\mathbb{R}^{\mathcal{I}}$, equipped with the norm

$$\|x\|_1 := \sum_{(i,j)\in\mathcal{I}} |x(i,j)| \quad \text{for all} \quad x \in \ell_1.$$

Throughout we assume that $\boldsymbol{s}_n(0)$ is finite, so $\boldsymbol{q}_n(0)$ takes values in ℓ_1. As a result, if we let $F_n := \{k/n : 0 \le k \le n\}$, then $\boldsymbol{q}_n$ takes values in the set

$$Q_n := \left\{ q \in F_n^{\mathcal{I}} \cap \ell_1 : q(i,j+1) \le q(i,j) \le q(i,0) = \alpha_n(i) \text{ for all } (i,j) \in \mathcal{I} \right\}.$$

If the load balancing policy is JLMU, then the state space of $\boldsymbol{q}_n$ is defined as the subset S_n of Q_n that is reachable from an empty occupancy state. If the load balancing policy is SLTA, then the state space of $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ is the subset S_n of $Q_n \times \{r \in \mathbb{N} : r \ge 1\}$ that is reachable from an empty occupancy state with $\boldsymbol{r}_n = 1$.

The notation used for the processes $\boldsymbol{s}_n$ and $\boldsymbol{q}_n$, for the state space S_n, and for some other objects that will be defined later, is exactly the same for JLMU and SLTA, but we always indicate which policy is being considered.

3.4. Asymptotic Optimality

Throughout the rest of the paper, we assume that there exist constants $\alpha(i) \in (0,1)$ and a random variable q_0 such that the following limits hold:

$$\lim_{n\to\infty} \alpha_n(i) = \alpha(i) \quad \text{for all} \quad i \quad \text{and} \quad \lim_{n\to\infty} \|\boldsymbol{q}_n(0) - q_0\|_1 = 0. \tag{9}$$

In analogy with (5) and (4), we consider the unique $\sigma_* = (i_{r_*}, j_{r_*}) \in \mathcal{I}_+$ such that

$$\sum_{(i,j)\rhd\sigma_*} \alpha(i) \le \rho < \sum_{(i,j)\trianglerighteq\sigma_*} \alpha(i), \tag{10}$$

and we define an occupancy state q_* in terms of σ_* by

$$q_*(i,j) := \begin{cases} 0 & \text{if} \quad (i,j) \lhd \sigma_*, \\ \alpha(i) & \text{if} \quad (i,j) \rhd \sigma_*, \\ \rho - \displaystyle\sum_{(i,j)\rhd\sigma_*} \alpha(i) & \text{if} \quad (i,j) = \sigma_*, \end{cases} \quad \text{for all} \quad (i,j) \in \mathcal{I}_+. \tag{11}$$

In Section 8.1, we establish that $\boldsymbol{q}_n$ and $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ have a unique stationary distribution for all n when the assignment policies are JLMU and SLTA, respectively. The following theorem is proved in Section 8.2 and implies that both policies are asymptotically optimal if the marginal utilities are bounded, and the service times are exponentially distributed; we also require that (9) holds, and we impose some mild technical assumptions, to be stated in Section 4.2.1. The condition on the marginal utilities always holds when the utility functions are nondecreasing due to the concavity of these functions.

Theorem 2. *Suppose that service times are exponentially distributed. Also, if the load balancing policy is SLTA, assume that the assumptions of Section 4.2.1 hold and that*

$$\lim_{n\to\infty} \beta_n = 0 \quad \text{and} \quad \liminf_{n\to\infty} n^{\gamma_0}\beta_n > 0 \quad \text{for some} \quad \gamma_0 \in (0, 1/2).$$

The following statements hold in the asymptotic regime described by (9).

(a) *Suppose that JLMU is used and let q_n have the stationary distribution of $\boldsymbol{q}_n$. Then the random variables q_n converge weakly in ℓ_1 to q_*.*

(b) *Suppose that SLTA is used and let (q_n, r_n) have the stationary distribution of $(\boldsymbol{q}_n, \boldsymbol{r}_n)$. The random variables (q_n, r_n) converge weakly in $\ell_1 \times \mathbb{N}$ to (q_*, r_*).*

(c) *Furthermore, if the load balancing policy is either JLMU or SLTA and the marginal utilities are bounded, then the random variables $u(q_n)$ are uniformly integrable and*

$$\lim_{n\to\infty} E\big[u(q_n)\big] = u(q_*) = \lim_{n\to\infty} u\big(q_n^*\big).$$

The claims concerning the stationary overall utilities are proved using (a) and (b), as well as the fact that $u(q)$ is a bounded linear functional of $q \in \ell_1$. To establish (a) and (b), we first use drift analysis to prove that the random variables in (a) and (b) are tight in ℓ_1 and $\ell_1 \times \mathbb{N}$, respectively. Then (a) is established through an interchange of limits argument based on a fluid limit and a global asymptotic stability result for the fluid dynamics; these two results are stated in Theorems 5 and 4, respectively. A different type of argument is used to prove (b). Namely, the fluid limit step is circumvented, and Theorem 6 serves as the counterpart of Theorems 4 and 5, as illustrated in Figure 6.

Deriving a fluid limit for a SLTA system would be inherently difficult due to the intricate interdependence between the dispatching rule and the learning scheme and because the actions of the learning scheme are triggered by excursions of the occupancy process that have vanishing size. We deal with these challenges using a methodology of Goldsztajn et al. (2021a) to derive the fluid approximation of Theorem 6, which consists of asymptotic bounds, over arbitrarily long intervals of time, for the occupancy state and the thresholds. As noted earlier, this fluid approximation serves as a counterpart of both the fluid limit and the global asymptotic stability results for JLMU.

3.4.1. Simulation Experiments. The asymptotic optimality of JLMU and SLTA is illustrated by Table 1, which shows estimates of the mean stationary overall utility $E[u(q_n)]$ for simulation experiments with different values of n. All the estimates correspond to systems with two server pool classes of equal size and utility functions of the form $u_i(x) = x\log(r(i)/x)$ for $r = (20, 30)$. Also, two different values of ρ are considered, so that the optimal

Figure 6. Schematic View of the Proofs of (a) and (b) of Theorem 2, on the Left and Right, Respectively

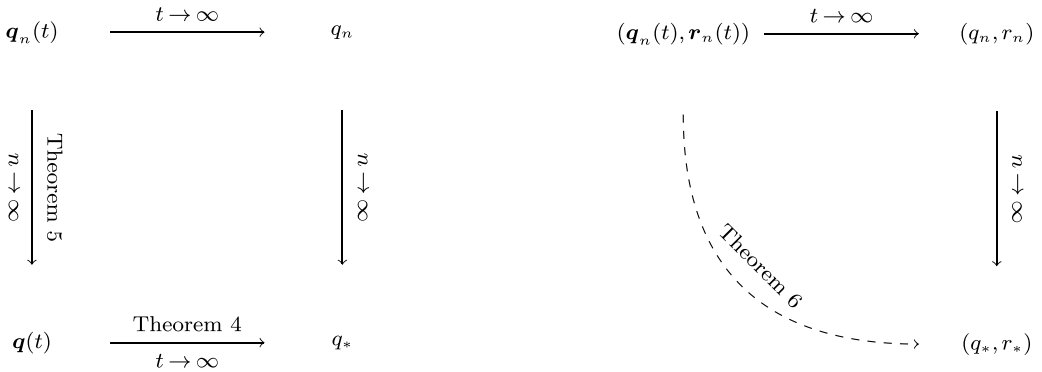


Table 1. Results of Simulation Experiments with Different Values of n

	$\rho = 9.75 \quad u^*(\rho) \cong 9.1731$			$\rho = 10 \quad u^*(\rho) \cong 9.1629$		
n	$u^*(\langle \boldsymbol{s}_n \rangle)$	$\langle u(\boldsymbol{q}_n) \rangle$ JLMU	$\langle u(\boldsymbol{q}_n) \rangle$ SLTA	$u^*(\langle \boldsymbol{s}_n \rangle)$	$\langle u(\boldsymbol{q}_n) \rangle$ JLMU	$\langle u(\boldsymbol{q}_n) \rangle$ SLTA
50	9.1752	9.1652	9.1649	9.1594	9.1433	9.1439
100	9.1746	9.1699	9.1698	9.1627	9.1499	9.1498
150	9.1738	9.1706	9.1705	9.1633	9.1534	9.1534
200	9.1743	9.1723	9.1723	9.1634	9.1556	9.1555
250	9.1735	9.1720	9.1720	9.1639	9.1572	9.1572
300	9.1738	9.1727	9.1726	9.1613	9.1548	9.1547
350	9.1742	9.1734	9.1734	9.1619	9.1563	9.1562
400	9.1734	9.1728	9.1728	9.1632	9.1578	9.1578
450	9.1737	9.1732	9.1731	9.1630	9.1576	9.1576
500	9.1726	9.1721	9.1721	9.1632	9.1583	9.1583

Notes. The systems considered have two server pool classes of equal size and utility functions of the form $u_i(x) = x \log(r(i)/x)$ with $r = (20, 30)$. Service times are exponentially distributed with $\mu = 1$ and the time averages $\langle \cdot \rangle$ are computed over intervals of length 180 with equilibrium initial conditions. The same sequences of interarrival and service times were used in the simulation experiments for JLMU and SLTA. Also, $\beta_n = 1/n^{0.45}$ in the simulations for SLTA. The upper bound of Theorem 1 for a normalized offered load of x is denoted by $u^*(x)$.

task assignment defined in (4) takes two distinct forms. For $\rho = 9.75$, the optimal task assignment is such that all server pools of class 1 have 8 tasks, half of the server pools of class 2 have 11 tasks, and the other half have 12 tasks. For $\rho = 10$, the optimal task assignment is such that all server pools of class 1 have 8 tasks and all server pools of class 2 have 12 tasks.

All the server pools of the same class have the same number of tasks when $\rho = 10$, and thus we say that the optimal task assignment q_n^* is *integral*; that is, the optimal task assignment is integral if $q_n^*(\sigma_n^*) = 0$. In contrast, the optimal task assignment is *fractional* when $\rho = 9.75$ because server pools of class 2 may have either 11 or 12 tasks. Server pools of class 1 behave similarly in the fractional and integral settings: almost all of the time all server pools of class 1 have precisely eight tasks when n is moderately large. However, the behavior of server pools of class 2 depends on the setting. In the fractional case, server pools typically have 11 or 12 tasks and the fractions of server pools with 11 and 12 tasks oscillate around a half. In the integral case, server pools typically have 12 tasks but a small number of server pools sometimes have 11 or 13 tasks instead. The aggregate utility of the system decreases by $\Delta(2, 11)$ whenever a server pool of class 2 goes from 12 to 11 tasks and increases by the same quantity when the server pool goes from 11 to 12 tasks. Therefore, the contributions to the average overall utility of the oscillations observed in the fractional case roughly balance each other. In the integral case, the aggregate utility increases by $\Delta(2, 12)$, instead of $\Delta(2, 11)$, if a server pool of class 2 goes from 12 to 13 tasks. Therefore, in the integral case, the contributions to the average overall utility of class 2 server pools that drop to 11 tasks or reach 13 tasks are amplified by different marginal utilities. As a result, the mean stationary overall utility $E[u(q_n)]$ is closer to the upper bound $u(q_n^*)$ in the fractional case, as reflected by the estimates in Table 1.

Although there is a difference between the fractional and integral settings, in both settings, the empirical mean of the overall utility $u(\boldsymbol{q}_n)$ is extremely close to the upper bound $u(q_n^*)$ across all the values of n listed in Table 1. Furthermore, the deviation of the empirical mean from the upper bound approaches zero as n increases in both cases. We also observe that the empirical mean of $u(\boldsymbol{q}_n)$ is almost the same for JLMU and SLTA in all the experiments and particularly for the largest values of n.

A final remark on the simulation experiments is that the empirical mean of $u(\boldsymbol{q}_n)$ is slightly larger than $u(q_n^*)$ in a few of the experiments within the fractional setting: both for JLMU and SLTA when $n = 350$ and just for JLMU when $n = 450$. It may be checked that the statement of Theorem 1 still holds if the stationary expectation sign is replaced by a time average and the upper bound is computed through (4) and (5) but with ρ replaced by the time average of $\boldsymbol{s}_n$; the proof does not change. The value of the upper bound when ρ is replaced by the time average of $\boldsymbol{s}_n$ is displayed in Table 1, and in all the experiments, the empirical mean of $u(\boldsymbol{q}_n)$ is indeed smaller than this empirical upper bound. Thus, the experiments where the empirical mean of $u(\boldsymbol{q}_n)$ slightly exceeds the upper bound $u(q_n^*)$ are an indication of how close the performance of JLMU and SLTA is to optimal.

4. Approximation Theorems

In this section, we assume exponential service times and we state several results used to prove Theorem 2. In Section 4.1, we specify a fluid model of a JLMU system, based on differential equations, and we state some properties of this model. In Section 4.2, we state limit theorems that characterize the asymptotic transient behavior of JLMU and SLTA.

4.1. Fluid Model of JLMU

Consider a large-scale system where the load balancing policy is JLMU and assume that $\alpha(i)$ is the fraction of server pools of class i. Then the occupancy state of the system remains within the set

$$Q := \left\{ q \in [0,1]^{\mathcal{I}} \cap \ell_1 : q(i,j+1) \le q(i,j) \le q(i,0) = \alpha(i) \text{ for all } (i,j) \in \mathcal{I} \right\}.$$

The evolution of the occupancy state of this large-scale system can be modeled through the system of differential equations introduced in the following definition.

Definition 1. We say that $\boldsymbol{q} : [0,\infty) \to Q$ is a fluid trajectory if the coordinate functions $\boldsymbol{q}(i,j)$ are absolutely continuous for all $(i,j) \in \mathcal{I}$ and the following conditions hold almost everywhere with respect to the Lebesgue measure:

$$\dot{\boldsymbol{q}}(i,j) = \Lambda(\boldsymbol{q},i,j) - \mu j \left[\boldsymbol{q}(i,j) - \boldsymbol{q}(i,j+1) \right], \tag{12a}$$

$$\Lambda(\boldsymbol{q},i,j) \ge 0 \quad \text{for all} \quad (i,j) \in \mathcal{I}_+, \tag{12b}$$

where $\Lambda : Q \times \mathcal{I}_+ \to \mathbb{R}$ is defined by

$$\Lambda(q,i,j) := \begin{cases} 0 & \text{if} \quad (i,j) \lhd \sigma(q), \\ \mu j \left[\alpha(i) - q(i,j+1) \right] & \text{if} \quad (i,j) \rhd \sigma(q), \\ \lambda - \sum_{(k,l) \rhd \sigma(q)} \mu l \left[\alpha(k) - q(k,l+1) \right] & \text{if} \quad (i,j) = \sigma(q). \end{cases}$$

In the latter definition, λ is the arrival rate of tasks normalized by the number of server pools, μ is the service rate of tasks, and $\boldsymbol{q}(i,j)$ represents the fraction of server pools that are of class i and have at least j tasks. Thus, the system of Differential Equations (12) has a simple interpretation. The right-most term of (12a) corresponds to the departure rate of tasks from server pools of class i with exactly j tasks, and $\Lambda(\boldsymbol{q},i,j)$ represents the arrival rate of tasks to server pools that belong to class i and have precisely $j-1$ tasks. The definition of Λ is motivated by the following remarks.

- Server pools of class i with exactly $j-1$ tasks are not assigned additional tasks if $(i,j) \lhd \sigma(\boldsymbol{q})$. Hence, we should have $\Lambda(\boldsymbol{q},i,j) = 0$ in this case.
- All server pools of class i have at least j tasks if $(i,j) \rhd \sigma(\boldsymbol{q})$. Therefore, $\Lambda(\boldsymbol{q},i,j)$ should be equal to the last term of (12a) in this case, because $\boldsymbol{q}(i,j)$ is at its maximum value and thus its derivative should be zero.
- The total arrival rate of tasks normalized by the number of server pools is equal to λ, and this determines the value of $\Lambda(\boldsymbol{q},i,j)$ when $(i,j) = \sigma(\boldsymbol{q})$.

4.1.1. Properties of Fluid Trajectories.

The two results stated here are proved in Section 6.1. The first one is a uniqueness theorem for the solutions of (12). Existence is ensured by Theorem 5 of Section 4.2.

Theorem 3. *Fix an initial condition $q \in Q$. If there exists a fluid trajectory $\boldsymbol{q}$ such that $\boldsymbol{q}(0) = q$, then this fluid trajectory is unique.*

To prove this theorem, we first show that all fluid trajectories satisfy an infinite system of integral equations, stated using Skorokhod one-dimensional reflection mappings. The theorem is then proved using a Lipschitz property of these mappings and a uniqueness result for certain Kolmogorov backward equations.

Besides uniqueness of solutions of (12), we also establish that there exists a unique equilibrium point and that this equilibrium point is globally asymptotically stable; that is, all fluid trajectories converge to the unique equilibrium over time.

Theorem 4. *Let q_* be as in (11). Then q_* is the unique equilibrium of (12). Furthermore, all fluid trajectories converge to q_* in ℓ_1 over time.*

Recall that (11) is the counterpart of (4), which is used to formulate the upper bound for the mean stationary overall utility provided in Theorem 1. It is not difficult to check that $u(q_n^*) \to u(q_*)$ as n grows large, which hints at the asymptotic optimality of JLMU.

4.2. Limit Theorems

In Section 5.1, we construct the processes defined in Section 3.3 on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for all n, in such a way that the sample paths of the occupancy processes lie in the space $D_{\ell_1}[0,\infty)$ of càdlàg functions with values in ℓ_1, which we endow with the topology of uniform convergence over compact sets. This construction is used to prove limit theorems that characterize the asymptotic transient behavior of JLMU and SLTA. Before stating these theorems, we introduce some mild technical assumptions.

4.2.1. Technical Assumptions. As indicated earlier, we assume that (9) holds with q_0 a random variable that takes values in Q and represents the limiting initial occupancy state. The initial number of tasks in the limit, normalized by the number of server pools, is defined as

$$s_0 := \sum_{(i,j)\in\mathcal{I}_+} q_0(i,j). \tag{13}$$

If the load balancing policy is SLTA, then we assume that the first inequality in (10) is strict and that there exists a constant $\gamma_0 \in (0,1/2)$ such that

$$\lim_{n\to\infty} \beta_n = 0 \quad \text{and} \quad \liminf_{n\to\infty} n^{\gamma_0}\beta_n > 0. \tag{14}$$

These assumptions are used to prove that the learning scheme reaches an equilibrium in all large enough systems with probability one. Finally, we adopt the following assumptions about the initial state of the system: There exists a random variable $R \geq 1$ such that

$$\boldsymbol{r}_n(0) \leq R, \tag{15a}$$

$$\boldsymbol{q}_n(0,i,j) < \alpha_n(i) \quad \text{for all} \quad (i,j) \trianglelefteq \left(i_{\boldsymbol{r}_n(0)}, j_{\boldsymbol{r}_n(0)}\right), \tag{15b}$$

for all n with probability one. We impose (15b) just to simplify the analysis; this property always holds after a certain time, which depends on the initial state of the system.

Remark 1. Property (15b) is preserved by arrivals and departures; thus, it holds at all times provided that it holds at time zero. Furthermore, every new task is sent to a server pool with coordinates $(i,j) \trianglerighteq (i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$ if the number of tokens is positive right before the arrival. Hence, (15b) implies that tasks are sent to server pools with coordinates $(i,j) \trianglerighteq (i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$ at all times and for all n with a probability of one.

4.2.2. Statements of the Theorems. First we state a fluid limit for JLMU, proven in Section 6.2. In view of Theorem 4, this fluid limit implies that, as n grows large, the occupancy processes of JLMU approach functions that converge over time to the unique equilibrium of (12).

Theorem 5. *Suppose that the load balancing policy is JLMU. Then there exists a set of probability one Γ with the following property. If $\omega \in \Gamma$, then $\boldsymbol{q}_n(\omega)$ converges in $D_{\ell_1}[0,\infty)$ to the unique fluid trajectory with initial condition $q_0(\omega)$.*

Because q_0 is arbitrary, the previous theorem implies that solutions to (12) exist for all initial conditions. Therefore, Theorems 3 and 5 imply that for each initial condition $q \in Q$ there exists a unique fluid trajectory with initial condition q.

The proof of Theorem 5 uses a methodology of Bramson (1998) to prove that, with a probability of one, every subsequence of $\{\boldsymbol{q}_n : n \geq 1\}$ has a further subsequence that converges uniformly over compact sets with respect to a metric for the product topology of $\mathbb{R}^{\mathcal{I}}$. Then we show that this convergence in fact holds with respect to $\|\cdot\|_1$ and that the limits of convergent subsequences are fluid trajectories, also with a probability of one.

The counterpart of Theorems 4 and 5 for SLTA is the following result. The proof is provided in Section 7.2 and is based on a methodology of Goldsztajn et al. (2021a).

Theorem 6. *Suppose that the load balancing policy is SLTA and let σ_* and r_* be as in (10). There exist $\tau_{\mathrm{eq}} : [0,\infty) \to \mathbb{R}$ and a set of probability one Γ with the following property. If $\omega \in \Gamma$ and $T \geq \tau > \tau_{\mathrm{eq}}(s_0(\omega))$, then the next limits hold:*

$$\lim_{n\to\infty} \sup_{t\in[\tau,T]} |\boldsymbol{r}_n(\omega) - r_*| = 0, \tag{16a}$$

$$\lim_{n\to\infty} \sup_{t\in[\tau,T]} n^{\gamma}\left|\alpha(i) - \boldsymbol{q}_n(\omega,t,i,j)\right| = 0 \quad \text{if} \quad (i,j) \triangleright \sigma_* \quad \text{and} \quad \gamma \in [0,1/2), \tag{16b}$$

$$\limsup_{n\to\infty} \sup_{t\in[\tau,T]} \sum_{(i,j)\trianglelefteq\sigma^*} \boldsymbol{q}_n(\omega,t,i,j)\mathrm{e}^{\mu(t-\tau)} \leq c(\omega,\tau), \tag{16c}$$

where $c(\omega,\tau)$ can be expressed in terms of ρ, τ, and $s_0(\omega)$.

5. Strong Approximations

In this section, we construct the processes defined in Section 3.3 on a common probability space for all n. In addition, we prove that $\{\boldsymbol{q}_n : n \geq 1\}$ is almost surely relatively compact in $D_{\ell_1}[0,\infty)$ both for JLMU and SLTA. The construction of the processes is carried out in Section 5.1, and the relative compactness results are provided in Section 5.2.

5.1. Coupled Construction of Sample Paths

Consider the following stochastic processes and random variables.

• *Driving Poisson processes:* A collection $\{\mathcal{N}\}\cup\{\mathcal{N}_\nu : \nu\in\mathcal{I}_+\}$ of independent Poisson processes with unit rate, for counting arrivals and departures. These processes are defined on a common probability space $(\Omega_D,\mathcal{F}_D,\ \mathbb{P}_D)$.

• *Selection variables:* A family $\{U_k : k\geq 1\}$ of independent random variables, uniformly distributed on $[0,1)$ and defined on a common probability space $(\Omega_S,\mathcal{F}_S,\ \mathbb{P}_S)$.

• *Initial conditions:* Sequences $\{\boldsymbol{q}_n(0) : n\geq 1\}$ and $\{\boldsymbol{r}_n(0) : n\geq 1\}$ for SLTA of random variables for describing the initial states of the systems, defined on a common probability space $(\Omega_I,\mathcal{F}_I,\ \mathbb{P}_I)$ and satisfying the assumptions of Section 4.2.1.

Denote the completion of the product probability space of $(\Omega_D,\mathcal{F}_D,\ \mathbb{P}_D)$, $(\Omega_S,\mathcal{F}_S,\ \mathbb{P}_S)$ and $(\Omega_I,\mathcal{F}_I,\ \mathbb{P}_I)$ by $(\Omega,\mathcal{F},\ \mathbb{P})$. The processes introduced in Section 3.3 are constructed on the latter space as deterministic functions of the stochastic primitives.

5.1.1. Construction for JLMU. Let $\mathcal{N}_n^\lambda(t) := \mathcal{N}(n\lambda t)$ for each $t\geq 0$ and each n. This quantity will be used to count the number of tasks arriving to the system with n server pools during the interval $[0,t]$. Also, denote the jump times of $\mathcal{N}_n^\lambda$ by $\{\tau_{n,k} : k\geq 1\}$ and define $\tau_{n,0} := 0$. For each function $\boldsymbol{q} : [0,\infty)\to\mathcal{Q}_n$ and each n, we define two counting processes, for arrivals and departures, denoted $\mathcal{A}_n(\boldsymbol{q})$ and $\mathcal{D}_n(\boldsymbol{q})$, respectively. The coordinates $(i,j)\in\mathcal{I}$ of these processes are identically zero if $j=0$, whereas the other coordinates are defined as follows:

$$\mathcal{A}_n(\boldsymbol{q},t,i,j) := \frac{1}{n}\sum_{k=1}^{\mathcal{N}_n^\lambda(t)} \mathbb{1}_{\{(i,j)=\sigma(\boldsymbol{q}(\tau_{n,k}^-))\}}, \tag{17a}$$

$$\mathcal{D}_n(\boldsymbol{q},t,i,j) := \frac{1}{n}\mathcal{N}_{(i,j)}\left(n\int_0^t \mu j\big[\boldsymbol{q}(s,i,j)-\boldsymbol{q}(s,i,j+1)\big]ds\right). \tag{17b}$$

For each n, the functional equation

$$\boldsymbol{q} = \boldsymbol{q}_n(0) + \mathcal{A}_n(\boldsymbol{q}) - \mathcal{D}_n(\boldsymbol{q}) \tag{18}$$

has a unique solution with a probability of one. More precisely, there exists a set of probability of one Γ_0 with the following property: For each $\omega\in\Gamma_0$ and each n, there exists a unique càdlàg function $\boldsymbol{q}_n(\omega) : [0,\infty)\to\mathcal{Q}_n$ that solves (18). This solution can be constructed by forward induction on the jump times of the driving Poisson processes. The assumption $\boldsymbol{q}_n(\omega,0)\in\mathcal{Q}_n$ implies that $\boldsymbol{q}_n(\omega,0)$ has finitely many nonzero coordinates and ensures that the constructed solution is defined on $[0,\infty)$ with probability one; that is, the constructed solution does not explode in finite time.

The occupancy processes are defined by extending the previous solutions to Ω, setting $\boldsymbol{q}_n(\omega,t)=0$ for all $t\geq 0$ and all $\omega\notin\Gamma_0$. In addition, we let

$$\mathcal{A}_n := \mathcal{A}_n(\boldsymbol{q}_n) \quad\text{and}\quad \mathcal{D}_n := \mathcal{D}_n(\boldsymbol{q}_n),$$

and we note that the sample paths of $\mathcal{A}_n$, $\mathcal{D}_n$ and $\boldsymbol{q}_n$ lie in $D_{\ell_1}[0,\infty)$. Functional Equation (18) can now be rewritten as follows:

$$\boldsymbol{q}_n(\omega) = \boldsymbol{q}_n(\omega,0) + \mathcal{A}_n(\omega) - \mathcal{D}_n(\omega) \quad\text{for all}\quad \omega\in\Gamma_0. \tag{19}$$

This construction endows the processes $\boldsymbol{q}_n$ with the intended statistical behavior. The processes $\mathcal{A}_n(i,j)$ count the arrivals to server pools of class i with precisely $j-1$ tasks and the processes $\mathcal{D}_n(i,j)$ count the departures from server pools of class i with exactly j tasks. Indeed, $\mathcal{A}_n(i,j)$ has a jump at the arrival epoch $\tau_{n,k}$ if and only if the incoming task should be assigned to a server pool of class i with $j-1$ tasks under the JLMU policy. In addition, the intensity of $\mathcal{D}_n(i,j)$ equals the total number of tasks in server pools of class i with exactly j tasks times the rate at which tasks are executed, and this totals the departure rate from server pools of class i with precisely j tasks.

5.1.2. Construction for SLTA. The processes $(\boldsymbol{q}_n,\boldsymbol{r}_n)$ are constructed to a large extent as in Section 5.1.1 when the load balancing policy is SLTA. The only differences arise in (17a) and (18). Namely, (17a) must be modified to capture the dispatching rule of SLTA, and (18) must be accompanied by another equation for describing the evolution of $\boldsymbol{r}_n$.

The counterpart of (17a), with an extra argument $\boldsymbol{r}:[0,\infty)\to\{r\in\mathbb{N}:r\geq 1\}$, is

$$\mathcal{A}_n(\boldsymbol{q},\boldsymbol{r},t,i,j):=\frac{1}{n}\sum_{k=1}^{\mathcal{N}_n^{\lambda}(t)}\eta_k(\boldsymbol{q}(\tau_{n,k}^-),\boldsymbol{r}(\tau_{n,k}^-),i,j)\quad\text{for all}\quad(i,j)\in\mathcal{I}_+.$$

The functions η_k are defined in Online Appendix B using the selection variables U_k, so that they have the following property. If (q,r) is the value of $(\boldsymbol{q}_n,\boldsymbol{r}_n)$ when the k^{th} task arrives, then SLTA sends this task to a server pool of class i with precisely $j-1$ tasks if and only if $\eta_k(q,r,i,j)=1$. Moreover, $\eta_k(q,r,h,l)=0$ for all $(h,l)\neq(i,j)$.

The analog of the functional Equation (18) is

$$\boldsymbol{q}(t)=\boldsymbol{q}_n(0)+\mathcal{A}_n(\boldsymbol{q},\boldsymbol{r},t)-\mathcal{D}_n(\boldsymbol{q},t),\tag{20a}$$

$$\boldsymbol{r}(t)=\boldsymbol{r}_n(0)+\sum_{k=1}^{\mathcal{N}_n^{\lambda}(t)}\left[\mathbb{1}_{I_{n,k}}-\mathbb{1}_{D_{n,k}}\right],\tag{20b}$$

where the sets $I_{n,k}$ and $D_{n,k}$ are defined formally in Online Appendix A. The former set indicates that $(\boldsymbol{q},\boldsymbol{r})$ corresponds to a system with no green tokens and at most one yellow token right before the k^{th} arrival. The latter set indicates that the number of green tokens is larger than or equal to $n\beta_n$ and that at least one of these tokens belongs to a server pool of class i_{r-1} right before the k^{th} arrival. Also, $\mathcal{D}_n(\boldsymbol{q})$ is defined as in (17b).

As in Section 5.1.1, there exists a set of probability of one Γ_0 with the next property. For each $\omega\in\Gamma_0$ and each n, there exists a unique pair of càdlàg functions $\boldsymbol{q}_n(\omega):[0,\infty)\to Q_n$ and $\boldsymbol{r}_n(\omega):[0,\infty)\to\{r\in\mathbb{N}:r\geq 1\}$ that solve (20); these functions can be constructed by forward induction on the jumps of the driving Poisson processes. The processes $(\boldsymbol{q}_n,\boldsymbol{r}_n)$ are defined by extending the above solutions to Ω, setting $\boldsymbol{q}_n(\omega,t)=0$ and $\boldsymbol{r}_n(\omega,t)=0$ for all $t\geq 0$ and all $\omega\notin\Gamma_0$. In addition, we define

$$\mathcal{A}_n:=\mathcal{A}_n(\boldsymbol{q}_n,\boldsymbol{r}_n)\quad\text{and}\quad\mathcal{D}_n:=\mathcal{D}_n(\boldsymbol{q}_n),$$

and we note that the sample paths of $\mathcal{A}_n$, $\mathcal{D}_n$ and $\boldsymbol{q}_n$ lie in $D_{\ell_1}[0,\infty)$. Functional Equation (20) can now be rewritten as follows:

$$\boldsymbol{q}_n(\omega,t)=\boldsymbol{q}_n(\omega,0)+\mathcal{A}_n(\omega,t)-\mathcal{D}_n(\omega,t),\tag{21a}$$

$$\boldsymbol{r}_n(\omega,t)=\boldsymbol{r}_n(\omega,0)+\sum_{k=1}^{\mathcal{N}_n^{\lambda}(\omega,t)}\left[\mathbb{1}_{I_{n,k}}(\omega)-\mathbb{1}_{D_{n,k}}(\omega)\right],\tag{21b}$$

for all $\omega\in\Gamma_0$ and all $t\geq 0$.

5.2. Relative Compactness Results

Let $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ denote the space of càdlàg functions on $[0,\infty)$ with values in $\mathbb{R}^{\mathcal{I}}$. We endow the space $\mathbb{R}^{\mathcal{I}}$ with the metric defined in Online Appendix B, which is compatible with the product topology, and we equip $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ with the topology of uniform convergence over compact sets. The following proposition is proved in Online Appendix B.

Proposition 2. *Suppose that the load balancing policy is JLMU or SLTA. There exists a set of probability of one Γ_∞, where*

$$\lim_{n\to\infty}\|\boldsymbol{q}_n(0)-q_0\|_1=0,\tag{22a}$$

$$\lim_{n\to\infty}\sup_{t\in[0,T]}n^{\gamma}\left|\frac{1}{n}\mathcal{N}_n^{\lambda}(t)-\lambda t\right|=0,\tag{22b}$$

$$\lim_{n\to\infty}\sup_{t\in[0,\mu jT]}n^{\gamma}\left|\frac{1}{n}\mathcal{N}_{(i,j)}(nt)-t\right|=0\quad\textit{for all}\quad(i,j)\in\mathcal{I}_+,\tag{22c}$$

for all $T\geq 0$ and $\gamma\in[0,1/2)$. Also, $\{\mathcal{A}_n(\omega):n\geq 1\}$, $\{\mathcal{D}_n(\omega):n\geq 1\}$ and $\{\boldsymbol{q}_n(\omega):n\geq 1\}$ are relatively compact subsets of $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ for all $\omega\in\Gamma_\infty$ and satisfy that the limit of every convergent subsequence is a function with locally Lipschitz coordinates. If the load balancing policy is SLTA, then there exists a random variable $R\geq 1$ such that, apart from the previous properties, we also have on Γ_∞ that

$$\boldsymbol{r}_n(0)\leq R\quad\textit{and}\quad\boldsymbol{q}_n(0,i,j)<\alpha_n(i)\quad\textit{for all}\quad(i,j)\trianglelefteq\left(i_{r_n(0)},j_{r_n(0)}\right)\quad\textit{and}\quad n.\tag{23}$$

The product topology of $\mathbb{R}^{\mathcal{I}}$ is coarser than the topology of ℓ_1; thus, convergence in $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ does not imply convergence in $D_{\ell_1}[0,\infty)$. The following technical lemma is used to demonstrate that $\{\mathcal{A}_n:n\geq 1\}$, $\{\mathcal{D}_n:n\geq 1\}$ and $\{\boldsymbol{q}_n:n\geq 1\}$ are relatively compact in $D_{\ell_1}[0,\infty)$ with a probability of one; the proof is given in Online Appendix A.

Lemma 2. *Suppose that the load balancing policy is JLMU or SLTA. There exists a set of probability of one $\Gamma \subset \Gamma_\infty$ with the following property. For each $\omega \in \Gamma$ and $T \geq 0$, there exist $j_T(\omega)$ and $n_T(\omega)$ such that*

$$\mathcal{A}_n(\omega, t, i, j) = 0 \quad \text{for all} \quad t \in [0,T], \quad 1 \leq i \leq m, \quad j > j_T(\omega) \quad \text{and} \quad n \geq n_T(\omega).$$

Also, if the load balancing policy is SLTA, then there exists $R_T(\omega)$ such that

$$\boldsymbol{r}_n(\omega, t) \leq R_T(\omega) \quad \text{for all} \quad t \in [0,T] \quad \text{and} \quad n \geq n_T(\omega).$$

Proposition 3. *The sequences $\{\mathcal{A}_n(\omega) : n \geq 1\}$, $\{\mathcal{D}_n(\omega) : n \geq 1\}$ and $\{\boldsymbol{q}_n(\omega) : n \geq 1\}$ are relatively compact in $D_{\ell_1}[0,\infty)$ for all $\omega \in \Gamma$.*

Proof. We fix some $\omega \in \Gamma$, which we omit from the notation. For $\{\mathcal{A}_n : n \geq 1\}$, the claim is a straightforward consequence of Proposition 2 and Lemma 2. Here we prove the claim for $\{\boldsymbol{q}_n : n \geq 1\}$. If the load balancing policy is JLMU, then (19) and (22a) imply that the claim also holds for $\{\mathcal{D}_n : n \geq 1\}$. If the load balancing policy is SLTA, then we must invoke (21a) instead of (19).

Consider any increasing sequence of natural numbers. By Proposition 2, there exists a subsequence $\mathcal{K}$ such that $\{\boldsymbol{q}_k : k \in \mathcal{K}\}$ converges in $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ to a function $\boldsymbol{q} \in D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ that satisfies $\boldsymbol{q}(0) = q_0$ and has locally Lipschitz coordinates. Therefore, it suffices to prove that the latter limit in fact holds in $D_{\ell_1}[0,\infty)$. More specifically, we have to demonstrate that $\boldsymbol{q} \in D_{\ell_1}[0,\infty)$ and that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \|\boldsymbol{q}_k(t) - \boldsymbol{q}(t)\|_1 = 0 \quad \text{for all} \quad T \geq 0.$$

For this purpose, fix arbitrary $T \geq 0$ and $\varepsilon > 0$. In addition, let j_T and n_T be as in the statement of Lemma 2, which implies that $\boldsymbol{q}_k(i,j)$ and $\boldsymbol{q}(i,j)$ are nonincreasing on $[0,T]$ provided that $j > j_T$ and $k \geq n_T$. The coordinates of $\boldsymbol{q}$ are continuous; thus, we may conclude from the monotone convergence theorem that $\|\boldsymbol{q}(s) - \boldsymbol{q}(t)\|_1 \to 0$ as $s \to t \in [0,T]$ monotonically, from above or below. Because T is arbitrary, $\boldsymbol{q}$ is continuous with respect to $\|\cdot\|_1$ and, in particular, $\boldsymbol{q} \in D_{\ell_1}[0,\infty)$.

For all $t \in [0,T]$, $l \geq j_T$ and $k \geq n_T$, we have

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j>l} \left|\boldsymbol{q}_k(t,i,j) - \boldsymbol{q}(t,i,j)\right| &\leq \sum_{i=1}^{m}\sum_{j>l} \boldsymbol{q}_k(t,i,j) + \sum_{i=1}^{m}\sum_{j>l} \boldsymbol{q}(t,i,j) \\
&\leq \sum_{i=1}^{m}\sum_{j>l} \boldsymbol{q}_k(0,i,j) + \sum_{i=1}^{m}\sum_{j>l} q_0(i,j) \\
&\leq \|\boldsymbol{q}_k(0) - q_0\|_1 + 2\sum_{i=1}^{m}\sum_{j>l} q_0(i,j).
\end{aligned}$$

Because $q_0 \in \ell_1$ and $\|\boldsymbol{q}_k(0) - q_0\|_1 \to 0$ with k, there exist $j_\varepsilon \geq j_T$ and $k_0 \geq n_T$ such that the following inequality holds for all $t \in [0,T]$ and $k \geq k_0$:

$$\sum_{i=1}^{m}\sum_{j>j_\varepsilon} \left|\boldsymbol{q}_k(t,i,j) - \boldsymbol{q}(t,i,j)\right| \leq \|\boldsymbol{q}_k(0) - q_0\|_1 + 2\sum_{i=1}^{m}\sum_{j>j_\varepsilon} q_0(i,j) \leq \frac{\varepsilon}{2}.$$

Convergence in $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ implies uniform convergence over compact sets of the coordinate functions. In particular, there exists $k_\varepsilon \geq k_0$ such that

$$\sup_{t\in[0,T]} \sum_{i=1}^{m}\sum_{j=0}^{j_\varepsilon} \left|\boldsymbol{q}_k(t,i,j) - \boldsymbol{q}(t,i,j)\right| \leq \frac{\varepsilon}{2} \quad \text{for all} \quad k \geq k_\varepsilon.$$

Therefore, we conclude that

$$\sup_{t\in[0,T]} \|\boldsymbol{q}_k(t) - \boldsymbol{q}(t)\|_1 \leq \varepsilon \quad \text{for all} \quad k \geq k_\varepsilon,$$

which completes the proof because T and ε are arbitrary. □

6. Limiting Behavior of JLMU

In this section we assume that JLMU is used and we prove Theorems 3, 4, and 5. The first two theorems are proved in Section 6.1, which is devoted to the study of fluid trajectories. The proof of Theorem 5 is provided in Section 6.2.

6.1. Properties of Fluid Trajectories

To prove the uniqueness of fluid trajectories, we show that every fluid trajectory satisfies a system of equations involving one-dimensional Skorokhod reflection mappings. Then we use a Lipschitz property of these mappings to prove that the system of equations cannot have multiple solutions for a given initial condition. The proof strategy is inspired by a fluid limit derived in Bhamidi et al. (2022) using Skorokhod reflection mappings; this fluid limit corresponds to a system of parallel single-server queues with a JSQ policy.

Consider the space $D[0,\infty)$ of all real càdlàg functions defined on $[0,\infty)$ and let

$$\|\boldsymbol{x}\|_T := \sup_{t\in[0,T]} |\boldsymbol{x}(t)| \quad \text{for all} \quad \boldsymbol{x}\in D[0,\infty) \quad \text{and} \quad T\geq 0.$$

The next lemma introduces the one-dimensional Skorokhod mappings with upper reflecting barrier; a proof is provided in Online Appendix A.

Lemma 3. *Fix $\alpha\in\mathbb{R}$ and suppose that $\boldsymbol{x}\in D[0,\infty)$ is such that $\boldsymbol{x}(0)\leq\alpha$. Then there exist unique $\boldsymbol{y},\boldsymbol{z}\in D[0,\infty)$ such that the following statements hold.*

(a) *We have $\boldsymbol{z}(t)=\boldsymbol{x}(t)-\boldsymbol{y}(t)\leq\alpha$ for all $t\geq 0$.*

(b) *The function $\boldsymbol{y}$ is nondecreasing, thus absolutely continuous, and $\boldsymbol{y}(0)=0$.*

(c) *The function $\boldsymbol{y}$ is flat off $\{t\geq 0:\boldsymbol{z}(t)=\alpha\}$, i.e., $\dot{\boldsymbol{y}}(t)\,\mathbb{1}_{\{\boldsymbol{z}(t)<\alpha\}}=0$ almost everywhere.*

The map $(\Psi_\alpha,\Phi_\alpha)$ such that $\Psi_\alpha(\boldsymbol{x})=\boldsymbol{y}$ and $\Phi_\alpha(\boldsymbol{x})=\boldsymbol{z}$ is called the one-dimensional Skorokhod mapping with upper reflecting barrier at α and satisfies

$$\Psi_\alpha(\boldsymbol{x})(t)=\sup_{s\in[0,t]}[\boldsymbol{x}(s)-\alpha]^+ \quad \textit{and} \quad \Phi_\alpha(\boldsymbol{x})(t)=\boldsymbol{x}(t)-\Psi_\alpha(\boldsymbol{x})(t). \tag{24}$$

In addition, if $\boldsymbol{x},\boldsymbol{y}\in D[0,\infty)$ are any two functions such that $\boldsymbol{x}(0),\boldsymbol{y}(0)\leq\alpha$, then for each $T\geq 0$ we have the following Lipschitz properties:

$$\|\Psi_\alpha(\boldsymbol{x})-\Psi_\alpha(\boldsymbol{y})\|_T\leq\|\boldsymbol{x}-\boldsymbol{y}\|_T \quad \textit{and} \quad \|\Phi_\alpha(\boldsymbol{x})-\Phi_\alpha(\boldsymbol{y})\|_T\leq 2\|\boldsymbol{x}-\boldsymbol{y}\|_T.$$

Consider càdlàg functions $\boldsymbol{x}:[0,\infty)\to\mathbb{R}^{\mathcal{I}}$ such that $\boldsymbol{x}(0,i,j)\leq\alpha(i)$ for all $(i,j)\in\mathcal{I}_+$ and families of càdlàg functions $\boldsymbol{v}_k:[0,\infty)\to\mathbb{R}$ such that $\boldsymbol{v}_k(0)=0$ for all $k\in\mathbb{N}$. Define for each $k\geq 1$ a mapping Θ_k as follows:

$$\Theta_k(\boldsymbol{x},\boldsymbol{v})(t):=\boldsymbol{x}(0,i_k,j_k)+\boldsymbol{v}_{k-1}(t)-\int_0^t\mu j_k\big[\boldsymbol{x}(s,i_k,j_k)-\boldsymbol{x}(s,i_k,j_k+1)\big]ds;$$

here recall the enumeration of $\mathcal{I}_+$ introduced in Section 3.2. The next lemma establishes that $(\boldsymbol{q},\boldsymbol{w})$ satisfies the following set of conditions if $\boldsymbol{q}$ is a fluid trajectory and $\boldsymbol{w}$ is defined suitably in terms of $\boldsymbol{q}$:

$$\boldsymbol{v}_k=\Psi_{\alpha(i_k)}[\Theta_k(\boldsymbol{x},\boldsymbol{v})] \quad \text{for all} \quad k\geq 1, \tag{25a}$$

$$\boldsymbol{x}(i_k,j_k)=\Phi_{\alpha(i_k)}[\Theta_k(\boldsymbol{x},\boldsymbol{v})] \quad \text{for all} \quad k\geq 1, \tag{25b}$$

$$\boldsymbol{v}_0(t)=\lambda t \quad \text{for all} \quad t\geq 0, \tag{25c}$$

$$\boldsymbol{x}(t,i,0)=\alpha(i) \quad \text{for all} \quad t\geq 0 \quad \text{and} \quad i. \tag{25d}$$

Lemma 4. *Let $\boldsymbol{q}$ be a fluid trajectory and define $\boldsymbol{r}$ such that*

$$\sigma\big(\boldsymbol{q}(t)\big)=\big(i_{\boldsymbol{r}(t)},j_{\boldsymbol{r}(t)}\big) \quad \textit{for all} \quad t\geq 0.$$

Also, consider the absolutely continuous functions $\boldsymbol{w}_k$ such that $\boldsymbol{w}_k(0)=0$ and

$$\dot{\boldsymbol{w}}_k(t)=\left[\lambda-\sum_{s=1}^{k}\mu j_s\big[\alpha(i_s)-\boldsymbol{q}(t,i_s,j_s+1)\big]\right]\mathbb{1}_{\{k<\boldsymbol{r}(t)\}} \quad \textit{for all} \quad k\geq 0.$$

Then $(\boldsymbol{q},\boldsymbol{w})$ satisfies (25).

Proof. It is clear that $(\boldsymbol{q},\boldsymbol{w})$ satisfies (25c) and (25d), so we only need to verify that (25a) and (25b) hold as well. By Lemma 3, it is enough to check the following properties.

(a) We have $\boldsymbol{q}(t,i_k,j_k) = \Theta_k(\boldsymbol{q},\boldsymbol{w})(t) - \boldsymbol{w}_k(t) \le \alpha(i_k)$ for all $t \ge 0$ and $k \ge 1$.
(b) We have $\boldsymbol{w}_k(0) = 0$ and $\dot{\boldsymbol{w}}_k(t) \ge 0$ almost everywhere for all $k \ge 1$.
(c) We have $\dot{\boldsymbol{w}}_k(t)\, \mathbb{1}_{\{\boldsymbol{q}(t,i_k,j_k)<\alpha(i_k)\}} = 0$ almost everywhere for all $k \ge 1$.

To establish (a), it suffices to show that

$$\dot{\boldsymbol{q}}(i_k,j_k) = \dot{\Theta}_k(\boldsymbol{q},\boldsymbol{w}) - \dot{\boldsymbol{w}}_k = \dot{\boldsymbol{w}}_{k-1} - \dot{\boldsymbol{w}}_k - \mu j_k\big[\boldsymbol{q}(i_k,j_k) - \boldsymbol{q}(i_k,j_k+1)\big] \tag{26}$$

holds almost everywhere. Indeed, (a) holds at $t = 0$ and $\boldsymbol{q}(i_k,j_k) \le \alpha(i_k)$ at all times. Because $\dot{\boldsymbol{w}}_{k-1} - \dot{\boldsymbol{w}}_k = \Lambda(\boldsymbol{q},i_k,j_k)$ and $\boldsymbol{q}$ is a fluid trajectory, we obtain (26) from (12a).

Property (b) is a consequence of the definition of $\boldsymbol{w}_k$ and (12b), and (c) follows from the following observation. If $\boldsymbol{q}(t,i_k,j_k) < \alpha(i_k)$, then $(i_k,j_k) \trianglelefteq \sigma\big(\boldsymbol{q}(t)\big)$ and thus $k \ge \boldsymbol{r}(t)$, which implies that $\dot{\boldsymbol{w}}_k(t) = 0$. □

Next we prove Theorem 3. As noted earlier, the proof relies on the Lipschitz property of the Skorokhod reflection mappings Ψ_α and Φ_α. In addition, a uniqueness of solutions result for certain Kolmogorov backward equations is used.

Proof of Theorem 3. Suppose that there exist two fluid trajectories $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $\boldsymbol{x}(0) = \boldsymbol{y}(0) = q$. Define $\boldsymbol{v}$ in terms of $\boldsymbol{x}$ and $\boldsymbol{w}$ in terms of $\boldsymbol{y}$ as in Lemma 4. It follows from the same lemma that $(\boldsymbol{x},\boldsymbol{v})$ and $(\boldsymbol{y},\boldsymbol{w})$ satisfy (25). Next we fix $T > 0$ and we prove that $\boldsymbol{x}(t) = \boldsymbol{y}(t)$ and $\boldsymbol{v}(t) = \boldsymbol{w}(t)$ for all $t \in [0,T]$.

As a first step, we demonstrate that there exists $M > 0$ such that

$$\boldsymbol{v}_k(t) = \boldsymbol{w}_k(t) = 0 \quad \text{for all} \quad t \in [0,T] \quad \text{and} \quad k \ge M. \tag{27}$$

Because $\dot{\boldsymbol{v}}_{k-1} - \dot{\boldsymbol{v}}_k = \Lambda(\boldsymbol{x},i_k,j_k) \ge 0$, we conclude that $\dot{\boldsymbol{v}}_k \le \dot{\boldsymbol{v}}_{k-1} \le \dot{\boldsymbol{v}}_0 = \lambda$ almost everywhere and for all $k \ge 1$. It follows from (12a), or equivalently from (25a) and (25b), that the following inequalities hold almost everywhere:

$$\dot{\boldsymbol{x}}(i_k,j_k) \le \dot{\boldsymbol{v}}_{k-1} - \dot{\boldsymbol{v}}_k \le \dot{\boldsymbol{v}}_{k-1} \le \lambda \quad \text{for all} \quad k \ge 1. \tag{28}$$

Define $\alpha_{\min} := \min\{\alpha(i) : 1 \le i \le m\}$ and note that there exist $\varepsilon > 0$ and k_0 such that $q(i_k,j_k) \le \varepsilon < \alpha_{\min}$ for all $k \ge k_0$ because $q \in Q \subset \ell_1$. Property (c) of Lemma 3 implies that $\boldsymbol{v}_k$ is zero until $\boldsymbol{x}(i_k,j_k)$ reaches $\alpha(i_k) \ge \alpha_{\min}$ for the first time. Using this remark and (28), it is possible to prove by induction on $k \ge k_0$ that

$$\boldsymbol{v}_k(t) = 0 \quad \text{for all} \quad t \in \left[0,(k+1-k_0)\frac{\alpha_{\min}-\varepsilon}{\lambda}\right).$$

The same property holds if $(\boldsymbol{x},\boldsymbol{v})$ is replaced by $(\boldsymbol{y},\boldsymbol{w})$; thus, (27) holds.

Next we show that $\boldsymbol{x}(t,i,j) = \boldsymbol{y}(t,i,j)$ for all $t \in [0,T]$ and $(i,j) \triangleleft (i_M,j_M)$. For this purpose, fix an arbitrary i and let $J_i := \min\big\{j \ge 1 : (i,j) \triangleleft (i_M,j_M)\big\}$. By (27), both $\boldsymbol{x}(i)$ and $\boldsymbol{y}(i)$ satisfy the following initial value problem:

$$\boldsymbol{z}.(j) = -\mu j\big[\boldsymbol{z}(j) - \boldsymbol{z}(j+1)\big] \quad \text{and} \quad \boldsymbol{z}(0,j) = q(i,j) \quad \text{for all} \quad j \ge J_i. \tag{29}$$

The previous system of differential equations are the backward Kolmogorov equations of the pure birth process with state space $E_i := \big\{j \in \mathbb{N} : j \ge J_i\big\}$ that has birth rate $\lambda_j := \mu j$ at state j. This process is nonexplosive because $\sum_{j \ge J_i} 1/\lambda_j = \infty$; hence, it follows from Fontes (1989) that the initial value problem (29) has a unique solution $\boldsymbol{z}$ such that $\boldsymbol{z}$ is bounded on $[0,t] \times E_i$ for all $t \ge 0$. Both $\boldsymbol{x}(i)$ and $\boldsymbol{y}(i)$ satisfy the latter condition because fluid trajectories take values in Q; thus, $\boldsymbol{x}(t,i,j) = \boldsymbol{y}(t,i,j)$ for all $t \ge 0$ and $j \ge J_i$.

We conclude by proving that $\boldsymbol{x}(i_k,j_k) = \boldsymbol{y}(i_k,j_k)$ and $\boldsymbol{v}_k = \boldsymbol{w}_k$ along the interval $[0,T]$ for all $k \le M$. Let $\boldsymbol{f}(i,j) := \boldsymbol{x}(i,j) - \boldsymbol{y}(i,j)$ and $\boldsymbol{g}_k := \boldsymbol{v}_k - \boldsymbol{w}_k$. The subsequent arguments are analogous to those in (Bhamidi et al. 2022, section 4.1).

For all $t \in [0,T]$ and $k \le M$, we have

$$\|\boldsymbol{f}(i_k,j_k)\|_t \le 2\|\boldsymbol{g}_{k-1}\|_t + 2\mu j_k \int_0^t \|\boldsymbol{f}(i_k,j_k)\|_s ds + 2\mu j_k \int_0^t \|\boldsymbol{f}(i_k,j_k+1)\|_s ds,$$
$$\|\boldsymbol{g}_k\|_t \le \|\boldsymbol{g}_{k-1}\|_t + \mu j_k \int_0^t \|\boldsymbol{f}(i_k,j_k)\|_s ds + \mu j_k \int_0^t \|\boldsymbol{f}(i_k,j_k+1)\|_s ds;$$

these inequalities follow from the Lipschitz properties of $\Psi_{\alpha(i_k)}$ and $\Phi_{\alpha(i_k)}$. Let us define $J := \max\big\{j_k : k \le M\big\}$ and $\boldsymbol{h}(t) := \max\big\{\|\boldsymbol{f}(i_k,j_k)\|_t : k \le M\big\}$, then

$$\|\boldsymbol{g}_k\|_t \le \|\boldsymbol{g}_{k-1}\|_t + 2\mu J \int_0^t \boldsymbol{h}(s) ds \le 2k\mu J \int_0^t \boldsymbol{h}(s) ds$$

for all $t \in [0,T]$ and $k \le M$. For the first inequality, $\boldsymbol{f}(i_k, j_k + 1)$ is identically zero along the interval $[0,T]$ if $(i_k, j_k + 1) = (i_l, j_l)$ for some $l > M$, and for the last inequality, observe that $\boldsymbol{g}_0$ is identically zero by (25c). In addition, we have

$$\|\boldsymbol{f}(i_k, j_k)\|_t \le 4k\mu J \int_0^t \boldsymbol{h}(s)ds + 4\mu J \int_0^t \boldsymbol{h}(s)ds \le 4(M+1)\mu J \int_0^t \boldsymbol{h}(s)ds$$

for all $t \in [0,T]$ and $k \le M$. We conclude that

$$\boldsymbol{h}(t) \le 4(M+1)\mu J \int_0^t \boldsymbol{h}(s)ds \quad \text{for all} \quad t \in [0,T].$$

Therefore, Grönwall's inequality yields $\boldsymbol{h}(t) = 0$ for all $t \in [0,T]$, and this in turn implies that $(\boldsymbol{x}, \boldsymbol{v}) = (\boldsymbol{y}, \boldsymbol{w})$ along the interval $[0,T]$. □

We conclude this section by establishing that (12) has a unique equilibrium point and that all fluid trajectories converge to this equilibrium point over time.

Proof of Theorem 4. First, we verify that q_* is an equilibrium of (12). To this end, $\sigma(q_*) = \sigma_*$, and the right-hand side of (12a) equals zero if $(i,j) \ne \sigma_*$ and $\boldsymbol{q} = q_*$. It only remains to be shown that this also holds for $(i_*, j_*) := \sigma_*$.

If $(i,j) = \sigma(q)$, then

$$\Lambda(q,i,j) - \mu j\big[q(i,j) - q(i,j+1)\big] = \lambda - \sum_{(r,s) \trianglerighteq (i,j)} \mu s\big[q(r,s) - q(r,s+1)\big].$$

Define $J^i := \max\{j \ge 1 : (i,j) \trianglerighteq \sigma_*\}$ for each i. If $(i,j) = (i_*, j_*)$ and $\boldsymbol{q}$ is replaced by q_*, then the right-hand side of (12a) equals

$$\begin{aligned}
\lambda - \sum_{(i,j) \trianglerighteq \sigma_*} \mu j\big[q_*(i,j) - q_*(i,j+1)\big] &= \lambda - \sum_{i=1}^{m} \sum_{j=1}^{J^i} \mu j\big[q_*(i,j) - q_*(i,j+1)\big] \\
&= \lambda - \sum_{i \ne i_*} \mu J^i \alpha(i) - \mu(j_* - 1)\alpha(i_*) - \mu q_*(i_*, j_*) \\
&= \sum_{(i,j) \triangleright \sigma_*} \mu \alpha(i) - \sum_{i \ne i_*} \mu J^i \alpha(i) - \mu(j_* - 1)\alpha(i_*).
\end{aligned}$$

The expression in the last line equals zero by definition of J^i, and therefore we conclude that q_* is indeed an equilibrium point of (12).

Next, we prove that all fluid trajectories converge coordinatewise to q_* over time; this implies, in particular, that q_* is the unique equilibrium of (12). Afterward we prove that all fluid trajectories in fact converge to q_* in ℓ_1.

If $\boldsymbol{q}$ is a fluid trajectory and $T \ge 0$, then there exists j_T such that $\Lambda(\boldsymbol{q}, i, j) = 0$ in $[0,T]$ for all i and $j \ge j_T$. This can be established directly from (12), but also using Lemma 2, Theorem 5, and the uniqueness of solutions. For each i and $k \ge j_T$,

$$\sum_{j=j_T}^{k} \dot{\boldsymbol{q}}(i,j) = -\sum_{j=j_T}^{k} \mu j\big[\boldsymbol{q}(i,j) - \boldsymbol{q}(i,j+1)\big] = -\mu(j_T - 1)\boldsymbol{q}(i,j_T) - \sum_{j=j_T}^{k} \boldsymbol{q}(i,j) + \mu k \boldsymbol{q}(i,k+1).$$

The last term vanishes as $k \to \infty$ because $\boldsymbol{q}$ takes values in ℓ_1. Also, $\boldsymbol{q}(i,j)$ does not increase in $[0,T]$ for all $j \ge j_T$ because $\Lambda(\boldsymbol{q}, i, j) = 0$ along $[0,T]$. Therefore,

$$\begin{aligned}
\left| \sum_{j=j_T}^{k} \dot{\boldsymbol{q}}(t,i,j) - \sum_{j=j_T}^{\infty} \dot{\boldsymbol{q}}(t,i,j) \right| &= \sum_{j=k+1}^{\infty} \boldsymbol{q}(t,i,j) + \mu k \boldsymbol{q}(t,i,k+1) \\
&\le \sum_{j=k+1}^{\infty} \boldsymbol{q}(0,i,j) + \mu k \boldsymbol{q}(0,i,k+1) \quad \text{for all} \quad t \in [0,T].
\end{aligned}$$

The right-hand side vanishes as $k \to \infty$ because $\boldsymbol{q}(0) \in \ell_1$; thus, the left-hand side converges uniformly to zero over $[0,T]$ and, by Rudin (1976, theorem 7.17), the derivative of

$$\sum_{j=j_T}^{\infty} \boldsymbol{q}(t,i,j) \quad \text{is} \quad \sum_{j=j_T}^{\infty} \dot{\boldsymbol{q}}(t,i,j) \quad \text{for all} \quad i \quad \text{and} \quad t \in (0,T).$$

This allows for the interchanges of summation and differentation that appear later.

Consider the function

$$f := \sum_{(i,j)\in\mathcal{I}_+} q(i,j).$$

It follows from (12a) that $\dot{f} = \lambda - \mu f$. Thus, $f(t) = \rho + [f(0) - \rho]e^{-\mu t}$ for all $t \geq 0$. We conclude from this identity and (10) that

$$t_0 := \min\left\{ t \geq 0 : f(t) \leq \sum_{(i,j)\trianglerighteq\sigma_*} \alpha(i) \right\}$$

exists and is finite. If $t > t_0$, then the inequality inside the minimum sign is strict, and this implies that $\sigma(q(t)) \trianglerighteq \sigma_*$ because $q(t,i,j) = \alpha(i)$ for all $(i,j) \triangleright \sigma(q(t))$.

Consider the function

$$g := \sum_{(i,j)\trianglelefteq\sigma_*} q(i,j).$$

As noted previously, if $t > t_0$, then $\sigma(q(t)) \trianglerighteq \sigma_*$, and thus we have

$$\begin{aligned}\dot{g}(t) &= \left[\lambda - \sum_{(i,j)\triangleright\sigma_*} \mu j\big[q(t,i,j) - q(t,i,j+1)\big]\right] \mathbb{1}_{\{\sigma(q(t))=\sigma_*\}} \\ &\quad - \sum_{(i,j)\trianglelefteq\sigma_*} \mu j\big[q(t,i,j) - q(t,i,j+1)\big],\end{aligned}$$

because $\Lambda(q(t),i,j) = 0$ for all $(i,j) \trianglelefteq \sigma_*$ except perhaps for $(i,j) = \sigma_*$. Hence,

$$\begin{aligned}\dot{g}(t) &= \left[\lambda - \sum_{(i,j)\in\mathcal{I}_+} \mu j\big[q(t,i,j) - q(t,i,j+1)\big]\right] \mathbb{1}_{\{\sigma(q(t))=\sigma_*\}} \\ &\quad - \left[\sum_{(i,j)\trianglelefteq\sigma_*} \mu j\big[q(t,i,j) - q(t,i,j+1)\big]\right] \mathbb{1}_{\{\sigma(q(t))\neq\sigma_*\}} \\ &= \left[\lambda - \sum_{(i,j)\in\mathcal{I}_+} \mu q(t,i,j)\right] \mathbb{1}_{\{\sigma(q(t))=\sigma_*\}} \\ &\quad - \left[\sum_{i=1}^{m}\sum_{j=J_i}^{\infty} \mu j\big[q(t,i,j) - q(t,i,j+1)\big]\right] \mathbb{1}_{\{\sigma(q(t))\neq\sigma_*\}},\end{aligned}$$

where $J_i := \min\{j \geq 1 : (i,j) \trianglelefteq \sigma_*\}$. Note that $q(t,i,j) = \alpha(i)$ if $(i,j) \triangleright \sigma(q(t))$. Thus,

$$\begin{aligned}\dot{g}(t) &= \left[\lambda - \mu \sum_{(i,j)\triangleright\sigma_*} \alpha(i) - \mu g(t)\right] \mathbb{1}_{\{\sigma(q(t))=\sigma_*\}} \\ &\quad - \left[\sum_{i=1}^{m} \left(\mu(J_i - 1)q(t,i,J_i) + \sum_{j=J_i}^{\infty} \mu q(t,i,j)\right)\right] \mathbb{1}_{\{\sigma(q(t))\neq\sigma_*\}} \\ &= \left[\lambda - \mu \sum_{(i,j)\triangleright\sigma_*} \alpha(i) - \mu g(t)\right] \mathbb{1}_{\{\sigma(q(t))=\sigma_*\}} \\ &\quad - \mu g(t)\, \mathbb{1}_{\{\sigma(q(t))\neq\sigma_*\}} - \mu \sum_{i=1}^{m} (J_i - 1)q(t,i,J_i) \mathbb{1}_{\{\sigma(q(t))\neq\sigma_*\}} \\ &\leq \left[\lambda - \mu \sum_{(i,j)\triangleright\sigma_*} \alpha(i)\right]^+ - \mu g(t).\end{aligned}$$

By definition of σ_*, we have $\theta := \sum_{(i,j)\triangleright\sigma_*}\alpha(i) \le \rho$, and from the above bound for $\dot{\boldsymbol{g}}$, we get

$$\boldsymbol{g}(t) \le \rho - \theta + \big[\boldsymbol{g}(t_0) - (\rho-\theta)\big]\mathrm{e}^{-\mu(t-t_0)} \quad \text{for all} \quad t \ge t_0.$$

We conclude that

$$\liminf_{t\to\infty}\sum_{(i,j)\triangleright\sigma_*}\boldsymbol{q}(t,i,j) = \liminf_{t\to\infty}\big[\boldsymbol{f}(t)-\boldsymbol{g}(t)\big] \ge \rho-(\rho-\theta) = \sum_{(i,j)\triangleright\sigma_*}\alpha(i).$$

This proves that $\boldsymbol{q}(t,i,j) \to \alpha(i) = q_*(i,j)$ over time for all $(i,j)\triangleright\sigma_*$.

Consider now the function

$$\boldsymbol{h} := \sum_{(i,j)\triangleleft\sigma_*}\boldsymbol{q}(i,j).$$

Recall that $\sigma\big(\boldsymbol{q}(t)\big)\trianglerighteq\sigma_*$ for all $t > t_0$. Therefore,

$$\dot{\boldsymbol{h}}(t) = -\sum_{(i,j)\triangleleft\sigma_*}\mu j\big[\boldsymbol{q}(t,i,j)-\boldsymbol{q}(t,i,j+1)\big] \le -\mu\boldsymbol{h}(t) \quad \text{for all} \quad t > t_0.$$

It follows that $\boldsymbol{h}(t) \le \boldsymbol{h}(t_0)\mathrm{e}^{-\mu(t-t_0)}$ for all $t \ge t_0$, and thus $\boldsymbol{q}(t,i,j) \to 0 = q_*(i,j)$ over time for all $(i,j)\triangleleft\sigma_*$. Furthermore, we have

$$\lim_{t\to\infty}\boldsymbol{q}(t,\sigma_*) = \lim_{t\to\infty}\left[\boldsymbol{f}(t)-\boldsymbol{h}(t)-\sum_{(i,j)\triangleright\sigma_*}\boldsymbol{q}(t,i,j)\right] = \rho - \sum_{(i,j)\triangleright\sigma_*}\alpha(i) = q_*(\sigma_*),$$

and thus $\boldsymbol{q}(t,i,j) \to q_*(i,j)$ for all $(i,j)\in\mathcal{I}$.

Finally, observe that

$$\lim_{t\to\infty}\sum_{(i,j)\triangleleft\sigma_*}\big|\boldsymbol{q}(t,i,j)-q_*(i,j)\big| = \lim_{t\to\infty}\boldsymbol{h}(t) = 0.$$

Consequently, $\boldsymbol{q}(t) \to q_*$ over time not only coordinatewise but also in ℓ_1. □

6.2. Proof of the Fluid Limit

To prove Theorem 5, it suffices to demonstrate, for each $\omega\in\Gamma$, that every subsequence of $\{\boldsymbol{q}_n(\omega) : n \ge 1\}$ has a further subsequence with a limit in $D_{\ell_1}[0,\infty)$ and that this limit is the unique fluid trajectory starting at $q_0(\omega)$. The first part is covered by Proposition 3, every subsequence of $\{\boldsymbol{q}_n(\omega) : n \ge 1\}$ has a further subsequence with a limit in $D_{\ell_1}[0,\infty)$. Next, we characterize the limits of the convergent subsequences.

Let us fix an arbitrary $\omega\in\Gamma$, which we omit from the notation for brevity, and an increasing sequence $\mathcal{K}\subset\mathbb{N}$ such that $\{\mathcal{A}_k : k\in\mathcal{K}\}$, $\{\mathcal{D}_k : k\in\mathcal{K}\}$ and $\{q_k : k\in\mathcal{K}\}$ converge in $D_{\ell_1}[0,\infty)$ to certain functions $\boldsymbol{a}$, $\boldsymbol{d}$ and $\boldsymbol{q}$, respectively, which have locally Lipschitz coordinates by Proposition 2. To characterize these three limits, it suffices to just characterize $\boldsymbol{a}$ and $\boldsymbol{d}$ because (19) and (22a) imply that

$$\boldsymbol{q} = q_0 + \boldsymbol{a} - \boldsymbol{d}. \tag{30}$$

Because $\boldsymbol{a}$ and $\boldsymbol{d}$ have locally Lipschitz coordinates, there exists $\mathcal{R}\subset(0,\infty)$ such that $\mathcal{R}^c$ has zero Lebesgue measure and the derivatives of $\boldsymbol{a}(i,j)$ and $\boldsymbol{d}(i,j)$ exist for all $(i,j)\in\mathcal{I}$ at all points in $\mathcal{R}$. These derivatives are zero if $j=0$ by the definitions of $\mathcal{A}_k$ and $\mathcal{D}_k$. The following lemma computes the derivatives for $(i,j)\in\mathcal{I}_+$.

Lemma 5. *Fix an arbitrary $t_0\in\mathcal{R}$, we have*

$$\dot{\boldsymbol{d}}(t_0,i,j) = \mu j\big[\boldsymbol{q}(t_0,i,j)-\boldsymbol{q}(t_0,i,j+1)\big] \quad \textit{for all} \quad (i,j)\in\mathcal{I}_+. \tag{31}$$

Furthermore, $\boldsymbol{q}(t_0)$ and the derivatives $\dot{\boldsymbol{a}}(t_0,i,j)$ satisfy

$$\Lambda(\boldsymbol{q}(t_0),i,j) = \dot{\boldsymbol{a}}(t_0,i,j) \ge 0 \quad \textit{for all} \quad (i,j)\in\mathcal{I}_+. \tag{32}$$

Proof. The sequences $\{\mathcal{D}_k(i,j) : k\in\mathcal{K}\}$ and $\{\boldsymbol{q}_k(i,j) : k\in\mathcal{K}\}$ converge uniformly over compact sets to $\boldsymbol{d}(i,j)$ and $\boldsymbol{q}(i,j)$, respectively, for all $(i,j)\in\mathcal{I}$. This remark, the definition of $\mathcal{D}_k$ and (22c) imply that

$$\boldsymbol{d}(t,i,j) = \int_0^t \mu j\big[\boldsymbol{q}(s,i,j)-\boldsymbol{q}(s,i,j+1)\big]ds \quad \text{for all} \quad t \ge 0.$$

It is clear that this identity establishes (31).

We now prove that

$$\sum_{(i,j)\in\mathcal{I}_+} \dot{\boldsymbol{a}}(t_0,i,j) = \lambda \quad \text{and} \quad \dot{\boldsymbol{a}}(t_0,i,j) \geq 0 \quad \text{for all} \quad (i,j)\in\mathcal{I}_+. \tag{33}$$

The derivatives $\dot{\boldsymbol{a}}(t_0,i,j)$ are nonnegative because the processes $\mathcal{A}_k(i,j)$ are nondecreasing, so we only need to show that the derivatives add up to λ. For this purpose, note that

$$\sum_{(i,j)\in\mathcal{I}_+} \mathcal{A}_k(t,i,j) = \mathcal{N}_k^{\lambda}(t) \quad \text{for all} \quad t\geq 0 \quad \text{and} \quad k\in\mathcal{K}.$$

Fix $T > t_0$, and let j_T and n_T be as in Lemma 2. The left-hand side has at most mj_T nonzero terms for all $k \geq n_T$ and $t \in [0,T]$. It follows from (22b) that

$$\sum_{(i,j)\in\mathcal{I}_+} \boldsymbol{a}(t,i,j) = \lambda t \quad \text{for all} \quad t\in[0,T].$$

This yields (33) because the left-hand side has at most mj_T nonzero terms.

It follows from (30) that the derivative of $\boldsymbol{q}(i,j)$ exists at t_0 and

$$\dot{\boldsymbol{q}}(t_0,i,j) = \dot{\boldsymbol{a}}(t_0,i,j) - \dot{\boldsymbol{d}}(t_0,i,j) \quad \text{for all} \quad (i,j)\in\mathcal{I}.$$

Note that $\boldsymbol{q}(i,j)$ is upper bounded by $\alpha(i)$, so $\boldsymbol{q}(t_0,i,j) = \alpha(i)$ implies $\dot{\boldsymbol{q}}(t_0,i,j) = 0$. Thus,

$$\dot{\boldsymbol{a}}(t_0,i,j) = \dot{\boldsymbol{d}}(t_0,i,j) = \mu j[\alpha(i) - \boldsymbol{q}(t_0,i,j+1)] \quad \text{if} \quad \boldsymbol{q}(t_0,i,j) = \alpha(i). \tag{34}$$

To prove the equality in (32), define $\sigma_0 = (i_0,j_0) := \sigma\big(\boldsymbol{q}(t_0)\big)$. If $(i,j)\triangleright\sigma_0$, then $\boldsymbol{q}(t_0,i,j) = \boldsymbol{q}(t_0,i,j-1)$ by (8). Moreover, for a fixed i, the marginal utility $\Delta(i,j)$ does not increase with j because u_i is a concave function. Therefore, $(i,j)\triangleright\sigma_0$ and $j > 1$ imply that $(i,j-1)\triangleright\sigma_0$. We conclude that

$$\boldsymbol{q}(t_0,i,j) = \boldsymbol{q}(t_0,i,0) = \alpha(i) \quad \text{for all} \quad (i,j)\triangleright\sigma_0.$$

The last property and (34) imply that (32) holds if $(i,j)\triangleright\sigma_0$.

Note that $\boldsymbol{q}(t_0,\sigma_0) < \alpha(i_0)$ by (8). Because $\boldsymbol{q}(\sigma_0)$ is continuous and $\boldsymbol{q}_k(\sigma_0)$ converges uniformly over compact sets to $\boldsymbol{q}(\sigma_0)$, there exist $\varepsilon > 0$ and $k_\varepsilon \in \mathcal{K}$ such that

$$\boldsymbol{q}_k(t,\sigma_0) < \alpha_k(i_0) \quad \text{for all} \quad t\in(t_0-\varepsilon,t_0+\varepsilon) \quad \text{and} \quad k\geq k_\varepsilon.$$

It follows from (8) and the last statement that

$$\sigma_0 \trianglelefteq \sigma\big(\boldsymbol{q}_k(t)\big) \quad \text{for all} \quad t\in(t_0-\varepsilon,t_0+\varepsilon) \quad \text{and} \quad k\geq k_\varepsilon.$$

Thus, $\mathcal{A}_k(i,j)$ is constant over $(t_0-\varepsilon,t_0+\varepsilon)$ if $k\geq k_\varepsilon$ and $(i,j)\triangleleft\sigma_0$. Indeed, server pools of class i with exactly $j-1$ tasks are not assigned incoming tasks in the system with k server pools if $(i,j)\triangleleft\sigma\big(\boldsymbol{q}_k\big)$. This proves (32) for $(i,j)\triangleleft\sigma_0$, and we conclude from (33) that (32) must also hold in the case $(i,j) = \sigma_0$. □

Here we complete the proof of Theorem 5.

Proof of Theorem 5. As previously, we fix some $\omega\in\Gamma$ that we omit from the notation. Every subsequence of $\{\boldsymbol{q}_n : n\geq 1\}$ has a further subsequence that converges in $D_{\ell_1}[0,\infty)$ by Proposition 3. It follows from (22a) and Lemma 5 that the limit $\boldsymbol{q}$ of this convergent subsequence is a fluid trajectory with $\boldsymbol{q}(0) = q_0$, which determines $\boldsymbol{q}$ by Theorem 3. □

7. Limiting Behavior of SLTA

In this section, we assume that the load balancing policy is SLTA, and we leverage a methodology developed in Goldsztajn et al. (2021a) to prove Theorem 6. The first steps of the proof are carried out in Section 7.1, where we establish that certain dynamical properties of the system hold asymptotically with probability one. The proof is completed in Section 7.2, where we analyze the evolution of the learning scheme over time. Although the arguments used here are more involved due to the heterogeneity of the system, most of the proofs are conceptually similar to those in Goldsztajn et al. (2021a) and hence are deferred to Online Appendix C.

7.1. Asymptotic Dynamical Properties

In this section, we establish asymptotic dynamical properties pertaining to the total and tail mass processes, which are defined as

$$\boldsymbol{s}_n := \sum_{(i,j)\in\mathcal{I}_+} \boldsymbol{q}_n(i,j) \quad \text{and} \quad \boldsymbol{v}_n(r) := \sum_{(i,j)\trianglelefteq(i_r,j_r)} \boldsymbol{q}_n(i,j) \quad \text{for all} \quad r\geq 1,$$

respectively. Recall that the total mass process was introduced in (1) and represents the total number of tasks in the system, normalized by the number of server pools. The following proposition is proved in Online Appendix C.

Proposition 4. *For each $\omega \in \Gamma$, the sequence $\{\boldsymbol{s}_n(\omega) : n \geq 1\}$ converges uniformly over compact sets to the unique function $\boldsymbol{s}(\omega)$ such that*

$$\dot{\boldsymbol{s}}(\omega) = \lambda - \mu \boldsymbol{s}(\omega) \quad \text{and} \quad \boldsymbol{s}(\omega, 0) = s_0(\omega), \tag{35}$$

where s_0 is as defined in (13). Explicitly, $\boldsymbol{s}(\omega, t) = \rho + \big[s_0(\omega) - \rho\big]\mathrm{e}^{-\mu t}$.

Although the previous law of large numbers is known to hold weakly, it is not straightforward that it holds with probability one under the coupled construction of sample paths adopted in Section 5.1; this fact is established in Proposition 4.

The next result is also proved in Online Appendix C, and it provides an asymptotic upper bound for certain tail mass processes, under specific conditions concerning $\boldsymbol{q}_n$ and $\boldsymbol{r}_n$. The upper bound implies at least an exponentially fast decay over time.

Proposition 5. *Suppose that the next conditions hold for a given $\omega \in \Gamma$ and a given increasing sequence $\mathcal{K}$ of natural numbers.*

(a) *The sequence $\big\{\boldsymbol{q}_k(\omega) : k \in \mathcal{K}\big\}$ converges in $D_{\ell_1}[0, \infty)$ to some function $\boldsymbol{q}$.*

(b) *There exist $r > 1$ and $0 \leq t_0 < t_1$ such that*

$$\boldsymbol{r}_k(\omega, t) \leq r \quad \text{and} \quad \sum_{(i,j) \trianglerighteq (i_r, j_r)} \boldsymbol{q}_k(\omega, t, i, j) < \sum_{(i,j) \trianglerighteq (i_r, j_r)} \alpha_k(i)$$

for all $t \in [t_0, t_1]$ and $k \in \mathcal{K}$.

Then $\boldsymbol{q}(i, j)$ is differentiable on (t_0, t_1) for all $(i, j) \trianglelefteq (i_r, j_r)$ and satisfies

$$\dot{\boldsymbol{q}}(t, i, j) = -\mu j\big[\boldsymbol{q}(t, i, j) - \boldsymbol{q}(t, i, j+1)\big] \quad \text{for all} \quad t \in (t_0, t_1).$$

Furthermore, the sequence of tail mass processes $\{\boldsymbol{v}_k(\omega, r) : k \in \mathcal{K}\}$ converges uniformly over compact sets to a function $\boldsymbol{v}(\omega, r)$ that satisfies

$$\boldsymbol{v}(\omega, t, r) < \boldsymbol{s}(\omega, t_0)\mathrm{e}^{-\mu(t - t_0)} \quad \text{for all} \quad t \in [t_0, t_1].$$

7.2. Evolution of the Learning Scheme

In this section, we complete the proof of Theorem 6. In Section 7.2.1, we establish that there exists a neighborhood of zero outside of which $\boldsymbol{r}_n$ is asymptotically upper bounded by r_* with a probability of one. This property partially proves (16a) and is used to obtain (16c). The proof of (16a) is finished in Section 7.2.2, where we also establish (16b).

7.2.1. Preliminary Results.

The following proposition states that $\boldsymbol{r}_n$ is asymptotically upper bounded by r_* outside of a neighborhood of zero with probability one; the proof is deferred to Online Appendix C.

Proposition 6. *There exists a function $\tau_{\mathrm{bd}} : [0, \infty) \to \mathbb{R}$ with the following property. If $\omega \in \Gamma$ and $T \geq \tau > \tau_{\mathrm{bd}}(s_0(\omega))$, then there exists $n_{\mathrm{bd}}^{\tau,T}(\omega)$ such that*

$$\boldsymbol{r}_n(\omega, t) \leq r_* \quad \text{for all} \quad t \in [\tau, T] \quad \text{and} \quad n \geq n_{\mathrm{bd}}^{\tau,T}(\omega).$$

The following corollary establishes (16c).

Corollary 1. *For each $\omega \in \Gamma$ and $T \geq \tau > \tau_{\mathrm{bd}}(s_0(\omega))$, we have*

$$\limsup_{n \to \infty} \sup_{t \in [\tau, T]} \boldsymbol{v}_n(\omega, t, r_* + 1)\mathrm{e}^{\mu(t-\tau)} \leq \boldsymbol{s}(\omega, \tau).$$

In particular, (16c) holds.

Proof. We fix $\omega \in \Gamma$ and $T \geq \tau > \tau_{\mathrm{bd}}(s_0(\omega))$, and we omit ω from the notation for brevity. Suppose that the statement of the corollary does not hold; then, there exist $\varepsilon > 0$ and an increasing sequence $\mathcal{K}$ of natural numbers such that

$$\sup_{t \in [\tau, T]} \boldsymbol{v}_k(t, r_* + 1)\mathrm{e}^{\mu(t-\tau)} > \boldsymbol{s}(\tau) + \varepsilon \quad \text{for all} \quad k \in \mathcal{K}.$$

By Propositions 3 and 6, we may assume that $\{\boldsymbol{q}_k : k \in \mathcal{K}\}$ has a limit in $D_{\ell_1}[0,\infty)$ and that $\boldsymbol{r}_k(t) \le r_*$ for all $t \in [\tau, T]$ and all $k \in \mathcal{K}$. The latter property implies that

$$\sum_{(i,j) \trianglerighteq (i_{r_*}, j_{r_*})} \boldsymbol{q}_k(t,i,j) < \sum_{(i,j) \trianglerighteq (i_{r_*}, j_{r_*})} \alpha_k(i) \quad \text{for all} \quad t \in [\tau, T],$$

because otherwise the number of tokens would be zero, which cannot occur by Remark 1. Therefore, Proposition 5 holds with $r = r_* + 1$ along the interval $[\tau, T]$, and in particular, there exists a function $\boldsymbol{v}(r_* + 1)$ such that

$$\boldsymbol{v}(t, r_* + 1) < \boldsymbol{s}(\tau) \mathrm{e}^{-\mu(t-\tau)} \quad \text{for all} \quad t \in [\tau, T],$$
$$\lim_{k \to \infty} \sup_{t \in [\tau, T]} |\boldsymbol{v}_k(t, r_* + 1) - \boldsymbol{v}(t, r_* + 1)| = 0.$$

This leads to a contradiction, so the statement of the corollary must hold. □

7.2.2. Proof of Theorem 6. Here we complete the proof of Theorem 6. For this purpose, let

$$\delta_n(t,r) := \frac{1}{n} \mathcal{N}_n^{\lambda}(t) - \lambda t - \sum_{(i,j) \triangleright (i_r, j_r)} \left[\mathcal{D}_n(t,i,j) - \int_0^t \mu j \big[\boldsymbol{q}_n(s,i,j) - \boldsymbol{q}_n(s,i,j+1) \big] ds \right] \tag{36}$$

for all $t \ge 0$ and $r \ge 1$. It follows from (22b) and (22c) that

$$\lim_{n \to \infty} \sup_{t \in [0,T]} n^{\gamma} |\delta_n(\omega, t, r)| = 0 \quad \text{for all} \quad \gamma \in [0, 1/2), \quad T \ge 0 \quad \text{and} \quad \omega \in \Gamma. \tag{37}$$

The following two technical lemmas are proved in Online Appendix C.

Lemma 6. *Fix $\omega \in \Gamma$, $T \ge 0$ and $r > 1$. Suppose that there exist an increasing sequence $\mathcal{K}$ of natural numbers and random times $0 \le \tau_{k,1} \le \tau_{k,2} \le T$ such that*

$$\boldsymbol{r}_k(\omega, t) \le r \quad \text{and} \quad \sum_{(i,j) \triangleright (i_r, j_r)} \boldsymbol{q}_k(\omega, t, i, j) < \sum_{(i,j) \trianglerighteq (i_r, j_r)} \alpha_k(i)$$

for all $t \in \big[\tau_{k,1}(\omega), \tau_{k,2}(\omega)\big)$ and $k \in \mathcal{K}$. Then

$$\sum_{(i,j) \trianglerighteq (i_r, j_r)} \boldsymbol{q}_k(\omega, t, i, j) - \sum_{(i,j) \trianglerighteq (i_r, j_r)} \boldsymbol{q}_k\big(\omega, \tau_{k,1}(\omega), i, j\big) \ge \big[t - \tau_{k,1}(\omega)\big] \left[\lambda - \mu \sum_{(i,j) \trianglerighteq (i_r, j_r)} \alpha_k(i) \right] - 2 \sup_{s \in [0,T]} |\delta_k(\omega, s, r)|$$

for all $t \in \big[\tau_{k,1}(\omega), \tau_{k,2}(\omega)\big]$ and $k \in \mathcal{K}$.

Lemma 7. *Fix $\omega \in \Gamma$, $T \ge 0$ and $1 \le r \le r_*$. Assume that there exist an increasing sequence $\mathcal{K}$ of natural numbers and random times $0 \le \zeta_{k,1} \le \zeta_{k,2} \le T$ such that*

$$\sum_{(i,j) \trianglerighteq (i_r, j_r)} \boldsymbol{q}_k\big(\omega, \zeta_{k,1}(\omega), i, j\big) = \sum_{(i,j) \trianglerighteq (i_r, j_r)} \alpha_k(i) \quad \text{and} \quad \boldsymbol{r}_k(\omega, t) \le r$$

for all $t \in \big[\zeta_{k,1}(\omega), \zeta_{k,2}(\omega)\big]$ and $k \in \mathcal{K}$. For each $\gamma \in [0, 1/2)$, we have

$$\boldsymbol{r}_k(\omega, t) = r \quad \text{and} \quad \boldsymbol{q}_k(\omega, t, i, j) \ge \alpha_k(i) - k^{-\gamma}$$

for all $(i,j) \triangleright (i_r, j_r)$, $t \in \big[\zeta_{k,1}(\omega), \zeta_{k,2}(\omega)\big]$ and all large enough $k \in \mathcal{K}$.

These lemmas are used to complete the proof of Theorem 6.

Proof of Theorem 6. We define τ_{eq} as follows:

$$\tau_{\mathrm{eq}}(s) := \tau_{\mathrm{bd}}(s) + \frac{1}{\mu} \log \left(\frac{\rho}{\rho - \sum_{(i,j) \trianglerighteq \sigma_*} \alpha(i)} \right) \quad \text{for all} \quad s \ge 0.$$

Fix $\omega \in \Gamma$ and $T \ge \tau > \tau_{\mathrm{eq}}(s_0(\omega))$ as in the statement of the theorem; we omit ω from the notation for brevity. Given $0 < \varepsilon < \rho - \sum_{(i,j) \trianglerighteq \sigma_*} \alpha(i)$, we define

$$\tau(\varepsilon) := \min\left\{ t \geq 0 : \rho\left(1 - \mathrm{e}^{-\mu t}\right) - \varepsilon \geq \sum_{(i,j)\triangleright\sigma_*} \alpha(i) \right\} = \frac{1}{\mu} \log\left(\frac{\rho}{\rho - \sum_{(i,j)\triangleright\sigma_*} \alpha(i) - \varepsilon} \right).$$

Fix $\tau_0 > \tau_{\mathrm{bd}}(s_0)$ and ε such that $\tau = \tau_0 + \tau(\varepsilon) + \varepsilon$. This is possible because $\tau_{\mathrm{bd}}(s_0) + \tau(\varepsilon) + \varepsilon$ decreases to $\tau_{\mathrm{eq}}(s_0)$ as $\varepsilon \to 0$. In addition, consider the random times

$$\xi_n := \inf\left\{ t \geq \tau_0 : \sum_{(i,j)\triangleright\sigma_*} \boldsymbol{q}_n(t,i,j) = \sum_{(i,j)\triangleright\sigma_*} \alpha_n(i) \right\}.$$

The proofs of (16a) and (16b) will be completed if we demonstrate that $\xi_n \leq \tau$ for all large enough n. Indeed, if this is established, then (16a) and (16b) follow from Lemma 7 with $r := r_*$, $\zeta_{n,1} := \xi_n$ and $\zeta_{n,2} := T$. The hypotheses of the lemma hold because $\boldsymbol{r}_n(t) \leq r_*$ for all $t \in [\tau_0, T]$ and all large enough n by the choice of τ_0 and Proposition 6.

To prove that $\xi_n \leq \tau$ for all large enough n, we show that

$$\limsup_{n\to\infty} \xi_n \leq \tau_0 + \tau(\varepsilon) < \tau_0 + \tau(\varepsilon) + \varepsilon = \tau. \tag{38}$$

If $r_* = 1$, then $\xi_n = \tau_0$ for all n and the above inequality holds, so suppose that $r_* > 1$.

Assume that (38) does not hold. Then there exists an increasing sequence $\mathcal{K}$ of natural numbers such that $\xi_k > \tau_0 + \tau(\varepsilon)$ for all $k \in \mathcal{K}$. Moreover, by Propositions 3 and 6, this sequence may be chosen so that the next two properties hold.

(i) The sequence $\{\boldsymbol{q}_k : k \in \mathcal{K}\}$ converges in $D_{\ell_1}[0,\infty)$.

(ii) We have $\boldsymbol{r}_k(t) \leq r_*$ for all $t \in [\tau_0, T]$ and $k \in \mathcal{K}$.

The definition of ξ_k implies that

$$\sum_{(i,j)\triangleright\sigma_*} \boldsymbol{q}_k(t,i,j) < \sum_{(i,j)\triangleright\sigma_*} \alpha_k(i) \quad \text{for all} \quad t \in [\tau_0, \tau_0 + \tau(\varepsilon)] \subset [\tau_0, \xi_k) \quad \text{and} \quad k \in \mathcal{K}.$$

The hypotheses of Proposition 5 hold with $r := r_*$, $t_0 := \tau_0$ and $t_1 := \tau_0 + \tau(\varepsilon)$, by the previous remark and properties (i) and (ii). Let $\boldsymbol{v}(r_*)$ be the function defined in this proposition, as the uniform limit of the tail processes $\boldsymbol{v}_k(r_*)$ over $[0,T]$. It follows from Propositions 4 and 5 that

$$\sup_{t\in[0,T]} |\boldsymbol{s}_k(t) - \boldsymbol{v}_k(t, r_*) - [\boldsymbol{s}(t) - \boldsymbol{v}(t, r_*)]| \leq \frac{\varepsilon}{2}$$

for all sufficiently large $k \in \mathcal{K}$. For each of these k, we have

$$\begin{aligned}
\sum_{(i,j)\triangleright\sigma_*} \boldsymbol{q}_k\left(\tau_0 + \tau(\varepsilon), i, j\right) &= \boldsymbol{s}_k(\tau_0 + \tau(\varepsilon)) - \boldsymbol{v}_k(\tau_0 + \tau(\varepsilon), r_*) \\
&\geq \boldsymbol{s}(\tau_0 + \tau(\varepsilon)) - \boldsymbol{v}(\tau_0 + \tau(\varepsilon), r_*) - \frac{\varepsilon}{2} \\
&> \rho + \left[\boldsymbol{s}(\tau_0) - \rho\right]\mathrm{e}^{-\mu\tau(\varepsilon)} - \boldsymbol{s}(\tau_0)\mathrm{e}^{-\mu\tau(\varepsilon)} - \frac{\varepsilon}{2} \\
&= \rho\left(1 - \mathrm{e}^{-\mu\tau(\varepsilon)}\right) - \frac{\varepsilon}{2} = \sum_{(i,j)\triangleright(i_r,j_r)} \alpha(i) + \frac{\varepsilon}{2}.
\end{aligned}$$

The third inequality follows from Proposition 5 and the last equality from the definition of $\tau(\varepsilon)$. It follows from (9) that the right-hand side is strictly larger than $\sum_{(i,j)\triangleright\sigma_*} \alpha_k(i)$ for all large enough $k \in \mathcal{K}$, which is a contradiction.

We conclude that (38) holds, which proves (16a) and (16b). We had already proved (16c) in Corollary 1; thus, the proof of the theorem is complete. □

8. Asymptotic Optimality

In this section, we prove Theorem 2. Specifically, in Section 8.1, we use drift analysis to demonstrate that the continuous-time Markov chains introduced in Section 3.3 are irreducible and positive-recurrent and to derive upper bounds for certain expectations and tail probabilities. In Section 8.2, we use these upper bounds to establish that the stationary distributions of the latter Markov chains are tight, and then we complete the proof of

Theorem 2 using the results of Sections 6 and 7. Finally, in Section 8.3, we demonstrate that JLMU is not optimal in general, although it is asymptotically optimal.

8.1. Drift Analysis

Denote the state space and the generator matrix of the continuous-time Markov chains defined in Section 3.3 by S_n and A_n, respectively. We use exactly the same notation for JLMU and SLTA, but we always indicate which policy is being considered. The drift of a function $f : S_n \to [0,\infty)$ is the function $A_n f$ defined by

$$A_n f(x) := \sum_{y\in S_n} A_n(x,y) f(y) = \sum_{y\neq x} A_n(x,y)\big[f(y) - f(x)\big] > -\infty \quad \text{for all} \quad x \in S_n.$$

The proof of the following proposition uses a Foster-Lyapunov argument, which is based on the drift of certain suitably chosen functions.

Proposition 7. *For each given n, the two continuous-time Markov chains introduced in Section 3.3 are irreducible and positive-recurrent. In particular, each of these Markov chains has a unique stationary distribution π_n.*

Proof. Suppose first that the load balancing policy is JLMU. Any occupancy state can reach the empty occupancy state after a finite number of consecutive departures. By the definition of S_n provided in Section 3.3, the latter remark implies that $\boldsymbol{q}_n$ is irreducible. Moreover, $\boldsymbol{q}_n$ is the empty occupancy state if and only if $\boldsymbol{s}_n = 0$, which implies that the empty occupancy state is positive-recurrent, because the $M/M/\infty$ queue $\boldsymbol{s}_n$ is irreducible and positive-recurrent. Thus, $\boldsymbol{q}_n$ is positive-recurrent.

Suppose now that the load balancing policy is SLTA. Any state $(q,r) \in S_n$ can reach the empty occupancy sate with $\boldsymbol{r}_n = r$ after a finite number of consecutive departures. Moreover, the latter state can reach the empty occupancy state with $\boldsymbol{r}_n = 1$ after a finite number of alternate arrivals and departures. We conclude from the definition of S_n provided in Section 3.3 that $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ is irreducible.

Next we use a Foster-Lyapunov argument to prove the positive recurrence. Consider the functions $f, g : S_n \to [0,\infty)$ defined by

$$f(q,r) := \sum_{(i,j)\in\mathcal{I}_+} q(i,j) \quad \text{and} \quad g(q,r) := r \quad \text{for all} \quad (q,r) \in S_n. \tag{39}$$

All server pools together form an infinite-server system; thus, $A_n f(q,r) = \lambda - \mu f(q,r)$ for all $(q,r) \in S_n$. In addition, we have

$$A_n g(q,r) = \lambda\big[\mathbb{1}_I(q,r) - \mathbb{1}_D(q,r)\big] \quad \text{for all} \quad (q,r) \in S_n.$$

Here I corresponds to those states (q, r) such that $\boldsymbol{r}_n$ increases if $(\boldsymbol{q}_n, \boldsymbol{r}_n) = (q,r)$ and the next event is an arrival. Specifically,

$$I := \big\{(q,r) \in S_n : q(i,j) = \alpha_n(i) \text{ for all } (i,j) \triangleright (i_r, j_r),\ nq(i_r, j_r) = n\alpha_n(i_r) - 1\big\}.$$

Also, D corresponds to those states (q, r) such that $\boldsymbol{r}_n$ decreases if $(\boldsymbol{q}_n, \boldsymbol{r}_n) = (q,r)$ and the next event is an arrival. Specifically,

$$D := \left\{(q,r) \in S_n : r > 1,\ n - \sum_{i=1}^{m} nq(i, \ell_i(r)) \geq n\beta_n,\ q(i_{r-1}, j_{r-1}) < \alpha_n(i_{r-1})\right\}.$$

Consider the function $h := f + 2g$ and let F be the set of those $(q,r) \in S_n$ that satisfy the following two conditions.

(i) We have $f(q,r) \leq 4\rho$.

(ii) We have $r = 1$ or $r > 1$ and $\big[\alpha_n(i_{r-1}) - \beta_n\big] j_{r-1} \leq 4\rho$.

The first condition holds for finitely many $q \in Q_n$ and the second condition holds for finitely many $r \geq 1$, thus F is finite. Next we establish that $A_n h \leq -\lambda + 4\lambda\mathbb{1}_F$. Note that $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ is nonexplosive because the infinite-server queue $\boldsymbol{s}_n$ has this property. Therefore, it follows from Hajek (2006, proposition 2.20.1) that $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ is positive-recurrent.

The latter inequality holds for all $(q,r) \in F$ since $A_n h \leq 3\lambda$. Hence, let us assume that $(q,r) \notin F$. Suppose that $(q,r) \notin F$ violates (i). Then

$$A_n h(q,r) \leq 3\lambda - \mu f(q,r) < 3\lambda - 4\lambda = -\lambda.$$

Assume now that $(q,r) \notin F$ satisfies (i). Then $(q,\ r)$ satisfies (i) and violates (ii), which implies that $r > 1$ and $\left[\alpha_n(i_{r-1}) - \beta_n\right] j_{r-1} > 4\rho \geq f(q,r)$. From this we conclude that $q(i_{r-1}, j_{r-1}) < \alpha_n(i_{r-1}) - \beta_n$, because otherwise $f(q,r) \geq j_{r-1} q(i_{r-1}, j_{r-1}) > 4\rho$. Thus,

$$n - \sum_{i=1}^{m} nq(i, \ell_i(r)) = \sum_{i=1}^{m} n\left[\alpha_n(i) - q(i, \ell_i(r))\right] \geq n\left[\alpha_n(i_{r-1}) - q(i_{r-1}, j_{r-1})\right] > n\beta_n.$$

It follows that $(q,r) \in D$; thus, $A_n h(q,r) = \lambda - \mu f(q,r) - 2\lambda \leq -\lambda$. □

Next we provide upper bounds for certain expectations and tail probabilities, which are used in the following section to demonstrate that the sequence of stationary distributions $\{\pi_n : n \geq 1\}$ is tight, both for JLMU and SLTA. First we state a technical lemma; the proof follows from Fubini's theorem and is provided in Online Appendix A.

Lemma 8. *Let x_n have the stationary distribution π_n, where $x_n = q_n$ if JLMU is used and $x_n = (q_n, r_n)$ if SLTA is used. If $f : S_n \to [0,\infty)$ satisfies*

$$E\left[\sum_{y \in S_n} \left|A_n(x_n, y)\right| f(y)\right] < \infty, \quad \textit{then} \quad E\left[A_n f(x_n)\right] = 0.$$

Consider the quantities

$$\theta_n^k := \sum_{(i,j) \trianglerighteq (i_k, j_k)} \alpha_n(i) \quad \textit{for all} \quad k \geq 1. \tag{40}$$

This lemma is used to prove the following two propositions.

Proposition 8. *Suppose that the load balancing policy is JLMU, fix n and consider the functions $f_k : S_n \to [0,\infty)$ defined by*

$$f_k(q) := \sum_{(i,j) \triangleleft (i_k, j_k)} q(i,j) \quad \textit{for all} \quad q \in S_n \quad \textit{and} \quad k \geq 1.$$

If q_n has the stationary distribution π_n, then

$$E\left[f_k(q_n)\right] \leq \rho e^{-n\left(\theta_n^k - 2\rho\right)} \quad \textit{for all} \quad k \geq 1.$$

Proof. Fix some $k \geq 1$ and define $J_i := \min\left\{j \geq 1 : (i,j) \triangleleft (i_k, j_k)\right\}$. The drift of f_k with respect to $\boldsymbol{q}_n$ satisfies

$$\begin{aligned}
A_n f_k(q) &= \lambda \mathbb{1}_{\{\sigma(q) \triangleleft (i_k, j_k)\}} - \sum_{(i,j) \triangleleft (i_k, j_k)} \mu j \left[q(i,j) - q(i,j+1)\right] \\
&= \lambda \mathbb{1}_{\{\sigma(q) \triangleleft (i_k, j_k)\}} - \sum_{i=1}^{m} \sum_{j=J_i}^{\infty} \mu j \left[q(i,j) - q(i,j+1)\right] \\
&= \lambda \mathbb{1}_{\{\sigma(q) \triangleleft (i_k, j_k)\}} - \sum_{i=1}^{m} \left[\mu(J_i - 1) q(i, J_i) + \mu \sum_{j=J_i}^{\infty} q(i,j)\right] \\
&= \lambda \mathbb{1}_{\{\sigma(q) \triangleleft (i_k, j_k)\}} - \mu f_k(q) - \mu \sum_{i=1}^{m} (J_i - 1) q(i, J_i) \\
&\leq \lambda \mathbb{1}_{\{f(q) \geq \theta_n^k\}} - \mu f_k(q) \quad \text{for all} \quad q \in S_n,
\end{aligned} \tag{41}$$

where $f(q)$ is as in (39). For the last step, $(i,j) \triangleright \sigma(q)$ implies that $q(i,j) = \alpha_n(i)$.

Observe that $|A_n(x,x)| = n\lambda + n\mu f(x)$ because $nf(x)$ is the total number of tasks at the occupancy state x and $n\lambda$ is the arrival rate of tasks. Hence,

$$\begin{aligned}
\sum_{y \in S_n} \left|A_n(x,y)\right| f_k(y) &= \left|A_n(x,x)\right| f_k(x) + \sum_{y \neq x} A_n(x,y) f_k(y) \\
&= 2\left|A_n(x,x)\right| f_k(x) + \sum_{y \in S_n} A_n(x,y) f_k(y) \\
&= 2\left[n\lambda + n\mu f(x)\right] f_k(x) + A_n f_k(x) \leq 2n\left[\lambda + \mu f(x)\right] f(x) + \lambda.
\end{aligned}$$

The right-hand side has a finite mean with respect to π_n because $nf(q_n)$ is the total number of tasks in the system in stationarity, which is Poisson distributed with mean $n\rho$. Thus, we conclude that $E\big[A_n f_k(q_n)\big] = 0$ by Lemma 8.

Taking expectations with respect to π_n on both sides of (41) and recalling that $nf(q_n)$ is Poisson distributed with mean $n\rho$, we obtain

$$E\big[f_k(q_n)\big] \le \rho E\Big[\mathbb{1}_{\{f(q_n)\ge\theta_n^k\}}\Big] = \rho P\Big(nf(q_n) \ge n\theta_n^k\Big) \le \rho \mathrm{e}^{-n\left(\theta_n^k - 2\rho\right)},$$

where the last inequality follows from a Chernoff bound. □

Proposition 9. *Suppose that the load balancing policy is SLTA, fix n and consider the functions $f_k : S_n \to [0,\infty)$ defined by*

$$f_k(q,r) := \sum_{(i,j)\lhd(i_k,j_k)} q(i,j) \quad \text{for all} \quad (q,r) \in S_n \quad \text{and} \quad k \ge 1.$$

Let (q_n, r_n) have the stationary distribution π_n. For each $k \ge 1$, we have

$$E\big[f_k(q_n,r_n)\big] \le \rho P(r_n > k), \tag{42a}$$

$$P(r_n > k) \le \mathrm{e}^{-n\left(\theta_n^k - 2\rho\right)} + \mathrm{e}^{-n\left[\theta_n^k - L(k)\beta_n - 2\rho\right]} + \mathrm{e}^{-n\left[\ell(k)\alpha_n^{\min} - 2\rho\right]}, \tag{42b}$$

where θ_n^k is defined as in (40), $\alpha_n^{\min} := \min\{\alpha_n(i) : 1 \le i \le m\}$,

$$L(k) := \max\{\ell_i(k+1) : 1 \le i \le m\} \quad \text{and} \quad \ell(k) := \min\{\ell_i(k+1) : 1 \le i \le m\}.$$

Proof. Fix $k \ge 1$. As in the proof of Proposition 8, we see that

$$A_n f_k(q,r) \le \lambda \mathbb{1}_{\{r>k\}} - \mu f_k(q,r) \quad \text{for all} \quad (q,r) \in S_n,$$

and that f_k satisfies the hypothesis of Lemma 8. Then we obtain (42a) by taking the expectation with respect to π_n on both sides of the latter inequality.

Consider the sets I and D defined in the proof of Proposition 7 and let

$$I_k := I \cap \big\{(q,r) \in S_n : r \ge k\big\} \quad \text{and} \quad D_k := D \cap \big\{(q,r) \in S_n : r > k\big\}.$$

The first step of the proof of (42b) is to establish that

$$P\big((q_n,r_n) \in I_k\big) = P\big((q_n,r_n) \in D_k\big). \tag{43}$$

Fix $l > k$ and consider the function $g_k^l : S_n \to [0,\infty)$ defined by

$$g_k^l(q,r) = (r-k)^+ \, \mathbb{1}_{\{r<l\}} + (l-k)\mathbb{1}_{\{r\ge l\}} \quad \text{for all} \quad (q,r) \in S_n.$$

As in the proof of Proposition 7, we obtain

$$A_n g_k^l(q,r) = \lambda\Big[\mathbb{1}_{I_k^l}(q,r) - \mathbb{1}_{D_k^l}(q,r)\Big], \tag{44}$$

where the sets I_k^l and D_k^l are defined by

$$I_k^l := I \cap \big\{(q,r) \in S_n : k \le r < l\big\} \quad \text{and} \quad D_k^l := D \cap \big\{(q,r) \in S_n : k < r \le l\big\}.$$

Define $f(q, r)$ as in (39) and note that $|A_n((x,r),(x,r))| = n\lambda + n\mu f(x,r)$. Thus,

$$\begin{aligned}\sum_{(y,s)\in S_n} \big|A_n\big((x,r),(y,s)\big)\big| g_k^l(y,s) &= 2\big|A_n\big((x,r),(y,s)\big)\big| g_k^l(x,r) + A_n g_k^l(x,r) \\ &\le 2n\big[\lambda + \mu f(x,r)\big](l-k) + \lambda.\end{aligned}$$

The right-hand side has a finite mean with respect to π_n because $nf(q_n,r_n)$ is the total number of tasks in stationarity, which is Poisson distributed with mean $n\rho$. Therefore, it follows from Lemma 8 and (44) that $P\big((q_n,r_n) \in I_k^l\big) = P\big((q_n,r_n) \in D_k^l\big)$. The sets I_k^l and D_k^l increase to I_k and D_k, respectively, as $l \to \infty$. This implies (43) because

$$P\big((q_n,r_n) \in I_k\big) = \lim_{l\to\infty} P\big((q_n,r_n) \in I_k^l\big) = \lim_{l\to\infty} P\big((q_n,r_n) \in D_k^l\big) = P\big((q_n,r_n) \in D_k\big).$$

Now we may write

$$\begin{aligned} P(r_n > k) &= P\big(r_n > k, (q_n, r_n) \in D\big) + P\big(r_n > k, (q_n, r_n) \notin D\big) \\ &= P\big((q_n, r_n) \in D_k\big) + P\big(r_n > k, (q_n, r_n) \notin D\big) \\ &= P\big((q_n, r_n) \in I_k\big) + P\big(r_n > k, (q_n, r_n) \notin D\big). \end{aligned} \tag{45}$$

Using the definition of I_k, we can bound the first term on the last line by

$$P\left(f(q_n, r_n) \geq \sum_{(i,j) \triangleright (i_{r_n}, j_{r_n})} \alpha_n(i), r_n > k\right) \leq P\Big(f(q_n, r_n) \geq \theta_n^k\Big). \tag{46}$$

The second term on the last line of (45) can be bounded by

$$\begin{aligned} &P\left(\sum_{i=1}^{m} q_n(i, \ell_i(r_n)) > 1 - \beta_n, r_n > k\right) + P\big(q_n(i_{r_n-1}, j_{r_n-1}) = \alpha_n(i_{r_n-1}), r_n > k\big) \leq \\ &P\Big(f(q_n, r_n) \geq \theta_n^k - L(k)\beta_n\Big) + P\big(f(q_n, r_n) \geq \ell(k)\alpha_n^{\min}\big). \end{aligned} \tag{47}$$

For the last inequality, observe that the condition inside the first probability sign of the left-hand side of (47) implies that

$$\sum_{i=1}^{m} q_n(i, \ell_i(k+1)) \geq \sum_{i=1}^{m} q_n(i, \ell_i(r_n)) > 1 - \beta_n,$$

and this in turn implies that

$$f(q_n, r_n) \geq \sum_{(i,j) \trianglerighteq (i_k, j_k)} \alpha_n(i) - L(k)\beta_n = \theta_n^k - L(k)\beta_n,$$

because $q_n(i,j)$ is nonincreasing in j for all i. In addition, the condition inside the second probability sign of the left-hand side of (47) implies that

$$f(q_n, r_n) \geq j_{r_n-1} q_n(i_{r_n-1}, j_{r_n-1}) = j_{r_n-1}\alpha_n(i_{r_n-1}) \geq \ell(k)\alpha_n^{\min}.$$

We obtain (42b) from (46) and (47), recalling that $nf(q_n, r_n)$ is Poisson distributed with mean $n\rho$ and applying Chernoff bounds. □

8.2. Proof of the Asymptotic Optimality

In this section, we prove Theorem 2. As a first step, we establish that the sequence of stationary distributions $\{\pi_n : n \geq 1\}$ is tight both for JLMU and SLTA.

Proposition 10. *If the load balancing policy is JLMU, then $\{\pi_n : n \geq 1\}$ is tight in ℓ_1. If the load balancing policy is SLTA, then $\{\pi_n : n \geq 1\}$ is tight in $\ell_1 \times \mathbb{N}$.*

Proof. Suppose first that the load balancing policy is JLMU and let q_n have the stationary distribution π_n for each n. The sequence $\{q_n : n \geq 1\}$ is tight with respect to the product topology because the random variables q_n take values in $[0,1]^{\mathcal{I}}$, which is compact with respect to the product topology. Therefore, as in Mukherjee et al. (2018, lemma 2), the tightness in ℓ_1 of $\{q_n : n \geq 1\}$ will follow if we establish that

$$\lim_{k\to\infty} \limsup_{n\to\infty} P\left(\sum_{(i,j) \triangleleft (i_k, j_k)} q_n(i,j) > \varepsilon\right) = 0 \quad \text{for all} \quad \varepsilon > 0. \tag{48}$$

By Proposition 8 and Markov's inequality, we have

$$P\left(\sum_{(i,j) \triangleleft (i_k, j_k)} q_n(i,j) > \varepsilon\right) \leq \frac{\rho}{\varepsilon} \mathrm{e}^{-n(\theta_n^k - 2\rho)} \quad \text{for all} \quad k, n \geq 1 \quad \text{and} \quad \varepsilon > 0.$$

For all sufficiently large k, the exponent on the right-hand side converges to minus infinity as n grows large and k is held fixed. Thus, $\{q_n : n \geq 1\}$ is tight in ℓ_1.

Suppose now that the load balancing policy is SLTA and let (q_n, r_n) have the stationary distribution π_n for each n. To prove that $\{(q_n, r_n) : n \geq 1\}$ is tight in $\ell_1 \times \mathbb{N}$, it suffices to show that $\{q_n : n \geq 1\}$ and $\{r_n : n \geq 1\}$ are tight in

ℓ_1 and $\mathbb{N}$, respectively. Indeed, if the latter properties hold, then for each $\varepsilon > 0$ there exist compact sets $K_q \subset \ell_1$ and $K_r \subset \mathbb{N}$ such that

$$\max\{P(q_n \notin K_q), P(r_n \notin K_r)\} \le \frac{\varepsilon}{2} \quad \text{for all} \quad n.$$

Therefore, the compact set $K_q \times K_r \subset \ell_1 \times \mathbb{N}$ satisfies

$$P((q_n, r_n) \notin K_q \times K_r) \le P(q_n \notin K_q) + P(r_n \notin K_r) \le \varepsilon \quad \text{for all} \quad n.$$

By Proposition 9 and Markov's inequality, we have

$$P\left(\sum_{(i,j)\lhd(i_k,j_k)} q_n(i,j) > \varepsilon\right) \le \frac{\rho}{\varepsilon} P(r_n > k) \quad \text{for all} \quad k, n \ge 1 \quad \text{and} \quad \varepsilon > 0,$$

$$P(r_n > k) \le \mathrm{e}^{-n(\theta_n^k - 2\rho)} + \mathrm{e}^{-n\left[\theta_n^k - L(k)\beta_n - 2\rho\right]} + \mathrm{e}^{-n\left[\ell(k)\alpha_n^{\min} - 2\rho\right]} \quad \text{for all} \quad k, n \ge 1.$$

For all large enough k, the right-hand side of the second inequality is summable over n and in particular vanishes with n. This implies that the sequences $\{q_n : n \ge 1\}$ and $\{r_n : n \ge 1\}$ are tight in ℓ_1 and $\mathbb{N}$, respectively. □

We also need the following technical lemma.

Lemma 9. *Fix $q \in \ell_1$ and suppose that the marginal utilities are bounded. Then*

$$u(q) = \sum_{i=1}^{m} u_i(0) q(i,0) + \sum_{(i,j)\in\mathcal{I}_+} \Delta(i, j-1) q(i,j).$$

Proof. Note that

$$\begin{aligned} u(q) &= \sum_{i=1}^{m} \sum_{j=0}^{\infty} u_i(j)\left[q(i,j) - q(i,j+1)\right] \\ &= \lim_{k\to\infty} \sum_{i=1}^{m} \sum_{j=0}^{k} u_i(j)\left[q(i,j) - q(i,j+1)\right] \\ &= \lim_{k\to\infty} \sum_{i=1}^{m} \left[u_i(0) q(i,0) + \sum_{j=1}^{k} \Delta(i, j-1) q(i,j) - u_i(k) q(i,k+1) \right], \\ &= \sum_{i=1}^{m} u_i(0) q(i,0) + \sum_{(i,j)\in\mathcal{I}_+} \Delta(i, j-1) q(i,j) - \lim_{k\to\infty} \sum_{i=1}^{m} u_i(k) q(i,k+1). \end{aligned}$$

The second term in the last line is absolutely convergent because the marginal utilities are bounded and $q \in \ell_1$. Moreover, there exists $M \ge 0$ such that

$$\lim_{k\to\infty} |u_i(k)| q(i,k+1) \le \lim_{k\to\infty} \left| u_i(0) + \sum_{j=0}^{k-1} \Delta(i,j) \right| q(i,k+1) \le \lim_{k\to\infty} (|u_i(0)| + Mk) q(i,k+1) \quad \text{for all} \quad i,$$

and the latter limit is zero for all i because $q \in \ell_1$. □

Now we are ready to prove Theorem 2.

Proof of Theorem 2. Suppose first that the load balancing policy is JLMU. It follows from Prokhorov's theorem and Proposition 10 that the stationary distributions $\{\pi_n : n \ge 1\}$ are relatively compact in ℓ_1; thus, every subsequence has a further subsequence that converges in distribution. To establish (a), it suffices to prove the following statement: if $\mathcal{K}$ is an increasing sequence of natural numbers such that $\{\pi_k : k \in \mathcal{K}\}$ converges weakly to π, then π is the Dirac measure concentrated at q_*. Similarly, the sequence $\{\pi_n : n \ge 1\}$ is relatively compact in $\ell_1 \times \mathbb{N}$ if the load balancing policy is SLTA, and to prove (b), it suffices to establish the following statement: if $\mathcal{K}$ is an

increasing sequence of natural number such that $\{\pi_k : k \in \mathcal{K}\}$ converges weakly to π, then π is the Dirac measure at (q_*, r_*). Here we prove (a) and (b) in parallel, proceeding as indicated previously.

Fix an arbitrary increasing sequence of natural numbers $\mathcal{K}$ such that $\{\pi_k : k \in \mathcal{K}\}$ converges weakly to a certain probability measure π. The following constructions use Skorokhod's representation theorem. If the load balancing policy is JLMU, then there exist random variables q_k and q, distributed as π_k and π, respectively, that are defined on a common probability space $(\Omega_I, \mathcal{F}_I, \mathbb{P}_I)$ and satisfy

$$\lim_{k\to\infty} \|q_k(\omega) - q(\omega)\|_1 = 0 \quad \text{for all} \quad \omega \in \Omega_I.$$

If the load balancing policy is SLTA, then there exist random variables (q_k, r_k) and (q, r), with distributions given by the probability measures π_k and π, respectively, that are defined on some common probability space $(\Omega_I, \mathcal{F}_I, \mathbb{P}_I)$ and satisfy

$$\lim_{k\to\infty} \|q_k(\omega) - q(\omega)\|_1 = 0 \text{ and } \lim_{k\to\infty} r_k(\omega) = r(\omega) \quad \text{for all} \quad \omega \in \Omega_I.$$

The second limit implies that there exists a random variable R such that $r_k(\omega) \le R(\omega)$ for all $k \in \mathcal{K}$ and $\omega \in \Omega_I$. Moreover, $q_k(i,j) < \alpha_k(i)$ for all $(i,j) \trianglelefteq (i_{r_k}, j_{r_k})$ on Ω_I by the definition of the state space S_k for SLTA. Indeed, recall from Remark 1 that the latter property is preserved by arrivals and departures and observe that it holds for the empty occupancy state with $\boldsymbol{r}_k = 1$.

If the load balancing policy is JLMU, then we may construct occupancy processes $\boldsymbol{q}_k$ on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as in Section 5.1.1, such that $\boldsymbol{q}_k(0) = q_k$ and (9) holds with $q_0 = q$. If the load balancing policy is SLTA, then we may construct processes $(\boldsymbol{q}_k, \boldsymbol{r}_k)$ on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as in Section 5.1.2, such that $\big(\boldsymbol{q}_k(0), \boldsymbol{r}_k(0)\big) = (q_k, r_k)$ and the assumptions of Section 4.2.1 hold with $q_0 = q$. For both load balancing policies, we may assume by Proposition 3 that $\boldsymbol{q}_k$ converges in $D_{\ell_1}[0,\infty)$ to a process $\boldsymbol{q}$ with a probability of one; this may require to replace $\mathcal{K}$ by a subsequence.

If the load balancing policy is JLMU, then Theorem 5 implies that $\boldsymbol{q}(\omega)$ is the unique fluid trajectory such that $\boldsymbol{q}(\omega, 0) = q(\omega)$ for each $\omega \in \Omega$. Moreover, by Theorem 4,

$$\lim_{t\to\infty} \big|\big|\boldsymbol{q}(\omega, t) - q_*\big|\big|_1 = 0 \quad \text{for all} \quad \omega \in \Omega. \tag{49}$$

If the load balancing policy is SLTA, then (16b) and (16c) hold on a set of probability of one by Theorem 6. Fix any $\omega \in \Omega$ such that $\boldsymbol{q}_k(\omega)$ converges to $\boldsymbol{q}(\omega)$ in $D_{\ell_1}[0,\infty)$ and such that (16b) and (16c) hold. Also, choose $\tau > \tau_{\mathrm{eq}}(s_0(\omega))$. It follows from (16b) that

$$\begin{aligned}|\alpha(i) - \boldsymbol{q}(\omega, T, i, j)| &\le \sup_{t\in[\tau,T]} |\alpha(i) - \boldsymbol{q}(\omega, t, i, j)| \\ &\le \lim_{k\to\infty} \sup_{t\in[\tau,T]} |\alpha(i) - \boldsymbol{q}_k(\omega, t, i, j)| + \lim_{k\to\infty} \sup_{t\in[\tau,T]} \|\boldsymbol{q}_k(\omega, t) - \boldsymbol{q}(\omega, t)\|_1 = 0,\end{aligned}$$

for all $(i,j) \triangleright \sigma_*$ and $T \ge \tau$. Thus, $\boldsymbol{q}(\omega, t, i, j) = q_*(i,j)$ for all $(i,j) \triangleright \sigma_*$ and $t \ge \tau$. Similarly, we conclude from (16c) that

$$\begin{aligned}\sum_{(i,j)\trianglelefteq\sigma_*} \boldsymbol{q}(\omega, T, i, j)\mathrm{e}^{\mu(T-\tau)} &\le \sup_{t\in[\tau,T]} \sum_{(i,j)\trianglelefteq\sigma_*} \boldsymbol{q}(\omega, t, i, j)\mathrm{e}^{\mu(t-\tau)} \\ &\le \limsup_{k\to\infty} \sup_{t\in[\tau,T]} \sum_{(i,j)\trianglelefteq\sigma_*} \boldsymbol{q}_k(\omega, t, i, j)\mathrm{e}^{\mu(t-\tau)} \\ &\quad + \lim_{k\to\infty} \sup_{t\in[\tau,T]} \|\boldsymbol{q}_k(\omega, t) - \boldsymbol{q}(\omega, t)\|_1 \mathrm{e}^{\mu(t-\tau)} \le c(\omega, \tau)\end{aligned}$$

for all $T \ge \tau$; for the last inequality, in the last line, $\mathrm{e}^{\mu(t-\tau)} \le \mathrm{e}^{\mu(T-\tau)}$ for all $t \in [\tau, T]$, so the last limit equals zero. Because T can be arbitrarily large, we have

$$\lim_{t\to\infty} \sum_{(i,j)\trianglelefteq\sigma_*} \boldsymbol{q}(\omega, t, i, j) \le \lim_{t\to\infty} c(\omega, \tau)\mathrm{e}^{-\mu(t-\tau)} = 0,$$

and in particular $\boldsymbol{q}(\omega, t, i, j) \to q_*(i,j)$ as $t \to \infty$ for all $(i,j) \triangleleft \sigma_*$. This also holds for $(i,j) = \sigma_*$ by Proposition 4, so we conclude that

$$\mathbb{P}\left(\lim_{t\to\infty} \|\boldsymbol{q}(t) - q_*\|_1 = 0\right) = 1. \tag{50}$$

It follows from the stationarity of π_k that $\boldsymbol{q}_k(t)$ is distributed as q_k for all $t \geq 0$ in the case of JLMU and that $(\boldsymbol{q}_k(t), \boldsymbol{r}_k(t))$ has the same distribution as (q_k, r_k) for all $t \geq 0$ if the load balancing policy is SLTA. Furthermore, recall that in either case, we have

$$\mathbb{P}\left(\lim_{k\to\infty} \|\boldsymbol{q}_k(t) - \boldsymbol{q}(t)\|_1 = 0\right) = 1 \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad q_k \Rightarrow q \quad \text{as} \quad k \to \infty,$$

which implies that $\boldsymbol{q}(t)$ is distributed as q for all $t \geq 0$. Moreover, $\boldsymbol{q}(t)$ converges weakly to q_* as $t \to \infty$ by (49) and (50). Therefore, q corresponds to the Dirac probability measure concentrated at q_* both for JLMU and SLTA.

This completes the proof of (a). To finish the proof of (b), observe that $s_0 = \rho$ with a probability of one because $q_0 = q$ is equal to q_* with a probability of one. It follows from (16a) that

$$\mathbb{P}\left(\lim_{k\to\infty} \boldsymbol{r}_k(t) = r_*\right) = 1 \quad \text{for all} \quad t > \tau_{\text{eq}}(\rho).$$

Hence, $\boldsymbol{r}_k(t)$ converges weakly to r_* for all $t > \tau_{\text{eq}}(\rho)$. Because $\boldsymbol{r}_k(t)$ has the same distribution as r_k for all $t \geq 0$, we conclude that r_k converges weakly to r_*. Note that q_* and r_* are deterministic; thus, q_k and r_k converge to q_* and r_*, respectively, in probability, which implies that (q_k, r_k) converges to (q_*, r_*) in probability. This completes the proof of (b).

Next we prove the statements about the stationary overall utilities, and here exactly the same arguments apply both for JLMU and SLTA. Suppose that q_n has the stationary distribution. By (a) and (b), q_n converges in probability to q_* in ℓ_1, and by Lemma 9,

$$u(x) = \sum_{i=1}^{m} u_i(0)x(i,0) + \sum_{(i,j)\in\mathcal{I}_+} \Delta(i,j-1)x(i,j) \quad \text{for all} \quad x \in \ell_1.$$

Because the marginal utilities are bounded, there exists a constant a such that

$$|u(x) - u(y)| \leq a\|x - y\|_1 \quad \text{for all} \quad x, y \in \ell_1.$$

This implies that $u(q_n)$ converges in probability to $u(q_*)$.

Finally, observe that $|u(q_n)| \leq b + cs_n$ with

$$b := \sum_{i=1}^{m} |u_i(0)|, \quad c := \max_{(i,j)\in\mathcal{I}} |\Delta(i,j)| \quad \text{and} \quad s_n := \sum_{(i,j)\in\mathcal{I}_+} q_n(i,j).$$

Because ns_n is Poisson distributed with mean $n\rho$, we have

$$E\left[u(q_n)^2\right] \leq b^2 + 2bcE[s_n] + c^2E\left[s_n^2\right] \leq b^2 + 2bc\rho + c^2\rho^2 + \frac{c^2\rho}{n} \leq b^2 + 2bc\rho + c^2\rho^2 + c^2\rho.$$

The last expression does not depend on n; therefore, we conclude that $\{u(q_n) : n \geq 1\}$ is uniformly integrable. As a result, we have

$$\lim_{n\to\infty} E\left[u(q_n)\right] = u(q_*).$$

This completes the proof. □

8.3. Suboptimality Result

In this section, we prove that JLMU is generally not optimal in the prelimit, although it becomes optimal as the number of server pools grows large. For this purpose, we construct an example in which JLMU is strictly outperformed by another policy.

Specifically, consider a system with two heterogeneous server pools and assume that the associated utility functions are of the following form:

$$u_1(x) := a\varepsilon x \quad \text{and} \quad u_2(y) := ay\mathbb{1}_{\{y<1\}} + a\mathbb{1}_{\{y\geq 1\}} \quad \text{for all} \quad x, y \geq 0,$$

where $a > 0$ and $\varepsilon \in (0,1)$ are constants. Note that

$$\Delta(1,0) < \Delta(2,0) \quad \text{but} \quad \Delta(1,j) > \Delta(2,j) \quad \text{for all} \quad j \geq 1.$$

As a result, JLMU sends tasks to server pool 2 if and only if this server pool is empty; thus, the number of tasks in server pool 2 is zero or one in stationarity. This assignment rule guarantees the largest increase in the

aggregate utility of the system at each arrival epoch. However, this increase can be very small when tasks are sent to server pool 1, whereas the decrease in the aggregate utility can be comparatively large when a departure leaves server pool 2 with zero tasks. Particularly, this is the case when ε is small.

Suppose that tasks arrive as a Poisson process of intensity λ with exponential service times of mean $1/\mu$. In addition, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the number of tasks in server pools 1 and 2, respectively. Next we provide an upper bound for the mean stationary overall utility of JLMU when the utility functions are as defined previously.

As already noted, a new task is sent to server pool 2 if and only if $\boldsymbol{Y} = 0$. Therefore, the following statements hold.

- We have $\boldsymbol{Y}(t) \in \{0, 1\}$ for all sufficiently large t.
- Conditional on $\boldsymbol{Y}(0) \in \{0, 1\}$, the process $\boldsymbol{Y}$ alone is a birth-death process with state space $\{0, 1\}$, birth rate λ, and death rate μ.

By Proposition 7, the process $(\boldsymbol{X}, \boldsymbol{Y})$ has a unique stationary distribution (X, Y), and by the previous observations, we have

$$P(Y = 0) = \frac{\mu}{\lambda + \mu} \quad \text{and} \quad P(Y = 1) = \frac{\lambda}{\lambda + \mu}.$$

If we let $U := u_1(X) + u_2(Y)$ denote the aggregate utility in stationarity, then

$$E[U] = a\varepsilon E[X] + aP(Y = 1) \le a\varepsilon E[X + Y] + aP(Y = 1) = a\left[\varepsilon\rho + \frac{\rho}{\rho + 1}\right].$$

Consider now the policy that sends all tasks to server pool 2. Whereas JLMU dispatches the new tasks in a greedy fashion, this other policy is conservative because it tries to avoid drops of $u_2(\boldsymbol{Y})$ from a to zero by keeping a positive number of tasks in server pool 2.

The mean stationary aggregate utility can be computed explicitly for the policy that we described previously because $\boldsymbol{Y}$ is now an $M/M/\infty$ queue and $\boldsymbol{X}(t) = 0$ for all sufficiently large t. Let (X, Y) be the stationary distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ and let $V := u_1(X) + u_2(Y)$ denote the aggregate utility of the system in stationarity, we have

$$E[V] = aP(Y > 0) = a(1 - \mathrm{e}^{-\rho}).$$

It is not difficult to verify that ε can be chosen so that

$$\varepsilon\rho + \frac{\rho}{\rho + 1} < 1 - \mathrm{e}^{-\rho}$$

for all ρ in some open interval contained in $(0, \infty)$. If ε and ρ are chosen so that the latter inequality holds, then $E[U] < E[V]$, and thus JLMU is strictly suboptimal.

In general, the right balance between greedy and conservative actions is difficult to determine and depends intricately on the set of utility functions. However, the benefits of conservative actions, which prevent the number of tasks in server pools of specific classes from dropping below certain occupancy levels, diminish as the scale of the system grows. Indeed, the average fraction of these server pools that have less tasks than the optimal quantity decreases as the number of server pools grows, even if the assignment policy is purely greedy as JLMU; essentially, this is a consequence of the increase in the number of server pools per class and the decrease in the coefficient of variation of the total number of tasks in the system. Therefore, the associated loss in mean stationary overall utility asymptotically vanishes as the scale of the system grows.

References

Benameur N, Fredj SB, Oueslati-Boulahia S, Roberts JW (2002) Quality of service and flow level admission control in the Internet. *Comput. Networks* 40(1):57–71.

Bhamidi S, Budhiraja A, Dewaskar M (2022) Near equilibrium fluctuations for supermarket models with growing choices. *Ann. Appl. Probab.* (Institute of Mathematical Statistics), 32(3):2083–2138.

Bramson M (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30(1): 89–140.

Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Oper. Res.* 43(3):867–886.

Fontes LR (1989) A note on Kolmogorov backward equations. *Brazilian J. Probability Statist.* (Institute of Mathematical Statistics), 3(1):59–65.

Gamarnik D, Tsitsiklis JN, Zubeldia M (2018) Delay, memory, and messaging tradeoffs in distributed service systems. *Stochastic Systems* 8(1):45–74.

Gamarnik D, Tsitsiklis JN, Zubeldia M (2020) A lower bound on the queueing delay in resource constrained load balancing. *Ann. Appl. Probability* 30(2):870–901.

Gardner K, Stephens C (2019) Smart dispatching in heterogeneous systems. *Performance Evaluation Rev.* 47(2):12–14.

Gardner K, Jaleel JA, Wickeham A, Doroudi S (2021) Scalable load balancing in the presence of heterogeneous servers. *Performance Evaluation Rev.* 48(3):37–38.

Goldsztajn D, Borst SC, Van Leeuwaarden JS (2021a) Learning and balancing unknown loads in large-scale systems. Preprint, submitted December 16, https://arxiv.org/abs/2012.10142.

Goldsztajn D, Ferragut A, Paganini F (2021b) Automatic cloud instance provisioning with quality and efficiency. *Performance Evaluation* 149–150:102209.

Goldsztajn D, Ferragut A, Paganini F, Jonckheere M (2018) Controlling the number of active instances in a cloud environment. *Performance Evaluation Rev.* 45(3):15–20.

Goldsztajn D, Borst SC, Van Leeuwaarden JS, Mukherjee D, Whiting PA (2022) Self-learning threshold-based load balancing. *INFORMS J. Comput.* 34(1):39–54.

Hajek B (2006) *Notes for ECE 467: Communication Network Analysis* (University of Illinois at Urbana-Champaign, Urbana).

Horváth IA, Scully Z, Van Houdt B (2019) Mean field analysis of join-below-threshold load balancing for resource sharing servers. *Proc. ACM on Measurement and Anal. of Comput. Systems* (ACM, New York), 3(3):1–21.

Jaleel JA, Wickeham A, Doroudi S, Gardner K (2022) A general "power-of-*d*" dispatching framework for heterogeneous systems. *Queueing Systems* 102(3):431–480.

Karthik A, Mukhopadhyay A, Mazumdar RR (2017) Choosing among heterogeneous server clouds. *Queueing Systems* 85(1):1–29.

Key P, Massoulié L, Bain A, Kelly F (2004) Fair Internet traffic integration: Network flow models and analysis. *Ann. Telecomm.* 59(11): 1338–1352.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

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

Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distribution Systems* 12(10):1094–1104.

Mukherjee D, Borst SC, Van Leeuwaarden JS, Whiting PA (2018) Universality of power-of-*d* load balancing in many-server systems. *Stochastic Systems* 8(4):265–292.

Mukherjee D, Borst SC, Van Leeuwaarden JS, Whiting PA (2020) Asymptotic optimality of power-of-*d* load balancing in large-scale systems. *Math. Oper. Res.* 45(4):1535–1571.

Mukherjee D, Dhara S, Borst SC, Van Leeuwaarden JS (2017) Optimal service elasticity in large-scale distributed systems. *Proc. ACM on Measurement and Anal. of Comput. Systems* (ACM, New York, Philadelphia), 1(1):1–28.

Mukhopadhyay A, Mazumdar RR, Guillemin F (2015a) The power of randomized routing in heterogeneous loss systems. *Proc. 27th Internat. Teletraffic Congress* (IEEE, New York), 125–133.

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

Rudin W (1976) *Principles of Mathematical Analysis*, vol. 3 (McGraw-Hill, New York).

Sparaggis PD, Towsley D, Cassandras C (1993) Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. *J. Appl. Probability* 30(1):223–236.

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80(4):341–361.

Van der Boor M, Borst SC, Van Leeuwaarden JS, Mukherjee D (2022) Scalable load balancing in networked systems: A survey of recent advances. *SIAM Rev.* (SIAM, Philadelphia), 64(3):554–622.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.

Winston W (1977) Optimality of the shortest line discipline. *J. Appl. Probability* 14(1):181–189.

Xie Q, Dong X, Lu Y, Srikant R (2015) Power of d choices for large-scale bin packing: A loss model. *Performance Evaluation Rev.* 43(1):321–334.

Zhou X, Tan J, Shroff N (2018) Heavy-traffic delay optimality in pull-based load balancing systems: Necessary and sufficient conditions. *Proc. ACM on Measurement and Anal. of Comput. Systems* (ACM, New York), 2(3):1–33.

Zhou X, Wu F, Tan J, Sun Y, Shroff N (2017) Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms. *Proc. ACM on Measurement and Anal. of Comput. Systems* (ACM, New York), 1(2):1–30.