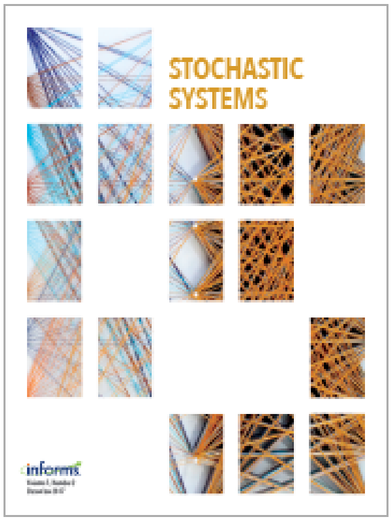


Stochastic Systems

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

Transform Methods for Heavy-Traffic Analysis

Daniela Hurtado-Lange, Siva Theja Maguluri

To cite this article:
Daniela Hurtado-Lange, Siva Theja Maguluri (2020) Transform Methods for Heavy-Traffic Analysis. Stochastic Systems 10(4):275-309. https://doi.org/10.1287/stsy.2019.0056

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

STOCHASTIC SYSTEMS
Vol. 10, No. 4, December 2020, pp. 275–309
ISSN 1946-5238 (online)

Transform Methods for Heavy-Traffic Analysis

Daniela Hurtado-Lange,[a] Siva Theja Maguluri[a]

[a] Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

Contact: d.hurtado@gatech.edu, https://orcid.org/0000-0002-1073-4716 (DH-L); siva.theja@gatech.edu, https://orcid.org/0000-0002-5797-1639 (STM)

Received: January 15, 2019
Revised: April 27, 2019; September 1, 2019; January 3, 2020
Accepted: February 5, 2020
Published Online in Articles in Advance: October 26, 2020

https://doi.org/10.1287/stsy.2019.0056

Abstract. The drift method was recently developed to study queuing systems in steady state. It was used successfully to obtain bounds on the moments of the scaled queue lengths that are asymptotically tight in heavy traffic and in a wide variety of systems, including generalized switches, input-queued switches, bandwidth-sharing networks, and so on. In this paper, we develop the use of transform techniques for heavy-traffic analysis, with a special focus on the use of moment-generating functions. This approach simplifies the proofs of the drift method and provides a new perspective on the drift method. We present a general framework and then use the moment-generating function method to obtain the stationary distribution of scaled queue lengths in heavy traffic in queuing systems that satisfy the complete resource pooling condition. In particular, we study load balancing systems and generalized switches under general settings.

Funding: This research was supported in part by the National Science Foundation [Grant National Science Foundation (NSF) - Computing and Communication Foundations (CCF) 1850439]. D. Hurtado-Lange has received partial funding from the Chilean National Agency for Research and Development (ANID)/DOCTORADO BECAS CHILE [Grant 2018–72190413].

Keywords: drift method • heavy-traffic analysis • state space collapse • complete resource pooling

1. Introduction

Exact analysis of queuing systems that arise in the study of stochastic processing networks (SPNs) is usually intractable, so it is common to analyze them in various asymptotic regimes to get insights into their behavior. A very popular regime in the literature is the heavy-traffic regime, where the system is loaded very close to its maximum capacity. This regime is sometimes called the *classical* or *conventional heavy-traffic regime*. One of the advantages of the heavy-traffic limit is that many queuing systems behave as if they live in a much lower-dimensional subspace of the state space in the limit. This phenomenon is known as *state space collapse* (SSC). If the heavy-traffic limit is taken such that exactly one resource constraint is made tight, then the system is said to satisfy the *complete resource pooling* (CRP) condition (Harrison and López 1999, Williams 2000, Dai and Lin 2008).

Over the past decades, several queuing systems that satisfy the CRP condition have been successfully and extensively studied using diffusion limits and Brownian motion processes. This approach was first developed by Kingman (1962a), where a $G/G/1$ queue was studied in heavy traffic. Later, it was successfully applied in a variety of systems that satisfy the CRP condition (Harrison 1988, 1998; Williams 1998, 2000; Harrison and López 1999; Stolyar 2004; Gamarnik and Zeevi 2006). In this approach, a scaled version of the queue lengths process is considered, and it is shown that it converges to a reflected Brownian motion (RBM) process. SSC is then established to show that this RBM process lives in a lower-dimensional subspace. Because the queue lengths cannot be negative, they *reflect* at the origin, so this lower-dimensional Brownian motion process is called an *RBM process*. Such a result is called *process-level convergence* and may be useful in approximating transient behavior. The next step is to obtain the stationary distribution of this RBM, which is usually the same as the heavy-traffic limiting stationary distribution of the original (unscaled) queuing system. However, this must be formally established by proving that the limit to steady state and the limit to heavy traffic (equivalently, the limit to the RBM) can be interchanged. Showing this interchange of limits is challenging in many systems because one needs to establish tightness of a sequence of probability measures. Although this method has been used successfully to study a wide variety of problems that satisfy the CRP condition, it is challenging to study systems where the CRP condition is not satisfied.

In addition, three different *direct methods* were developed to study queuing systems in heavy traffic without considering the scaled process and the diffusion limits (Dai 2018). Therefore, none of these direct methods requires the interchange of limits step. They are the drift method (Eryilmaz and Srikant 2012, Maguluri and Srikant 2016, Maguluri et al. 2018, Wang et al. 2018, Zhou et al. 2018), Stein's method (Gurvich 2014, Braverman et al. 2017a, Braverman and Dai 2017), and the basic adjoint relationship (BAR) method (Braverman et al. 2017b). We briefly describe each of them.

The main idea in the drift method is to carefully choose a test function and to equate the expected value of the test function in steady state to the same value at the following time step. Equating the expected value of the test function in two different time steps is also known as *setting to zero the drift of the test function* (see Definition 1 for a formal definition of this expression). Because this method does not involve the use of diffusion limits, SSC must be established independently, and this is done using the Lyapunov drift arguments and the moment bounds developed by Hajek (1982) and Bertsimas et al. (2001). When selecting a test function, one needs to keep in mind that one of the reasons to perform heavy-traffic analysis is SSC. Therefore, test functions that depend on the geometry of the region where SSC occurs yield bounds that are tight in heavy traffic. Usually, if quadratic test functions are used, bounds on the mean of the queue lengths are obtained. To obtain bounds on the mth moments, polynomial test functions of degree $(m+1)$ are used. The complete steady-state distribution in heavy traffic is obtained once all the moments are obtained inductively, under some mild conditions (see section 4.10 in Gut (2012) for a formal discussion of these conditions). For example, in the case of a single-server queue, the test functions $q, q^2, q^3, \ldots$ are used inductively, where q denotes the queue length.

This approach was first used to reprove known heavy-traffic results in a class of queuing systems that satisfy the CRP condition (Eryilmaz and Srikant 2012) and include a load-balancing system and an ad hoc wireless network in the presence of interference and fading (time-varying channel conditions). The drift method was later applied successfully to obtain the heavy-traffic mean of the sum queue lengths even in systems that do not satisfy the CRP condition, such as the input-queued switch (Maguluri and Srikant 2016, Maguluri et al. 2018) and bandwidth-sharing networks (Wang et al. 2018). However, it was recently shown that when the CRP condition is not satisfied, the drift method with polynomial test functions does not have all the information needed to obtain all the higher moments and the distribution of the queue lengths (Hurtado-Lange and Maguluri 2019).

In this paper, we develop the moment-generating function (MGF) method in systems that satisfy the CRP condition by generalizing the drift method to directly study the stationary distribution (as opposed to the moments) in heavy traffic. The key insight is that instead of using the polynomial test functions of increasing degrees inductively as in the drift method, all the polynomials can be combined in Taylor series to obtain an exponential test function. For example, in the case of a single-server queue, combining $q, q^2, q^3, \ldots$ in Taylor series (with appropriate coefficients), we obtain $e^{\theta q}$ for some constant θ, and $E[e^{\theta} q]$ is the MGF of q. The MGF method is similar to the drift method in the sense that we use the same notion of SSC and we set to zero the drift of a carefully chosen test function in steady state. However, in the drift method, one needs to perform an inductive argument to compute the stationary distribution, whereas the MGF method immediately yields the stationary distribution.

While the drift method is based on setting the drift of carefully chosen polynomial test functions to zero, the BAR method uses carefully chosen exponential functions. The focus in the BAR method is to choose the exponential functions to get a handle on the jumps in a continuous time system under general arrivals and services. In this paper, we illustrate how the MGF method can be thought of as a natural generalization of the drift method using exponential test functions and, in that sense, is similar in spirit to the BAR method. Using the BAR method, it was shown by Braverman et al. (2017b) that in heavy traffic, the stationary distribution of a generalized Jackson network is identical to that of an appropriately defined RBM. In contrast, the focus in this paper is to incorporate SSC and to evaluate the closed-form stationary distribution in heavy traffic in a variety of systems under the CRP condition. Moreover, although the BAR method was developed to study continuous time systems, we focus on studying discrete time systems in this paper.

The drift method and the BAR method are focused on computing the stationary distribution of the scaled queue lengths in heavy traffic. By contrast, Stein's method is focused on computing rates of convergence to the limiting distribution. Stein's method for studying queuing systems was first introduced by Gurvich (2014). Erlang-A and Erlang-C queuing models were studied using Stein's method by Braverman et al. (2017a), and $M/Ph/n+M$ systems were studied by Braverman and Dai (2017). Similar to the MGF method, a key step in using Stein's method for some results is in establishing SSC. Stein's method was used to study load-balancing systems in the mean field regime (Ying 2016, 2017), in the Halfin–Whitt regime (Braverman 2020), and in the sub-Halfin–Whitt regimes (Liu and Ying 2019). Universal approximations for queues with abandonment were obtained using Stein's method by Huang and Gurvich (2018). More recently, a single-server queue in heavy traffic was studied using Stein's method by Gaunt and Walton (2020). Gurvich et al. (2013) studies Erlang-A system and obtains universal approximations using excursion-based analysis as opposed to using Stein's method.

In this paper, we develop the MGF method and illustrate its power to study a variety of queuing systems that satisfy the CRP condition. In order to introduce the method, and to showcase its simplicity, we first present a sketch of the MGF method in the case of a single-server queue operating in discrete time in Section 3.2. We show that the stationary distribution of scaled queue lengths in heavy-traffic limit converges to an exponential distribution. This is, of course, a classic result first proved by Kingman (1962a) using the diffusion-limit method and later by Eryilmaz and Srikant (2012) using the drift method.

We then develop the MGF method framework and apply it to load-balancing systems and generalized switches. In both cases, we study the queuing systems under some general conditions, and we exemplify with specific systems that satisfy those conditions. In Section 4, we study load-balancing systems and identify that the join-the-shortest-queue (JSQ; Winston 1977, Foschini and Salz 1978) and the power-of-two choices (Mitzenmacher 1996, 2001; Vvedenskaya et al. 1996) routing policies satisfy the assumptions. In Section 5, we study generalized switches (Stolyar 2004) under the CRP condition, operating under the maxweight scheduling algorithm (Tassiulas and Ephremides 1992). We also show that ad hoc wireless networks operating under the maxweight scheduling algorithm satisfy our assumptions. All these systems are assumed to satisfy the CRP condition, and they are operated under algorithms that ensure that SSC occurs into a one-dimensional subspace. We show that the stationary distribution of this one-dimensional component is exponential. In addition to MGFs, which are the two-sided Laplace transforms of the probability distribution, one may use other transforms such as one-sided Laplace transforms and characteristic functions. We present a brief discussion about other transform methods in Remark 2 at the end of Section 3.3.

The primary contribution of this paper is the development of the MGF method, which is a simple framework to compute the stationary distribution of the scaled vector of queue lengths in heavy traffic. This is done by considering the previously mentioned set of systems. The paper also shows how the MGF method can be thought of as a generalization of the drift method by considering a richer class of test functions. This class of test functions leads to substantially different proofs that are much simpler than in the drift method, as will be illustrated in the following sections. However, unlike the drift method, the MGF method does not involve an art of picking a test function because the test function is essentially the MGF. Although most of the results that we present have already been established in the literature using diffusion limits and the drift method the purpose of this paper is to develop a framework based on transform techniques and illustrate its power and simplicity. A secondary contribution is that the load-balancing system we consider is allowed to have correlated servers, and the generalized switch is allowed to have correlated arrival processes. Under the CRP condition and control algorithms that ensure SSC to a one-dimensional subspace, we show that even under correlated arrivals or services, the heavy-traffic-scaled stationary distribution continues to be exponential (Theorems 2 and 3, respectively). It is possible to allow for this generalization using other methods, but we illustrate the simplicity of such generalizations using the MGF method.

The focus of this paper is on queuing systems that satisfy the CRP condition. However, the long-term goal of developing the MGF method is to characterize the heavy-traffic stationary distribution of systems that do not satisfy the CRP condition, such as input-queued switches (Maguluri and Srikant 2016, Maguluri et al. 2018). This will form the basis for future work on input-queued switches, which is briefly discussed in Section 6. This approach is similar to the one taken in the development of the drift method, which was first proposed by Eryilmaz and Srikant (2012) to prove known results in systems under the CRP condition. The drift method was later generalized to study the input-queued switch when the CRP condition is not satisfied (Maguluri and Srikant 2016, Maguluri et al. 2018).

1.1. Notation

In this section, we introduce the notation that we will use in this paper. We use $P[A]$ to denote the probability of the event A, $E[X]$ to denote the expected value of the random variable X, $\text{Cov}[X, Y]$ to denote the covariance between the random variables X and Y, and $\text{Var}[X]$ to denote the variance of the random variable X. The indicator function of an event A is $\mathbb{1}_{\{A\}}$; that is, $\mathbb{1}_{\{A\}}$ is one if A is true and zero otherwise. Convergence in distribution is denoted by $\Longrightarrow$.

We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{R}^n$ to denote the set of n-dimensional vectors with real components. We use $\mathbb{R}_+$ and $\mathbb{R}^n_+$ to denote the set of nonnegative numbers and the set of n-dimensional vectors with nonnegative elements, respectively. Vectors are written in bold letters, and we use the same letter, but not bold, and with a subindex to denote their elements. For example, for a positive integer n, the vector $\mathbf{x} \in \mathbb{R}^n$ has elements $x_i \in \mathbb{R}$ for $i \in \{1, \dots, n\}$. We use $\mathbf{1}$ to denote a vector of ones and $\mathbf{0}$ to denote a vector of zeroes; that is, if $\boldsymbol{x} = \mathbf{1}$, then $x_i = 1$ for all $i \in \{1, \dots, n\}$, and if $\boldsymbol{x} = \mathbf{0}$, then $x_i = 0$ for all $i \in \{1, \dots, n\}$. The dot product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, and the Euclidean norm is denoted by $\|\boldsymbol{x}\|$; that is, $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. For each $i \in \{1, \dots, n\}$, we use $\boldsymbol{e}^{(i)}$ to denote the ith canonical vector, that is, a vector with elements $e_i^{(i)} = 1$ and $e_j^{(i)} = 0$ for all $j \neq i$. Given a fixed vector $\boldsymbol{c} \in \mathbb{R}^n$ and a parameter $b \in \mathbb{R}$, the set $\{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{c}, \boldsymbol{x} \rangle = b\}$ is a hyperplane, and the set $\{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{c}, \boldsymbol{x} \rangle \leq b\}$ is a half-space.

We say that $f(x)$ is $O(g(x))$ if $\lim_{x \downarrow 0} |f(x)/g(x)|$ is finite, and we say that $f(x)$ is $o(g(x))$ if $\lim_{x \downarrow 0} (f(x)/g(x)) = 0$.

2. Related Work

In this section, we present an overview of related work on heavy-traffic analysis of queuing systems in general, as well as the different systems that we study in particular.

MGFs have been used in the literature to study queuing systems, such as the classical analysis of the $M/G/1$ queue (Harrison and Patel 1992). However, here we use the MGF to study heavy-traffic-scaled queue lengths because the queue lengths go to infinity in the heavy-traffic limit. There has been only a little work in the literature that uses transform methods for heavy-traffic analysis. Characteristic functions were used by Köllerström (1974) and Kingman (1961) to study heavy-traffic queuing systems, and MGFs were used by Lehoczky (1996, 1997). In contrast, the primary focus of this work is to develop transform methods for heavy-traffic analysis that incorporate SSC.

The single-server queue was first studied in heavy traffic by Kingman (1961) using characteristic functions and tools from complex analysis. Köllerström (1974) also used characteristic functions to study single-server queues. The diffusion limits method to study queuing systems was developed by studying the single-server queue (Kingman 1962a). The well-known Kingman bound for the expected waiting time in a single-server queue was developed in the 1960s (Kingman 1962b), and later, Marshall (1968) used similar arguments to compute bounds on the second moment. These formed the basis for the drift method, which was developed by Eryilmaz and Srikant (2012). The single-server queue was also presented as an illustrative example of the BAR method (Braverman et al. 2017b). Most of these papers study the delay in $G/G/1$ queue in continuous time, which evolves according to Lindley's equation (Lindley 1952). Similar to Eryilmaz and Srikant (2012), in this paper, we study the queue lengths in discrete time. The queue lengths process evolves according to (3), which is equivalent to Lindley's equation for the waiting time of the $(k+1)$th customer in a $G/G/1$ queue. Consequently, the results established for queue lengths in discrete time can be easily extended to delay in continuous time.

The load-balancing system (also known as the *supermarket checkout model*) has been widely studied since the 1970s. It was shown that the JSQ policy minimizes the mean delay among the class of policies that do not know the job durations (Winston 1977, Weber 1978, Ephremides et al. 1980). Heavy-traffic optimality of the JSQ policy in a system with two servers was established by Foschini and Salz (1978) using the diffusion limits approach, where they also introduced the notion of SSC. Since then, the load-balancing system has been studied extensively both to improve performance and to decrease the complexity of the algorithms (Lu et al. 2011; Chen and Ye 2012; Ying 2016, 2017; Stolyar 2017; Ying et al. 2017; Eschenfeldt and Gamarnik 2018; Li et al. 2018; Zhou et al. 2018; Braverman 2020). One lower-complexity algorithm that has received attention is the power-of-two-choices algorithm (Mitzenmacher 1996, 2001; Vvedenskaya et al. 1996; Chen and Ye 2012). An exhaustive survey of the literature on load balancing is presented by van der Boor et al. (2018). The most relevant work for our purposes is the study of the JSQ policy under the drift method by Eryilmaz and Srikant (2012) and that of the power-of-two-choices algorithm by Maguluri et al. (2014).

The maxweight algorithm was first proposed by Tassiulas and Ephremides (1992) in the context of scheduling for downlinks in wireless base stations. This algorithm was later adapted to be used in a variety of systems, including ad hoc wireless networks, input-queued switches (McKeown et al. 1996), and cloud computing (Maguluri et al. 2014), and was generalized into the backpressure algorithm (Tassiulas and Ephremides 1992) in networks and was studied extensively by Stolyar (2004), Gupta and Shroff (2010), and Meyn (2008). The generalized switch model subsumes many of these systems and has been studied under the CRP condition using the diffusion limits approach (Stolyar 2004) and the drift method (Eryilmaz and Srikant 2012). We use the notion of SSC as developed by Eryilmaz and Srikant (2012). Dai and Lin (2008) generalize the results in Stolyar (2004) to SPNs where the jobs can join a queue after being served.

3. The MGF Method

In this section, we introduce the MGF method to compute the distribution of scaled queue lengths in heavy traffic. This section is organized as follows. In Section 3.1, we define a general queuing model; in Section 3.2, we introduce the method with a single-server queue as a simple example; and in Section 3.3, we describe the MGF method as a step-by-step procedure so that it can be applied in the context of a variety of queuing systems.

3.1. A General Queuing Model

We first introduce a general queuing model for an SPN that includes the single-server queue, the load-balancing system, and the generalized switch as special cases. We provide the details of each system in the corresponding section.

Consider a single-hop queuing system operating in discrete time, with n separate servers. Each server has an infinite buffer, where jobs line up if the server is busy. For $k \geq 1$ and $i \in \{1, \ldots, n\}$, let $q_i(k)$ be the number of jobs in the ith queue at the beginning of time slot k, that is, the number of jobs waiting to be served and the job that is being served (if any). Let $\boldsymbol{q}(k)$ be an n-dimensional vector with elements $q_i(k)$ for $i \in \{1, \ldots, n\}$. Given that the vector of queue lengths in time slot k is $\boldsymbol{q}(k)$, let $a_i(\boldsymbol{q}(k))$ be the number of arrivals to the ith queue in time slot k and $s_i(\boldsymbol{q}(k))$ be the potential number of jobs that can be served from queue i in time slot k. We say that $s_i(\boldsymbol{q}(k))$ is potential service because if there are not enough jobs in line, then fewer than $s_i(\boldsymbol{q}(k))$ jobs are processed. For ease of exposition, and with a slight abuse of notation, from now on we write $\boldsymbol{a}(k)$ and $\boldsymbol{s}(k)$ instead of $\boldsymbol{a}(\boldsymbol{q}(k))$ and $\boldsymbol{s}(\boldsymbol{q}(k))$, respectively. We assume that $a_i(k)$ and $s_i(k)$ are upper bounded by constants. Specifically, let $A_{\max}$ and $S_{\max}$ be finite constants such that $a_i(k) \leq A_{\max}$ and $s_i(k) \leq S_{\max}$ with probability one for all $i \in \{1, \ldots, n\}$ and all $k \geq 1$. The difference between potential and actual service is called *unused service*, which we denote as $u_i(\boldsymbol{q}(k))$. We also write $\boldsymbol{u}(k)$ instead of $\boldsymbol{u}(\boldsymbol{q}(k))$ from now on for ease of exposition. In some queuing systems, the control problem is to decide the vector $\boldsymbol{a}(k)$ in each time slot (e.g., the load-balancing system) and, in others, the vector $\boldsymbol{s}(k)$ (e.g., the generalized switch). We give more details about these selection processes in the systems that we study in Sections 4 and 5, respectively.

In each time slot, the order of events is as follows. First, queue lengths are observed. Second, given the vector of queue lengths $\boldsymbol{q}(k)$, the control problem is solved. Then arrivals occur, and at the end of each time slot, jobs are processed by the servers. Therefore, the dynamics of the queues are as follows:

$$q_i(k+1) = \max\{q_i(k) + a_i(k) - s_i(k), 0\}, \qquad \forall i \in \{1, \ldots, n\},\ \forall k \geq 1. \tag{1}$$

For each $i \in \{1, \ldots, n\}$, the variables $a_i(k)$ and $s_i(k)$ depend only on $\boldsymbol{q}(k)$ (or they are independent of $\boldsymbol{q}(k)$); then (1) implies that the process $\{\boldsymbol{q}(k) : k \geq 1\}$ is a Markov chain.

We can also describe the dynamics of the queues using unused service instead of the maximum as follows:

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k), \qquad \forall i \in \{1, \ldots, n\},\ \forall k \geq 1. \tag{2}$$

Observe that (2) implies that

$$q_i(k+1)u_i(k) = 0, \qquad \forall i \in \{1, \ldots, n\},\ \forall k \geq 1, \tag{3}$$

because the unused service is nonzero only when the potential service is larger than the number of jobs available to be served (queue lengths and arrivals), and in this case, the queue is empty in the next time slot. If $i \neq j$, then we do not necessarily have $q_i(k+1)u_j(k) = 0$ because the fact that queue j is empty at the end of time slot k does not imply that queue i will be empty at the beginning of time slot $k+1$, and vice versa. It turns out that getting a handle on the unused service plays an important role in heavy-traffic analysis, and (3) will be an important tool in the analysis. Equation (3) can be thought of as a defining property of the queuing process and is analogous to the Skorohod map (Skorokhod 1961).

In this paper, we add a line on top of the variables and vectors to denote steady state. Specifically, let $\bar{\boldsymbol{q}}$, $\bar{\boldsymbol{a}} \triangleq \boldsymbol{a}(\bar{\boldsymbol{q}})$, $\bar{\boldsymbol{s}} \triangleq \boldsymbol{s}(\bar{\boldsymbol{q}})$, and $\bar{\boldsymbol{u}} \triangleq \boldsymbol{u}(\bar{\boldsymbol{q}})$ be steady-state vectors that represent the queue lengths at the beginning of a time slot and arrivals, potential service, and unused service in one time slot in steady state, respectively. Let $\bar{\boldsymbol{q}}^+ \triangleq \bar{\boldsymbol{q}} + \bar{\boldsymbol{a}} - \bar{\boldsymbol{s}} + \bar{\boldsymbol{u}}$ denote the queue length at time $k+1$ in terms of the queue length, arrival, and service at time k, assuming that the system is in steady state. The precise definition of each of these steady-state vectors depends on the control problem, so we provide them in Section 3.2 for the single-server queue, in Section 4 for the load-balancing system, and in Section 5 for the generalized switch.

The MGF method will be used to compute the joint distribution of the scaled vector of queue lengths in heavy traffic, so before introducing the framework, we specify what we mean by heavy traffic and how we parametrize the queuing systems to obtain the limit. The heavy-traffic limit is the limit as the arrival-rate vector approaches the boundary of the capacity region of the system. The capacity region of an SPN is the set of arrival-rate vectors such that the system can be positive recurrent. In other words, if the arrival-rate vector is in the interior of the capacity region, there exists an algorithm that solves the control problem and is such that the queue lengths process is positive recurrent; if the vector of arrival rates is outside the capacity region, no algorithm can ensure positive recurrence. We use $\mathscr{C}$ to denote the capacity region, and we parametrize the heavy-traffic limit as follows. Take $\epsilon > 0$ and consider a set of queuing systems parametrized by ϵ. The parametrization is such that ϵ represents how far away the vector of arrival rates is from a fixed point $\boldsymbol{r}$ in the boundary of $\mathscr{C}$. Then the heavy-traffic limit is the limit as $\epsilon \downarrow 0$. In this paper, we add a superscript (ϵ) when we refer to the parametrized queuing system. More details on the parametrization of each queuing system will be provided once the models are completely specified, that is, in Section 3.2 for the single-server queue, in Section 4 for the load-balancing system, and in Section 5 for the generalized switch.

Before introducing the MGF framework in the context of a single-server queue, we formally define the drift of a function, and we explain what *set the drift to zero* means.

Definition 1 (Drift of a Function). Let $V : \mathbb{R}^n \to \mathbb{R}_+$ be a function. We define the drift of V at $\boldsymbol{q}$ as

$$\Delta V(\boldsymbol{q}) \triangleq \big(V(\boldsymbol{q}(k+1)) - V(\boldsymbol{q}(k))\big)\mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}}.$$

If $E[V(\boldsymbol{q}(k))] < \infty$ for k such that the Markov chain $\{\boldsymbol{q}(k) : k \geq 1\}$ is in steady state, we say that we set the drift of V to zero when we use the property

$$E\big[\Delta V(\boldsymbol{q}(k))\big] = 0,$$

where the expected value is taken with respect to the stationary distribution.

Observe that we can set to zero the drift of any function with finite expected value by definition of steady state.

3.2. MGF Method in the Single-Server Queue

Before presenting the details of the MGF framework, we use it in the simplest queuing system: a single-server queue. We provide a proof of the well-known result that the scaled queue length of a single-server queue has an exponential distribution in heavy traffic to illustrate the method and to show its simplicity. We do not provide all the details of our proofs because the single-server queue is a special case of the load-balancing system ($n = 1$), and this system is studied in detail in Section 4.

Consider a single-server queue operating in discrete time. Arrivals and potential service in each time slot are assumed to be independent sequences of independent and identically distributed (i.i.d.) random variables. Because they are also assumed to be finite with probability one (as specified in Section 3.1), their MGFs $E[e^{\theta a(1)}]$ and $E[e^{\theta s(1)}]$ exist for all $\theta \in \mathbb{R}$.

Let $\lambda \triangleq E\big[a(1)\big]$ and $\mu \triangleq E\big[s(1)\big]$. Observe that λ and μ are the rates of arrival and service, respectively, because they are the expected number of arrival/services in one time slot. Then the capacity region of the single-server queue is $\mathscr{C} = \{\lambda \in \mathbb{R}_+ : \lambda \leq \mu\}$. We consider a set of single-server queues parametrized by ϵ with a fixed service process of rate μ and arrival rate $\lambda^{(\epsilon)} \triangleq \mu - \epsilon$.

Let $\bar{a}^{(\epsilon)}$ and $\bar{s}$ be steady-state random variables that have the same distribution as $a^{(\epsilon)}(1)$ and $s(1)$, respectively. Then $\lambda^{(\epsilon)} = E[\bar{a}^{(\epsilon)}]$ and $\mu = E[\bar{s}]$. Let $(\sigma_a^{(\epsilon)})^2 = \mathrm{Var}[\bar{a}^{(\epsilon)}]$ and $\sigma_s^2 = \mathrm{Var}[\bar{s}]$.

In the rest of this section, we prove Theorem 1. This is a well-known result, and there are proofs using diffusion limits (Kingman 1962a) and the drift method (Eryilmaz and Srikant 2012) in the literature. We present an alternate proof that is simpler than the two proofs mentioned previously and will serve as a template for the MGF method.

Theorem 1. *Let $\epsilon \in (0, \mu)$, and consider a set of single-server queues parametrized by ϵ, as described previously. Let $\bar{q}^{(\epsilon)}$ be a steady-state random variable such that $\{q^{(\epsilon)}(k) : k \geq 1\}$ converges in distribution to $\bar{q}^{(\epsilon)}$ as $k \uparrow \infty$. Furthermore, assume that $\lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a$. Then $\epsilon \bar{q}^{(\epsilon)} \Rightarrow \Upsilon$ as $\epsilon \downarrow 0$, where Υ is an exponential random variable with mean $(\sigma_a^2 + \sigma_s^2)/2$.*

It is well known that for all $\epsilon \in (0, \mu)$, the Markov chain $\{q^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent. For instance, the reader can find a proof using the Foster–Lyapunov theorem in Srikant and Ying (2014, theorem 3.4.2). Then $\bar{q}^{(\epsilon)}$ is well defined.

Before presenting the proof, we prove two lemmas. The first lemma is a different version of (3) and is key in the MGF method. For other queuing systems, we use a weaker version of this lemma that is sufficient for the MGF method (see Step 1 in Section 3.3 for more details).

Lemma 1. *Consider a single-server queue parametrized by* ϵ*, as described previously. Then, for all* $\alpha, \beta \in \mathbb{R}$ *and each* $k \geq 1$*, we have*

$$\left(e^{\alpha q^{(\epsilon)}(k+1)} - 1\right)\left(e^{-\beta u^{(\epsilon)}(k)} - 1\right) = 0.$$

Proof. It follows from (3) and because $e^x - 1 = 0$ if and only if $x = 0$. □

The next lemma gives the first moment of the unused service in steady state, and it will be used at the end of the proof of Theorem 1.

Lemma 2. *Consider a single-server queue parametrized by* $\epsilon \in (0, \mu)$*, as described previously. Then*

$$E\left[\bar{u}^{(\epsilon)}\right] = \epsilon.$$

Proof. We set to zero the drift of the linear test function $V_1(q) = q$, and we obtain

$$\begin{aligned} 0 &= E\left[\left(\bar{q}^{(\epsilon)}\right)^{+} - \bar{q}^{(\epsilon)}\right] \\ &= E\left[\left(\bar{q}^{(\epsilon)} + \bar{a}^{(\epsilon)} - \bar{s} + \bar{u}^{(\epsilon)}\right) - \bar{q}^{(\epsilon)}\right], \end{aligned}$$

where the last equality holds by definition of $(\bar{q}^{(\epsilon)})^{+}$. Rearranging terms, we obtain

$$E\left[\bar{u}^{(\epsilon)}\right] = E\left[\bar{s} - \bar{a}^{(\epsilon)}\right] = \mu - (\mu - \epsilon) = \epsilon. \quad \square$$

Now we prove Theorem 1.

Proof of Theorem 1. If we expand the product in Lemma 1 and rearrange terms, we obtain

$$e^{\theta\epsilon q^{(\epsilon)}(k+1)} - e^{\theta\epsilon\left(q^{(\epsilon)}(k) + a^{(\epsilon)}(k) - s(k)\right)} = 1 - e^{-\theta\epsilon u^{(\epsilon)}(k)}. \tag{4}$$

Observe that (4) holds for all $k \geq 1$. In particular, it holds in steady state. Also, it can be shown that $E[e^{\theta\epsilon\bar{q}^{(\epsilon)}}] < \infty$ in an interval around zero. We omit the proof because in Lemma 11 we provide a proof for the load-balancing system, which is a more general case. Therefore, $E[e^{\theta\epsilon(\bar{q}^{(\epsilon)})^{+}}] = E[e^{\theta\epsilon\bar{q}^{(\epsilon)}}]$. Taking the expected value with respect to the stationary distribution in (4), we obtain

$$E\left[e^{\theta\epsilon\bar{q}^{(\epsilon)}}\left(1 - e^{\theta\epsilon\left(\bar{a}^{(\epsilon)} - \bar{s}\right)}\right)\right] = 1 - E\left[e^{-\theta\epsilon\bar{u}^{(\epsilon)}}\right].$$

Because $\bar{a}^{(\epsilon)}$ and $\bar{s}$ are independent of the queue length, rearranging terms, we obtain

$$E\left[e^{\theta\epsilon\bar{q}^{(\epsilon)}}\right] = \frac{1 - E\left[e^{-\theta\epsilon\bar{u}^{(\epsilon)}}\right]}{1 - E\left[e^{\theta\epsilon\left(\bar{a}^{(\epsilon)} - \bar{s}\right)}\right]}. \tag{5}$$

Now we take the heavy-traffic limit. Observe that the right-hand side yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$. Then we take Taylor series of each term with respect to θ, around $\theta = 0$. The technical details of why this expansion can be done are established in Lemma 3, which is presented in Section 3.3. For the numerator, we obtain

$$\begin{aligned} 1 - E\left[e^{-\theta\epsilon\bar{u}^{(\epsilon)}}\right] &= \theta\epsilon E\left[\bar{u}^{(\epsilon)}\right] - \frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{u}^{(\epsilon)}\right)^2\right] + O\left(\epsilon^3\right) \\ &= \theta\epsilon^2 + O\left(\epsilon^3\right), \end{aligned} \tag{6}$$

where the last equality holds by Lemma 2 and because $E\left[(\bar{u}^{(\epsilon)})^2\right]$ is $O(\epsilon)$. Details of this argument will be provided in Section 4 for the load-balancing system (see Claim 1), but the main idea is that $\bar{u}^{(\epsilon)}$ is a bounded random variable. For the denominator, we obtain

$$\begin{aligned} 1 - E\left[e^{\theta\epsilon\left(\bar{a}^{(\epsilon)} - \bar{s}\right)}\right] &= -\theta\epsilon E\left[\bar{a}^{(\epsilon)} - \bar{s}\right] - \frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{a}^{(\epsilon)} - \bar{s}\right)^2\right] + O\left(\epsilon^3\right) \\ &= \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sigma_s^2 + \epsilon^2\right) + O\left(\epsilon^3\right), \end{aligned} \tag{7}$$

where the last step holds because $E[\bar{a}^{(\epsilon)}] = \mu - \epsilon$ and by the definition of variance.

If we replace (6) and (7) in (5) and cancel out $\theta\epsilon^2$ from the numerator and denominator, we obtain

$$E\left[e^{\theta\epsilon\bar{q}^{(\epsilon)}}\right] = \frac{1 + O(\epsilon)}{1 - \frac{\theta}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sigma_s^2\right) + O(\epsilon)}.$$

Therefore, taking the heavy-traffic limit, we obtain

$$\lim_{\epsilon\downarrow 0} E\left[e^{\theta\epsilon\bar{q}^{(\epsilon)}}\right] = \frac{1}{1 - \theta\left(\frac{\sigma_a^2+\sigma_s^2}{2}\right)}. \tag{8}$$

Because the right-hand side is the MGF of an exponential random variable with mean $(\sigma_a^2 + \sigma_s^2)/2$, Equation (8) implies that $\epsilon\bar{q}^{(\epsilon)}$ converges in distribution to such an exponential random variable (Gut 2012, theorem 9.5 in section 5). □

In this section, we exemplified the MGF method in an intuitive fashion for the simplest queuing system. In the next section, we generalize these steps for other queuing systems that satisfy the CRP condition.

3.3. General Framework

In the preceding section, we proved a well-known result using the MGF method in the case of the simplest queuing system, that is, the single-server queue. In this section, we describe the method in detail for more general queuing systems that satisfy the CRP condition. Before presenting the framework, we present a formal definition of the CRP condition. We use the definition provided by Stolyar (2004).

Definition 2 (CRP Condition). Consider a set of queuing systems parametrized by ϵ, as described in Section 3.1, where the capacity region is $\mathscr{C}$. Suppose that in heavy traffic (i.e., as $\epsilon\downarrow 0$), the vector of arrival rates approaches a point $\boldsymbol{r}$ in the boundary of $\mathscr{C}$. We say that the queuing system *satisfies* the CRP condition if the outer normal vector to $\mathscr{C}$ at $\boldsymbol{r}$ is unique up to a scalar coefficient.

This implies that the system can be operated such that all the servers pool together in the heavy-traffic limit (Harrison and López 1999, Williams 2000, Dai and Lin 2008). Intuitively, this means that the queuing system behaves as a one-dimensional queuing system (i.e., as a single-server queue) if it is operated under a good control algorithm. Therefore, the MGF method is essentially similar to the proof of Theorem 1 after one establishes SSC on a one-dimensional subspace of the state space.

In order to use the MGF method, one needs to make sure that two prerequisites are satisfied. We state them before presenting the framework.

Prerequisite 1 (Positive Recurrence). Prove that the Markov chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent for $\epsilon > 0$.

Positive recurrence is a requirement to make sure that there exists a steady-state random vector $\bar{\boldsymbol{q}}^{(\epsilon)}$ such that the queue lengths process $\{\boldsymbol{q}^{(\epsilon)}(k) : k \geq 1\}$ converges in distribution to $\bar{\boldsymbol{q}}^{(\epsilon)}$ as $k \uparrow \infty$.

Prerequisite 2 (SSC). Prove SSC into a one-dimensional subspace.

Let $\boldsymbol{c} \geq \boldsymbol{0}$ be the direction into which SSC occurs. For simplicity, we assume that $\|\boldsymbol{c}\| = 1$. Then $\mathscr{K} \triangleq \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \alpha\boldsymbol{c}, \alpha \geq 0\}$ is the cone where the state space collapses in heavy traffic. For any n-dimensional vector $\boldsymbol{x}$, let $\boldsymbol{x}_{\|} \triangleq \langle \boldsymbol{x}, \boldsymbol{c}\rangle \boldsymbol{c}$ be the projection of $\boldsymbol{x}$ on $\mathscr{K}$, and let $\boldsymbol{x}_{\perp} \triangleq \boldsymbol{x} - \boldsymbol{x}_{\|}$. In this step, it should be proved that $E[\|\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^2]$ is $o(1/\epsilon^2)$, which is equivalent to proving that $\epsilon^2 E[\|\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^2]$ is $o(1)$.

The queuing systems that we study in this paper actually exhibit a stronger form of SSC, where $E[\|\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^m]$ is $O(1)$ for all $m = 1, 2, \ldots$. However, a weaker form of SSC is studied by Wang et al. (2017, 2018).

From this notion of SSC, we conclude that

$$\lim_{\epsilon\downarrow 0} \epsilon^2 E\left[\|\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^2\right] = 0;$$

that is, $\epsilon\|\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|$ converges to zero in the mean-squares sense and therefore in probability.

In the case of the single-server queue, we did not have to verify Prerequisite 2 because the state space is already one-dimensional. Now we present the MGF method.

Step 1. Prove an equation of the form

$$E\left[\left(e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}-1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}-1\right)\right] \quad \text{is } o\left(\epsilon^2\right), \tag{9}$$

and compute an expression for the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$.

The key in the MGF method is to handle unused service and its interaction with the queue lengths, arrivals, and potential service. In the drift method, the unused service is handled with (3). However, in this case, we want to work with an exponential transform of the queue lengths, so we need to write (3) in a different way. In the case of the single-server queue, we used Lemma 1, which, in fact, is much stronger than what we actually use in the MGF method. For more general queuing systems, we use (9).

To prove an equation of the form of (9), it is essential to use SSC. After proving (9), we need to obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ that is valid for all traffic. We sketch some algebraic steps that are useful to do so. Expanding the product on the left-hand side of (9), we obtain

$$\begin{aligned}
&E\left[\left(e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}-1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}-1\right)\right]\\
&\quad = E\left[e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+-\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]-E\left[e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}\right]+1-E\left[e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right] && (10)\\
&\quad \overset{(a)}{=} E\left[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}+\bar{\boldsymbol{a}}^{(\epsilon)}-\bar{\boldsymbol{s}}^{(\epsilon)}\rangle}\right]-E\left[e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}\right]+1-E\left[e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]\\
&\quad \overset{(b)}{=} E\left[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}+\bar{\boldsymbol{a}}^{(\epsilon)}-\bar{\boldsymbol{s}}^{(\epsilon)}\rangle}\right]-E\left[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle}\right]+1-E\left[e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right], && (11)
\end{aligned}$$

where (a) holds by the dynamics of the queues described in (2) and by definition of $(\bar{\boldsymbol{q}}^{(\epsilon)})^+$, and (b) holds if the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists in an interval around zero (this must be proved). In such a case, by definition of steady state, we have $E[e^{\theta\epsilon\langle \boldsymbol{c},(\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}]=E[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle}]$, which is equivalent to setting to zero the drift of the test function $V(\boldsymbol{q})=e^{\theta\epsilon\langle \boldsymbol{c},\boldsymbol{q}\rangle}$.

Observe that when we first expand the product in (10), we obtain two terms that are related to the unused service (the first and last terms). We use (2) to deal with the first one, and we write $(\bar{\boldsymbol{q}}^{(\epsilon)})^+-\bar{\boldsymbol{u}}^{(\epsilon)}$ in terms of $\bar{\boldsymbol{q}}^{(\epsilon)}$, $\bar{\boldsymbol{a}}^{(\epsilon)}$, and $\bar{\boldsymbol{s}}^{(\epsilon)}$. The last term is the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle$, and we deal with it in the second step of the MGF method.

Using (11) in (9) and reorganizing terms, we obtain

$$E\left[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle}\left(1-e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{a}}^{(\epsilon)}-\bar{\boldsymbol{s}}^{(\epsilon)}\rangle}\right)\right]=1-E\left[e^{-\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]+o\left(\epsilon^2\right). \tag{12}$$

From (12), we can obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ that is valid for all traffic. However, the steps to obtain it depend on the properties of each queuing system. For example, in the case of the single-server queue, we know that the arrival and potential service processes are independent of the queue lengths. Then we can separate the product on the left-hand side, and we obtain (5).

Step 2. Bound unused service and take heavy-traffic limit.

Observe that the MGFs of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{a}}^{(\epsilon)}\rangle$ and $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{s}}^{(\epsilon)}\rangle$ exist for all $\theta\in\mathbb{R}$ because the random variables are bounded by assumption. Furthermore, by definition of unused service, we have $\boldsymbol{0}\le\bar{\boldsymbol{u}}^{(\epsilon)}\le\bar{\boldsymbol{s}}^{(\epsilon)}$ component-wise. Then the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{u}}^{(\epsilon)}\rangle$ exists for all $\theta\in\mathbb{R}$. Also, in Step 1 (before obtaining (11)), it was proved that the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists in an interval around zero. Therefore, as $\epsilon\downarrow 0$, Equation (12) yields $0=0$. As mentioned previously, depending on the queuing system, we will use different approaches to obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ that is valid for all traffic from (12). For example, in the case of the single-server queue, we obtained (5), which yields a $\frac{0}{0}$ form in the limit as $\epsilon\downarrow 0$. Therefore, to compute the heavy-traffic limit, we take Taylor series of each term around $\theta=0$, except for the MGF of $\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}^{(\epsilon)}\rangle$. To do this, we use the following lemma.

Lemma 3. *Let $X^{(\epsilon)}$ be a set of random variables indexed by $\epsilon>0$. Assume that $X^{(\epsilon)}$ is bounded for all ϵ; that is, there exists a constant $K_{\max}$ (that does not depend on ϵ) such that $X^{(\epsilon)}\le K_{\max}$ with probability one. Define $f_{\epsilon,X}(\theta)\triangleq e^{\theta\epsilon X^{(\epsilon)}}$. Then*

$$\left|E\left[f_{\epsilon,X}(\theta)\right]-1-\theta\epsilon E\left[X^{(\epsilon)}\right]-\frac{(\theta\epsilon)^2}{2}E\left[\left(X^{(\epsilon)}\right)^2\right]\right|\le C\epsilon^3,$$

where C is a finite constant. With a slight abuse of notation, we write the inequality as follows:

$$E\left[f_{\epsilon,X}(\theta)\right]=1+\theta\epsilon E\left[X^{(\epsilon)}\right]+\frac{(\theta\epsilon)^2}{2}E\left[\left(X^{(\epsilon)}\right)^2\right]+O\left(\epsilon^3\right). \tag{13}$$

We present the proof of Lemma 3 in Appendix A.

Remark 1. Because we are working with a bounded random variable, the proof that we presented of Lemma 3 was straightforward. However, in general, one needs an assumption on the existence of the MGF.

Expanding each term on the right-hand side of (12) in Taylor series according to Lemma 3 yields terms related to the moments of the unused service. As illustrated in the case of the single-server queue, it suffices to handle the first moment. To do this, we set to zero the drift of the linear test function $V_1(\boldsymbol{q}) = \langle \boldsymbol{c}, \boldsymbol{q}\rangle$; that is, we set $E[\langle \boldsymbol{c}, (\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle] = E[\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle]$ (which is finite because in Step 1 it was proved that the MGF of $\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists in an interval around zero). For example, see Lemma 2 in the case of the single-server queue, which is used in (6).

From this step, we obtain an expression for the limit as $\epsilon\downarrow 0$ of the MGF of $\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle$. This proves convergence in the distribution of $\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ to a random variable Υ, which turns out to be exponential in the cases we study in this paper. Then $\epsilon\bar{\boldsymbol{q}}_{\parallel}^{(\epsilon)} = \epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle \boldsymbol{c} \Rightarrow \Upsilon \boldsymbol{c}$ as $\epsilon\downarrow 0$ because $\boldsymbol{c}$ is a fixed vector. We also know from SSC in Prerequisite 2 that $\epsilon\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)} \to 0$ in probability as $\epsilon\downarrow 0$. Then, by Slutsky's theorem (Gut 2012, theorem 11.4 in section 5), we obtain that $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} = \epsilon\bar{\boldsymbol{q}}_{\parallel}^{(\epsilon)} + \epsilon\boldsymbol{q}_{\perp}^{(\epsilon)} \Rightarrow \Upsilon\boldsymbol{c}$ as $\epsilon\downarrow 0$.

Remark 2. In order to set $E[e^{\theta\epsilon\langle \boldsymbol{c}, (\bar{\boldsymbol{q}}^{(\epsilon)})^+\rangle}] = E[e^{\theta\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle}]$ in Step 1, one must first prove the existence of the MGF of $\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle$ in an interval around zero. An alternative approach (where this difficulty does not arise) is to use characteristic functions because they always exist. However, working with characteristic functions involves the use of complex analysis. Another way to overcome this difficulty is to use a one-sided Laplace transform, that is, to consider $\theta < 0$. The one-sided Laplace transform of $\epsilon\langle \boldsymbol{c}, \boldsymbol{q}\rangle$ always exists because ϵ, $\boldsymbol{c}$, and $\boldsymbol{q}$ are nonnegative. If one chooses to work with other transforms, such as the characteristic function or one-sided Laplace transform, to get around the issue of the existence of the MGF, then one needs to assume that certain moments exist in a counterpart of Lemma 3. For instance, theorem 2.3.3 in Lukacs (1970) can be used when one is working with characteristic functions.

4. Load-Balancing Systems

In this section, we use the MGF method in the context of load-balancing systems, also known as supermarket-checkout *systems*. We first define the model, and then we use the MGF method to prove that the steady-state distribution of the scaled vector of queue lengths is exponential in heavy traffic.

4.1. Load-Balancing Model

Consider a system with n separate queues, as described in Section 3.1. For each $i \in \{1, \ldots, n\}$, $\{s_i(k) : k \geq 1\}$ is a sequence of i.i.d. random variables with $\mu_i \triangleq E[s_i(1)]$, and let $\mu_\Sigma \triangleq \sum_{i=1}^n \mu_i$. We consider this system in a general setting, so we do not assume independence of the servers. For $i, j \in \{1, \ldots, n\}$, let $\mathrm{Cov}[s_i, s_j]$ be the covariance between $s_i(1)$ and $s_j(1)$. There is a single stream of arrivals that we model as a sequence $\{a(k) : k \geq 1\}$ of i.i.d. random variables such that $a(k)$ is the number of arrivals to the system in time slot k. In this queuing system, the control problem is to route the arrivals to one of the n queues in each time slot. We assume that the routing policy is fixed for all $k \geq 1$, but we do not assume any particular policy. After routing, $a_i(k)$ is the number of arrivals routed to the ith queue in time slot k for $i \in \{1, \ldots, n\}$. We assume that $a(k) \leq A_{\max}$ with probability one for all $k \geq 1$ and that the arrival process is independent of the queue length and service processes. The dynamics of the queues are according to (2). It is well known that the capacity region of the load-balancing system is $\mathscr{C} = \{\lambda \in \mathbb{R}_+ : \lambda \leq \mu_\Sigma\}$. A proof can be found in appendix A of Eryilmaz and Srikant (2012).

To study the heavy-traffic limit of this queuing system, we parametrize the arrival process as follows. For $\epsilon \in (0, \mu_\Sigma)$, we consider a load-balancing system as described previously, where the arrival process $\{a^{(\epsilon)}(k) : k \geq 1\}$ is such that $E[a^{(\epsilon)}(1)] = \mu_\Sigma - \epsilon$ and $\mathrm{Var}[a^{(\epsilon)}(1)] = (\sigma_a^{(\epsilon)})^2$. In other words, the arrival rate approaches the point $r = \mu_\Sigma$ in the boundary of $\mathscr{C}$ as $\epsilon\downarrow 0$. Because the capacity region $\mathscr{C}$ of the load-balancing system is one-dimensional, the CRP condition (as defined in Definition 2) is trivially satisfied.

4.2. MGF Method Applied to Load-Balancing Systems

In this section, we state the main theorem of this section and provide some examples, and in the next section, we prove the theorem using the MGF method as developed in Section 3.3. Before presenting the formal statement of the result, we introduce the following definitions.

Definition 3 (Throughput Optimality)**.** A routing algorithm $\mathscr{A}$ is *throughput optimal* for the load-balancing system described in Section 4.1 if the Markov chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \geq 1\}$ operating under $\mathscr{A}$ is positive recurrent for all $\epsilon \in (0, \mu_\Sigma)$.

Definition 4 (SSC). Consider a routing algorithm $\mathscr{A}$, and let

$$\mathscr{K} \triangleq \{x \in \mathbb{R}^n : \; x_i = x_j, \quad \forall i, j \in \{1, \dots, n\}\};$$

that is, $c = \frac{1}{\sqrt{n}}\mathbf{1}$. For any vector $y \in \mathbb{R}^n$, let $y_\parallel$ be the projection of y on $\mathscr{K}$, and let $y_\perp \triangleq y - y_\parallel$. We say that the algorithm $\mathscr{A}$ *satisfies* SSC if the load-balancing system described in Section 4.1 operating under $\mathscr{A}$ satisfies the following property:

$$E\left[\|\bar{q}_\perp^{(\epsilon)}\|^2\right] \text{ is } o\left(\frac{1}{\epsilon^2}\right),$$

where $\bar{q}^{(\epsilon)}$ is a steady-state random vector such that $\{q^{(\epsilon)}(k) : \; k \geq 1\}$ converges in distribution to $\bar{q}^{(\epsilon)}$ if it is positive recurrent.

Observe that if an algorithm $\mathscr{A}$ satisfies SSC (as defined previously), then SSC occurs in the one-dimensional space $\mathscr{K}$. Therefore, a load-balancing system operating under such an $\mathscr{A}$ behaves as a single-server queue in the heavy-traffic limit.

Now we formally present the result that we prove using the MGF method.

Theorem 2. *Let $\epsilon \in (0, \mu_\Sigma)$, and consider a set of load-balancing systems parametrized by ϵ, as described in Section 4.1. Suppose that the routing algorithm is throughput optimal and that it satisfies SSC. For each $\epsilon \in (0, \mu_\Sigma)$, let $\bar{q}^{(\epsilon)}$ be a steady-state random vector such that the queue lengths process $\{q^{(\epsilon)}(k) : \; k \geq 1\}$ converges in distribution to $\bar{q}^{(\epsilon)}$. Assume that the MGF of $\epsilon \sum_{i=1}^n \bar{q}_i$ exists; that is, $E[e^{\theta\epsilon \sum_{i=1}^n \bar{q}_i^{(\epsilon)}}] < \infty$ for $\theta \in [-\Theta, \Theta]$, where $\Theta > 0$ is a finite number, and that $\lim_{\epsilon\downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a$. Then $\epsilon\bar{q}^{(\epsilon)} \Longrightarrow \widetilde{\Upsilon}\mathbf{1}$ as $\epsilon\downarrow 0$, where $\widetilde{\Upsilon}$ is an exponential random variable with mean $(\sigma_a^2 + \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}[s_i, s_j])/(2n)$.*

Now we introduce two routing policies that satisfy SSC, as defined previously. We first define the policies.

Definition 5 (JSQ and Power of Two Choices). Consider a load-balancing system as described in Section 4.1. Then, for each $k \geq 1$, given the vector of queue lengths $q^{(\epsilon)}(k)$, a routing policy selects $i^*(k)$ and sends arrivals according to the following formula:

$$a_i^{(\epsilon)}(k) = \begin{cases} a^{(\epsilon)}(k), & \text{if } i = i^*(k), \\ 0, & \text{otherwise.} \end{cases}$$

a. The routing policy JSQ sends all arrivals in time slot k to the queue with the least number of jobs, breaking ties at random. Formally, under JSQ routing policy,

$$i^*(k) \in \underset{i \in \{1,\dots,n\}}{\arg\min} \left\{q_i^{(\epsilon)}(k)\right\},$$

breaking ties at random.

b. The routing policy power of two choices selects two queues uniformly at random, say $i_1, i_2 \in \{1, \dots, n\}$, and sends all arrivals in time slot k to the queue with the least number of jobs between those two, breaking ties at random. Formally, under the power of two choices, if queues i_1 and i_2 are selected, then

$$i^*(k) \in \underset{i \in \{i_1, i_2\}}{\arg\min} \left\{q_i^{(\epsilon)}(k)\right\},$$

breaking ties at random.

In the following two corollaries, we show that these routing policies satisfy the assumptions of Theorem 2, and therefore, the scaled vector of queue lengths in a load-balancing system operating under any of these policies has an exponential distribution in heavy traffic.

Corollary 1. *Consider a set of load-balancing systems parametrized by $\epsilon \in (0, \mu_\Sigma)$, as described in Section 4.1, operating under JSQ routing policy. Then $\epsilon\bar{q}^{(\epsilon)} \Longrightarrow \widetilde{\Upsilon}_1\mathbf{1}$ as $\epsilon\downarrow 0$, where $\widetilde{\Upsilon}_1$ is an exponential random variable with mean $(\sigma_a^2 + \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}[s_i, s_j])/(2n)$.*

A particular case of the queuing system described in Corollary 1 is the load-balancing system operating under JSQ with independent servers. In this case, $\sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}[s_i, s_j]$ reduces to the sum of variances of the servers. This is one of the systems studied by Eryilmaz and Srikant (2012).

Proof. We only need to show that JSQ is throughput optimal, that it satisfies SSC, and that there exists $\Theta > 0$ such that $E[e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i^{(\epsilon)}}] < \infty$ for all $\theta \in [-\Theta, \Theta]$. Eryilmaz and Srikant (2012) prove throughput optimality and SSC in the case of independent servers. However, their proofs hold for correlated servers. The proof of throughput optimality can be found in appendix A of Eryilmaz and Srikant (2012).

The SSC result proved by Eryilmaz and Srikant (2012) is stronger than the property presented in Definition 4. In fact, they prove that $E[\|\bar{\boldsymbol{q}}_\perp^{(\epsilon)}\|^m]$ is upper bounded by a constant for each $m = 1, 2, \ldots$. This clearly implies that Definition 4 is satisfied. We provide a sketch of their proof of SSC in Section B.1 of Appendix B.

The existence of an MGF of $\epsilon\sum_{i=1}^n \bar{q}_i^{(\epsilon)}$ in an interval around zero is proved in Section B.2 of Appendix B. □

Corollary 2. *Consider a set of load-balancing systems parametrized by ϵ, as described in Section 4.1, operating under power of two choices and where all the servers are identical. Then $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} \Longrightarrow \widetilde{\Upsilon}_2\mathbf{1}$ as $\epsilon\downarrow 0$, where $\widetilde{\Upsilon}_2$ is an exponential random variable with mean $(\sigma_a^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i, s_j])/(2n)$.*

Proof. Similar to the proof of Corollary 1, we check throughput optimality, SSC, and existence of MGF. Maguluri et al. (2014) prove SSC in the case of independent servers in section 4.3 of their article, but their proof holds true if this assumption is dropped. Their proof is along the lines of the proof for JSQ in Section B.1 of Appendix B, so we do not present it here. Throughput optimality can be proved using the Foster–Lyapunov theorem and the calculations that Maguluri et al. (2014) develop in the proof of SSC, and existence of MGF is similar to the case of JSQ. We omit these proofs in this paper because our goal is to introduce the MGF method. □

Observe that the assumption of identical servers is essential for the power-of-two-choices algorithm to be throughput optimal. The case where the servers are not identical was studied by Chen and Ye (2012) using the diffusion-limits approach. The routing policy there randomly selects d servers in each time slot, where the probability of choosing server i is proportional to its service rate μ_i for all $i \in \{1, \ldots, n\}$. Then the arrivals are sent to the server with the shortest queue among the d selected servers. Chen and Ye (2012) prove that this queuing system satisfies the CRP condition and that the distribution of the scaled vector of queue lengths is exponential. A similar result can be obtained using the MGF method once the SSC as in Definition 4 is established. This is a straightforward extension, and we do not present the details here because the focus is on illustrating the MGF approach.

In this section, we presented the main theorem of this section and two examples where the assumptions of the theorem are satisfied. Observe that in both cases we only needed to check that the conditions of the theorem are satisfied. In fact, if we want to prove that the scaled vector of queue lengths of the load-balancing system operating under any other routing policy has an exponential distribution, we only need to check these three assumptions.

4.3. Proof of Theorem 2

In the rest of this section, we prove Theorem 2 using the MGF method. Before presenting the proof, we specify notation.

Let $\bar{a}^{(\epsilon)}$ be a steady-state random variable with the same distribution as $a^{(\epsilon)}(1)$, and let $\bar{\boldsymbol{a}}^{(\epsilon)} \triangleq \boldsymbol{a}^{(\epsilon)}(\bar{\boldsymbol{q}})$ be the vector of arrivals to each queue after routing in steady state. The vector $\bar{\boldsymbol{u}}^{(\epsilon)}$ is defined as in Section 3.1. Observe that in this case the vector $\bar{\boldsymbol{s}}$ is independent of $\bar{\boldsymbol{q}}^{(\epsilon)}$, and it has the same distribution as $\boldsymbol{s}(1)$ because the potential service sequences $\{s_i(k) : k \geq 1\}$ are i.i.d. and independent of the queue lengths processes for each $i \in \{1, \ldots, n\}$.

Proof of Theorem 2. For ease of exposition, we omit the dependence on ϵ of the variables in this proof. We use the MGF method. Before applying the steps, we need to verify that the prerequisites are satisfied; that is, we need to check positive recurrence and SSC. In fact, one of the assumptions of the theorem is that the routing policy is throughput optimal. Therefore, for any $\epsilon > 0$, the Markov chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent. Also, SSC is satisfied by assumption. Now we go through the steps of the MGF method.

Step 1. Prove an equation of the form of (9), and compute an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle$.

We first prove the following lemma.

Lemma 4. *Consider a load-balancing system parametrized by ϵ, as described in Theorem 2. Then there exists $\theta_{\max} > 0$ finite such that for any real number $\theta \in [-\theta_{\max}, \theta_{\max}]$, we have*

$$E\left[\left(e^{\theta\epsilon\sum_{i=1}^n (\bar{q}_i^{(\epsilon)})^+} - 1\right)\left(e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i^{(\epsilon)}} - 1\right)\right] \quad \textit{is } o(\epsilon^2).$$

We present the proof of Lemma 4 in Section B.3 of Appendix B.

Because $\langle \boldsymbol{c}, \boldsymbol{q}\rangle = (\sum_{i=1}^n q_i)/\sqrt{n}$ proving an equation of the form of (9) is equivalent to Lemma 4 using $\theta/\sqrt{n}$ instead of θ. For ease of exposition, we work with θ in the rest of this proof.

Note that $P\big[\bar{a} - \sum_{i=1}^n \bar{s}_i \neq 0\big] > 0$ whenever $\epsilon > 0$. If we expand the product in the expression of Lemma 4 and we follow the steps sketched after Step 1 in Section 3.3, we obtain

$$E\Big[e^{\theta\epsilon \sum_{i=1}^n \bar{q}_i}\Big(1 - e^{\theta\epsilon \sum_{i=1}^n (\bar{a}_i - \bar{s}_i)}\Big)\Big] = 1 - E\Big[e^{-\theta\epsilon \sum_{i=1}^n \bar{u}_i}\Big] + o(\epsilon^2). \tag{14}$$

Recall that $\sum_{i=1}^n \bar{a}_i = \bar{a}$ and that $\bar{a}$ and $\bar{\mathbf{s}}$ are independent of $\bar{\boldsymbol{q}}$ by definition. Therefore, reorganizing terms, we obtain

$$E\Big[e^{\theta\epsilon \sum_{i=1}^n \bar{q}_i}\Big] = \frac{1 - E\Big[e^{-\theta\epsilon \sum_{i=1}^n \bar{u}_i}\Big] + o(\epsilon^2)}{1 - E\Big[e^{\theta\epsilon(\bar{a} - \sum_{i=1}^n \bar{s}_i)}\Big]}, \tag{15}$$

which gives an expression for the MGF of $\epsilon \sum_{i=1}^n \bar{q}_i$ that is valid for all traffic.

Step 2. Bound unused service and take heavy-traffic limit. Equation (15) yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$, just like (5), in the case of the single-server queue. Equivalently, we can observe that (14) yields $0 = 0$ in the limit as $\epsilon \downarrow 0$. Then we take Taylor series of the numerator and the denominator of (15) at $\theta = 0$ to obtain the limit. To take Taylor expansion, we use Lemma 3.

In order to bound the numerator, we need to compute $E\big[\sum_{i=1}^n \bar{u}_i\big]$, so we start with a lemma.

Lemma 5. *Consider a load-balancing system parametrized by $\epsilon \in (0, \mu_\Sigma)$, as described in* Section 4.1, *operating under a throughput optimal routing policy. Then*

$$E\left[\sum_{i=1}^n \bar{u}_i^{(\epsilon)}\right] = \epsilon.$$

Proof. We set to zero the drift of $V_1(\boldsymbol{q}) = \langle \boldsymbol{c}, \boldsymbol{q}\rangle$ in steady state. In this case, from the definition of $\mathcal{K}$ in Definition 4, we have $\boldsymbol{c} = (1/\sqrt{n})\mathbf{1}$. Then we obtain

$$\begin{aligned}
0 &= E\big[V_1(\bar{\boldsymbol{q}}^+) - V_1(\bar{\boldsymbol{q}})\big] \\
&= \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^n \bar{q}_i^+ - \sum_{i=1}^n \bar{q}_i\right] \\
&\overset{(a)}{=} \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^n (\bar{q}_i + \bar{a}_i - \bar{s}_i + \bar{u}_i) - \sum_{i=1}^n \bar{q}_i\right] \\
&\overset{(b)}{=} \frac{1}{\sqrt{n}} E\left[\bar{a} - \sum_{i=1}^n \bar{s}_i + \sum_{i=1}^n \bar{u}_i\right],
\end{aligned}$$

where (a) holds by definition of $\bar{\boldsymbol{q}}^+$, and (b) holds because $\bar{a} = \sum_{i=1}^n \bar{a}_i$ by definition of $\bar{a}$ and $\bar{a}_i$. Rearranging terms and canceling $1/\sqrt{n}$, we obtain

$$\begin{aligned}
E\left[\sum_{i=1}^n \bar{u}_i\right] &= \sum_{i=1}^n E[\bar{s}_i] - E[\bar{a}] \\
&\overset{(a)}{=} \sum_{i=1}^n \mu_i - (\mu_\Sigma - \epsilon) \\
&\overset{(b)}{=} \epsilon,
\end{aligned}$$

where (a) holds because $E[\bar{a}] = \mu_\Sigma - \epsilon$, and (b) holds by definition of μ_Σ. $\square$

Now we expand the numerator and denominator of (15) in Taylor series. We start with the numerator, and we obtain

$$\begin{aligned} 1 - E\left[e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i}\right] &= 1 - E\left[f_{\epsilon,-\sum_{i=1}^n \bar{u}_i}(\theta)\right] \\ &= \theta\epsilon E\left[\sum_{i=1}^n \bar{u}_i\right] - \frac{(\theta\epsilon)^2}{2} E\left[\left(\sum_{i=1}^n \bar{u}_i\right)^2\right] + O(\epsilon^3) \\ &= \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2} E\left[\left(\sum_{i=1}^n \bar{u}_i\right)^2\right] + O(\epsilon^3), \end{aligned} \tag{16}$$

where the last equality holds by Lemma 5. Now we need to bound the second moment of the sum of unused services.

Claim 1. *Consider a load-balancing system as described in Theorem 2. Then*

$$\frac{(\theta\epsilon)^2}{2} E\left[\left(\sum_{i=1}^n \bar{u}_i^{(\epsilon)}\right)^2\right] \text{ is } O(\epsilon^3).$$

We prove the claim in Section D.1 of Appendix D. Using the claim in (16), we obtain

$$1 - E\left[e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i}\right] = \theta\epsilon^2 + O(\epsilon^3). \tag{17}$$

For the denominator, we obtain

$$\begin{aligned} 1 - E\left[e^{\theta\epsilon\left(\bar{a} - \sum_{i=1}^n \bar{s}_i\right)}\right] &= 1 - E\left[f_{\epsilon,\left(\bar{a} - \sum_{i=1}^n \bar{s}_i\right)}(\theta)\right] \\ &= -\theta\epsilon E\left[\bar{a} - \sum_{i=1}^n \bar{s}_i\right] - \frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{a} - \sum_{i=1}^n \bar{s}_i\right)^2\right] + O(\epsilon^3) \\ &= \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i, s_j] + \epsilon^2\right) + O(\epsilon^3), \end{aligned} \tag{18}$$

where the last step holds because $E[\bar{a}] = \mu_\Sigma - \epsilon$, $E[\sum_{i=1}^n \bar{s}_i] = \mu_\Sigma$, and by the definition of covariance.

Using (17) and (18) in (15), and because $O(\epsilon^3)$ is $o(\epsilon^2)$, we obtain

$$E\left[e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i}\right] = \frac{\theta\epsilon^2 + o(\epsilon^2)}{\theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i,s_j] + \epsilon^2\right) + O(\epsilon^3)}.$$

Canceling $\theta\epsilon^2$ from the numerator and denominator, we obtain

$$E\left[e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i}\right] = \frac{1 + o(1)}{1 - \frac{\theta}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i,s_j]\right) + O(\epsilon)}.$$

Therefore, taking the limit, we obtain

$$\lim_{\epsilon\downarrow 0} E\left[e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i}\right] = \frac{1}{1 - \frac{\theta}{2}\left(\sigma_a^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i, s_j]\right)},$$

which is the MGF of an exponential random variable with mean $(\sigma_a^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i, s_j])/2$. Then $\epsilon\langle \mathbf{c}, \bar{\mathbf{q}}\rangle\mathbf{c} = (\epsilon/n)\sum_{i=1}^n \bar{q}_i\mathbf{1} \Rightarrow \widetilde{\Upsilon}\mathbf{1}$ as $\epsilon\downarrow 0$, where $\widetilde{\Upsilon}$ is an exponential random variable with mean $(\sigma_a^2 + \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}[s_i, s_j])/(2n)$.

Therefore, we conclude that $\epsilon\bar{\mathbf{q}}^{(\epsilon)} = \epsilon\bar{\mathbf{q}}_{\|}^{(\epsilon)} + \epsilon\bar{\mathbf{q}}_{\perp}^{(\epsilon)} \Rightarrow \widetilde{\Upsilon}\mathbf{1}$ as $\epsilon\downarrow 0$. □

5. Generalized Switch

In this section, we apply the MGF method in the context of a generalized switch operating under maxweight. We compute the distribution of the scaled vector of queue lengths in heavy traffic under the assumption that CRP is satisfied. The generalized switch is a model that was first introduced by Stolyar (2004), and it represents a generalization of a variety of queuing systems, such as the input-queued switch (McKeown et al. 1996), cloud computing (Maguluri et al. 2014), downlinks in wireless base stations (Tassiulas and Ephremides 1992), and so on.

5.1. Generalized Switch Model

Consider a system with n separate queues, as described in Section 3.1. For each $i \in \{1, \dots, n\}$, let $\{a_i(k) : k \geq 1\}$ be a sequence of i.i.d. random variables such that $a_i(k)$ is the number of arrivals to queue i in time slot k. For $i, j \in \{1, \dots, n\}$, let $\mathrm{Cov}\left[a_i, a_j\right]$ be the covariance between $a_i(1)$ and $a_j(1)$. The servers interfere with each other. Then the vector of service rates must satisfy feasibility constraints in each time slot. Additionally, there are conditions of the environment that affect these constraints. We group all these conditions in a random variable called *channel state*. For each $k \geq 1$, let $T(k)$ be the channel state in time slot k. The sequence of random variables $\{T(k) : k \geq 1\}$ is i.i.d., and it is independent of the queue lengths and arrival processes. We assume that the state space of the channel state is a finite set $\mathcal{T}$, and we let $\boldsymbol{\psi}$ be the probability mass function of $T(1)$; that is, for each $t \in \mathcal{T}$, the probability of observing state t is $\psi_t \triangleq P[T(1) = t]$. For each $t \in \mathcal{T}$, let $\mathcal{S}^{(t)}$ be the set of feasible service-rate vectors under channel state t. We also assume that if $\boldsymbol{x} \in \mathcal{S}^{(t)}$ for some $t \in \mathcal{T}$, then all vectors that are strictly dominated by $\boldsymbol{x}$ are feasible. In other words, if $\boldsymbol{y}$ is a nonnegative vector that satisfies $\boldsymbol{y} \leq \boldsymbol{x}$ component-wise, then $\boldsymbol{y}$ is also a feasible service rate vector if the channel state is t. In particular, the projection of $\boldsymbol{x} \in \mathcal{S}^{(t)}$ on each of the coordinate axes is a feasible service rate vector as well. We assume that $\mathcal{S}^{(t)}$ is finite for each $t \in \mathcal{T}$, so we only consider maximal feasible schedules and their projection on the coordinate axes in $\mathcal{S}^{(t)}$. With this assumption, we do not lose much generality because the vector $\boldsymbol{s}(k)$ is the potential (not actual) service rate vector, and we are interested in the heavy-traffic limit.

In this queuing system, the control problem (which is a scheduling problem) is to select $\boldsymbol{s}(k)$ in each time slot after realizing the channel state. Let $\boldsymbol{s}(k)$ be the solution of the scheduling problem in time slot k. Because $\mathcal{S}^{(t)}$ is finite for each $t \in \mathcal{T}$ and $\mathcal{T}$ is also finite, there exists a constant $S_{\max}$ such that $s_i(k) \leq S_{\max}$ for all $i \in \{1, \dots, n\}$ and all $k \geq 1$.

It is known (Eryilmaz and Srikant 2012) that the capacity region of this queuing system is

$$\mathcal{C} = \sum_{t \in \mathcal{T}} \psi_t \,\mathrm{ConvexHull}\left\{\mathcal{S}^{(t)}\right\}. \tag{19}$$

Providing a formal proof of (19) is beyond the scope of this paper, but we intuitively explain why it holds. First, suppose that the channel state is fixed and that the set of feasible service rate vectors is $\mathcal{S}^{(1)}$. Then the capacity region should have all vectors $\boldsymbol{x}$ that satisfy $\boldsymbol{x} \leq \boldsymbol{s}$ for all $\boldsymbol{s} \in \mathcal{S}^{(1)}$. Because $\mathcal{S}^{(1)}$ contains the projection of its elements on the coordinate axis, the set of such vectors $\boldsymbol{x}$ is $\mathrm{ConvexHull}\{\mathcal{S}^{(1)}\}$. Now, if we consider the channel state as a random variable, recall that ψ_t is the probability that the channel state is t, and if the channel state is t, then the set of feasible service rate vectors is $\mathcal{S}^{(t)}$. Then (19) just gives the capacity region associated with each channel state, weighted by the probability that each channel state is observed.

Recall that, by assumption, each set $\mathcal{S}^{(t)}$ is finite. Then, for each $t \in \mathcal{T}$, the set $\mathrm{ConvexHull}\{\mathcal{S}^{(t)}\}$ is the convex hull of finitely many points. Therefore, $\mathrm{ConvexHull}\{\mathcal{S}^{(t)}\}$ is a polytope, that is, a bounded polyhedron. Also, the state space of the channel state $\mathcal{T}$ is finite by assumption. Then (19) is the weighted sum of finitely many polytopes. This implies that $\mathcal{C}$ is also a polytope. In order to exploit this structure, we describe it as the intersection of a finite number of half-spaces, where each half-space defines a facet of $\mathcal{C}$. Let L be the minimal number of half-spaces required to describe $\mathcal{C}$, and for each $\ell \in \{1, \dots, L\}$, let $\boldsymbol{c}^{(\ell)} \in \mathbb{R}^n$ and $b^{(\ell)} \in \mathbb{R}$ be the parameters that define each facet of the polytope. In other words, we describe $\mathcal{C}$ as follows:

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}^n_+ : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \leq b^{(\ell)}, \text{ for all } \ell \in \{1, \dots, L\}\right\}. \tag{20}$$

Without loss of generality, we can assume that $\boldsymbol{c}^{(\ell)} \geq \mathbf{0}$, $\|\boldsymbol{c}^{(\ell)}\| = 1$, and $b^{(\ell)} > 0$ for all $\ell \in \{1, \dots, L\}$ because we assumed that the sets $\mathcal{S}^{(t)}$ contain the projection on the coordinate axes of all their feasible vectors. Therefore, the capacity region is coordinate convex. For each $\ell \in \{1, \dots, L\}$, let $\mathcal{F}^{(\ell)} \triangleq \{\boldsymbol{x} \in \mathcal{C} : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle = b^{(\ell)}\}$ be the ℓ^{th} facet of the polytope $\mathcal{C}$.

In this paper, we assume that the scheduling problem is solved using the maxweight algorithm in each time slot; that is, if the channel state is t, then the selected schedule satisfies

$$\boldsymbol{s}(k) \in \arg\max_{\boldsymbol{x}\in\mathcal{S}^{(t)}} \langle \boldsymbol{x}, \boldsymbol{q}(k)\rangle, \tag{21}$$

and ties are broken at random.

From (19) and (21), observe that the service rate vector $\boldsymbol{s}(k)$ does not necessarily belong to the capacity region $\mathscr{C}$ because $\psi_t \le 1$ for all $t \in \mathcal{T}$. To overcome this difficulty, we define the following random variable. For each $\ell \in \{1, \dots, L\}$ and each $t \in \mathcal{T}$, define the *maximum* $\boldsymbol{c}^{(\ell)}$*-weighted service rate available in channel state* t (Eryilmaz and Srikant 2012) as

$$b^{(t,\ell)} = \max_{\boldsymbol{s}\in\mathcal{S}^{(t)}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{s}\rangle. \tag{22}$$

In other words, given that the channel state is t, $b^{(t,\ell)}$ is a real number such that the hyperplane $\mathscr{H}^{(t,\ell)} = \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle = b^{(t,\ell)}\}$ is tangent to the boundary of ConvexHull$\{\mathcal{S}^{(t)}\}$. Let $\{B_\ell(k) : k \ge 1\}$ be a sequence of i.i.d. random variables such that $P[B_\ell(k) = b^{(t,\ell)}] = \psi_t$ and $\sigma^2_{B_\ell} \triangleq \mathrm{Var}[B_\ell(k)]$. In the next lemma, we present the relation between the random variable $B_\ell(1)$ and the parameter $b^{(\ell)}$ for each $\ell \in \{1, \dots, L\}$.

Lemma 6. *Consider a generalized switch as described previously. Then, for each* $\ell \in \{1, \dots, L\}$,

$$E\big[B_\ell(1)\big] = b^{(\ell)}.$$

Proof. By definition of the random variable $B_\ell(1)$, we have

$$\begin{aligned} E\big[B_\ell(1)\big] &= \sum_{t\in\mathcal{T}} \psi_t b^{(t,\ell)} \\ &\overset{(a)}{=} \sum_{t\in\mathcal{T}} \psi_t \max_{\boldsymbol{s}\in\mathcal{S}^{(t)}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{s}\rangle \\ &\overset{(b)}{=} \max_{\boldsymbol{s}\in\mathscr{C}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{s}\rangle \\ &\overset{(c)}{=} b^{(\ell)}, \end{aligned}$$

where (a) holds by the definition of $b^{(t,\ell)}$ given in (22), (b) holds by definition of the capacity region $\mathscr{C}$ given in (19), and (c) holds by definition of the ℓ^{th} facet and because the objective function in the maximization problem is linear. □

To perform heavy-traffic analysis, we fix a facet $\mathscr{F}^{(\ell)}$, and we study a set of generalized switches where the vector of arrival rates approaches a fixed point in the relative interior of $\mathscr{F}^{(\ell)}$. Formally, we fix $\boldsymbol{r}^{(\ell)}$ in the relative interior of $\mathscr{F}^{(\ell)}$, and we let $\epsilon \in (0,1)$. Then the system parametrized by ϵ is such that $E[\boldsymbol{a}^{(\epsilon)}(k)] = \boldsymbol{r}^{(\ell)} - \epsilon\boldsymbol{c}^{(\ell)}$, and $Cov[a_i^{(\epsilon)}, a_j^{(\epsilon)}]$ is the covariance between $a_i^{(\epsilon)}(1)$ and $a_j^{(\epsilon)}(1)$ for each $i, j \in \{1, \dots, n\}$. In this case, because the point $\boldsymbol{r} = \boldsymbol{r}^{(\ell)}$ of the boundary of the capacity region $\mathscr{C}$ is in the relative interior of the facet $\mathscr{F}^{(\ell)} = \{\boldsymbol{x} \in \mathscr{C} : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle = b^{(\ell)}\}$, the unique outer normal vector to the capacity region $\mathscr{C}$ at $\boldsymbol{r}$ is the outer normal vector to the facet $\mathscr{F}^{(\ell)}$; that is, it is $\boldsymbol{c}^{(\ell)}$. Therefore, the CRP condition as defined in Definition 2 is satisfied. Observe that if $\boldsymbol{r}$ is in the intersection of two (or more) facets, then the CRP condition is not satisfied because there is a range of vectors that are normal to $\mathscr{C}$ at $\boldsymbol{r}$.

5.2. MGF Method Applied to Generalized Switches

In this section, we state the main theorem of this section and provide some examples. In the next section, we prove the theorem.

Theorem 3. *Let* $\epsilon \in (0,1)$. *Given the* ℓ^{th} *facet of* $\mathscr{C}$, $\mathscr{F}^{(\ell)}$, *and a vector* $\boldsymbol{r}^{(\ell)}$ *in the relative interior of* $\mathscr{F}^{(\ell)}$, *consider a set of generalized switches operating under the maxweight algorithm parametrized by* ϵ, *as described in Section 5.1. For each* ϵ, *let* $\bar{\boldsymbol{q}}^{(\epsilon)}$ *be a steady-state vector such that the queue length process* $\{\boldsymbol{q}^{(\epsilon)}(k) : k \ge 1\}$ *converges in distribution to* $\bar{\boldsymbol{q}}^{(\epsilon)}$. *Furthermore, let* $\lim_{\epsilon\downarrow 0} \mathrm{Cov}[a_i^{(\epsilon)}, a_j^{(\epsilon)}] = \mathrm{Cov}[a_i, a_j]$ *for each* $i, j \in \{1, \dots, n\}$. *Then* $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} \Rightarrow \overline{\Upsilon}\boldsymbol{c}^{(\ell)}$ *as* $\epsilon\downarrow 0$, *where* $\overline{\Upsilon}$ *is an exponential random variable with mean* $\left(\sum_{i=1}^n \sum_{j=1}^n c_i^{(\ell)} c_j^{(\ell)} \mathrm{Cov}[a_i, a_j] + \sigma^2_{B_\ell}\right)/2$, *where* $c_i^{(\ell)}$ *is the ith element of* $\boldsymbol{c}^{(\ell)}$ *for each* $i \in \{1, \dots, n\}$.

In the next corollary, we present a particular example of a generalized switch operating under maxweight.

Corollary 3. *Consider a set of generalized switches parametrized by ϵ, as described in Section 5.1, operating under the maxweight algorithm. Suppose that $\mathcal{T}$ has only one element; that is, the channel state is fixed over time. Then $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} \Rightarrow \overline{\Upsilon}_2 \boldsymbol{c}^{(\ell)}$, where $\overline{\Upsilon}_2$ is an exponential random variable with mean $(\sum_{i=1}^n \sum_{j=1}^n c_i^{(\ell)} c_j^{(\ell)} \mathrm{Cov}[a_i, a_j])/2$.*

The queuing system described in Corollary 3 is also known as an *ad hoc wireless network*. In an ad hoc wireless network, we have $\sigma^2_{B_\ell} = 0$ because the channel state is not a random variable anymore. The input-queued switch or a cross-bar switch (Srikant and Ying 2014, Maguluri and Srikant 2016, Maguluri et al. 2018) is yet another system that is well studied. When only one port of the switch is saturated, it satisfies the CRP condition (Stolyar 2004) and forms a special case of Corollary 3. In the next section, we present the model and formalize this result.

5.3. MGF Method Applied to the Input-Queued Switch

An input-queued switch is a generalized switch where n is a perfect square; that is, there exists an integer N such that $n = N^2$. Then it can be represented as a square matrix, where the rows are input ports and the columns are output ports. The feasibility constraints are such that, in each time slot, at most one queue can be served from each input and output port, and all jobs take exactly one time slot to be processed. Therefore, the set of feasible service-rate vectors is analogous to permutation matrices of $N \times N$.

For each $i \in \{1, \ldots, N\}$, let $\boldsymbol{\chi}^{(i)}$ be the normalized indicator vector of row i; that is, it is such that for each $i' \in \{1, \ldots, n\}$, we have $\chi^{(i)}_{i'} = 1/\sqrt{N}$ if queue i' corresponds to row i of the switch and $\chi^{(i)}_{i'} = 0$ otherwise. Similarly, for each $j \in \{1, \ldots, N\}$, let $\tilde{\boldsymbol{\chi}}^{(j)}$ be the normalized indicator vector of column j. With this notation, we can write the capacity region of the input-queued switch as

$$\mathcal{C}_{\mathrm{switch}} \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^n_+ :\ \langle \boldsymbol{\chi}^{(i)}, \boldsymbol{x} \rangle \le 1,\ \langle \tilde{\boldsymbol{\chi}}^{(j)}, \boldsymbol{x} \rangle \le 1\ ,\ \forall i, j \in \{1, \ldots, N\} \right\},$$

which is the intersection of $L = 2N$ half-spaces.

Only one port can be saturated in heavy traffic to ensure that CRP condition is satisfied. Without loss of generality, assume that input port 1 is saturated; that is, we consider a vector $\boldsymbol{r}^{(1)} \in \mathcal{F}^{(1)}$, where $\mathcal{F}^{(1)} \triangleq \{\boldsymbol{x} \in \mathcal{C}_{\mathrm{switch}} : \langle \boldsymbol{\chi}^{(1)}, \boldsymbol{x} \rangle = 1\}$. For simplicity, we let $\boldsymbol{r}^{(1)} = \boldsymbol{\chi}^{(1)}$. Then the heavy-traffic parametrization for $\epsilon \in (0,1)$ is such that $\boldsymbol{\lambda}^{(\epsilon)} = (1-\epsilon)\boldsymbol{\chi}^{(1)}$. Unlike the generalized switch, for the input-queued switch, we do not give the scheduling algorithm. Instead, we write the result in terms of the conditions that this algorithm must satisfy (similar to the load-balancing case).

Similar to the case of the load-balancing system, we say that an algorithm $\mathcal{A}$ is throughput optimal for the input-queued switch if $\{\boldsymbol{q}^{(\epsilon)}(k) : k \ge 1\}$ is positive recurrent for all $\epsilon \in (0,1)$. Also, defining $\boldsymbol{x}_\parallel \triangleq \langle \boldsymbol{\chi}^{(1)}, \boldsymbol{x} \rangle \boldsymbol{\chi}^{(1)}$ and $\boldsymbol{x}_\perp \triangleq \boldsymbol{x} - \boldsymbol{x}_\parallel$ for any vector $\boldsymbol{x}$, we say that the switch operating under a scheduling algorithm $\mathcal{A}$ satisfies SSC if

$$E\left[\|\bar{\boldsymbol{q}}^{(\epsilon)}_\perp\|^2 \right] \text{ is } o\left(\frac{1}{\epsilon^2}\right).$$

In the next proposition, we compute the distribution of the scaled vector of queue lengths in heavy traffic.

Proposition 1. *Let $\epsilon \in (0,1)$, and consider a set of input-queued switches parametrized by ϵ, as described previously. Suppose that the scheduling algorithm is throughput optimal and that it satisfies SSC. For each $\epsilon \in (0,1)$, let $\bar{\boldsymbol{q}}^{(\epsilon)}$ be a steady-state random vector such that the queue length process $\{\boldsymbol{q}^{(\epsilon)}(k) : k \ge 1\}$ converges in distribution to $\bar{\boldsymbol{q}}^{(\epsilon)}$. Assume that the MGF of $\epsilon\langle \boldsymbol{\chi}^{(1)}, \bar{\boldsymbol{q}}^{(\epsilon)} \rangle$ exists. Then $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} \Longrightarrow \overline{\Upsilon}_s \boldsymbol{\chi}^{(1)}$ as $\epsilon \downarrow 0$, where $\overline{\Upsilon}_s$ is an exponential random variable with mean $(\sum_{i=1}^n \sum_{j=1}^n \chi_i^{(1)} \chi_j^{(1)} \mathrm{Cov}[a_i, a_j])/2$.*

Sketch of Proof. For ease of exposition, we do not write the dependence on ϵ of the variables. We use the MGF method. We only present a sketch of this proof because it is similar to the proofs of Theorems 2 and 3. We only show the main differences.

Both prerequisites are satisfied by assumption. Now we go through the steps.

Step 1. Prove an equation of the form of (9), and compute an expression for the MGF of $\epsilon\langle \boldsymbol{\chi}^{(1)}, \bar{\boldsymbol{q}}^{(\epsilon)} \rangle$.

Proving an equation of the form of (9) is similar to the proof of Lemmas 4 and 7. Then, following the steps sketched in Step 1 in Section 3.3, we obtain

$$E\left[e^{\theta\epsilon\langle \boldsymbol{\chi}^{(1)}, \bar{\boldsymbol{q}} \rangle} \left(1 - e^{\theta\epsilon\langle \boldsymbol{\chi}^{(1)}, \bar{\boldsymbol{a}} - \bar{\boldsymbol{s}} \rangle} \right) \right] = 1 - E\left[e^{-\theta\epsilon\langle \boldsymbol{\chi}^{(1)}, \bar{\boldsymbol{u}} \rangle} \right] + o\left(\epsilon^2\right).$$

Because $\bar{s}$ is a function of the queue lengths obtained through the scheduling problem, $\bar{s}$ is not independent of $\bar{q}$. However, $\langle \chi^{(1)}, \bar{s}\rangle = 1/\sqrt{N}$ because all the feasible schedules $\bar{s}$ are analogous to permutation matrices. Then the sum of all the elements of $\bar{s}$ corresponding to the first input port (row 1 of the switch) is one. Then $\langle \chi^{(1)}, \bar{s}\rangle$ is independent of the vector of queue lengths $\bar{q}$. Also, the vector of arrivals is independent of $\bar{q}$. Therefore, reorganizing terms, we obtain

$$E\left[e^{\theta\epsilon\langle \chi^{(1)},\bar{q}\rangle}\right] = \frac{1 - E\left[e^{-\theta\epsilon\langle \chi^{(1)},\bar{u}\rangle}\right] + o\left(\epsilon^2\right)}{1 - E\left[e^{\theta\epsilon\langle \chi^{(1)},\bar{a}-\bar{s}\rangle}\right]}.$$

Step 2. Bound unused service and take heavy-traffic limit.

This step is equivalent to Step 2 in the proofs of Theorems 2 and 3, so we omit the details. □

In the case of a generalized switch, one of the difficulties is to handle the dependence on the queue lengths of the potential service vector. In the case of an input-queued switch, this difficulty does not arise because, even though $\bar{s}^{(\epsilon)}$ depends on the queue lengths, the projection $s_{\|}^{(\epsilon)} \triangleq \langle \chi^{(1)}, \bar{s}^{(\epsilon)}\rangle \chi^{(1)}$ is independent of $\bar{q}^{(\epsilon)}$. Therefore, we do not need to assume that the scheduling problem is solved with maxweight. In general, for any special case of the generalized switch such that $s_{\|}^{(\epsilon)}$ is independent of the queue lengths, we can obtain a result similar to Proposition 1, that is, where we assume properties of the scheduling algorithm but not a specific algorithm.

5.4. Proof of Theorem 3

In the rest of this section, we prove Theorem 3 using the MGF method. Before presenting the proof, we introduce some notation.

Let $\bar{T}$ and $\bar{B}$ be steady-state random variables with the same distribution as $T(1)$ and $B_\ell(1)$, respectively.

Proof of Theorem 3. For ease of exposition, we omit the dependence on ϵ of the variables in this proof. We use the MGF method. Similar to the proof of Theorem 2, we first need to verify that the prerequisites are satisfied.

Prerequisite 1 (Positive Recurrence)**.** In fact, the maxweight algorithm is throughput optimal (Stolyar 2004, Eryilmaz and Srikant 2012). Then, for each $\epsilon > 0$, the Markov chain $\{q^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent.

Prerequisite 2 (SSC)**.** Let $\mathcal{K} = \{x \in \mathbb{R}^n_+ : x = \alpha c^{(\ell)}, \ \alpha \geq 0\}$. Using the notation introduced in Prerequisite 2 in Section 3.3, we have $c = c^{(\ell)}$, $\bar{q}_{\|}^{(\epsilon)} = \langle c^{(\ell)}, \bar{q}^{(\epsilon)}\rangle c^{(\ell)}$, and $\bar{q}_{\perp}^{(\epsilon)} = \bar{q}^{(\epsilon)} - \bar{q}_{\|}^{(\epsilon)}$. Eryilmaz and Srikant (2012) proved that $E[e^{\theta^*\|\bar{q}_\perp\|}]$ is bounded for some finite θ^*.[1] Then, for each $m = 1, 2, \ldots$, there exists a constant M_m such that $E[\|\bar{q}_{\perp}^{(\epsilon)}\|^m] \leq M_m$. Therefore, SSC as defined in Section 3.3 is satisfied, and it occurs in the one-dimensional subspace $\mathcal{K}$. In fact, in this case, $E[\|\bar{q}_{\perp}^{(\epsilon)}\|^m]$ is $O(1)$, which is stronger.

Now we go through the steps of the MGF method.

Step 1. Prove an equation of the form of (9), and compute an expression for the MGF of $\epsilon\langle c, \bar{q}^{(\epsilon)}\rangle$. We first prove Lemma 7.

Lemma 7. *Consider a generalized switch parametrized by ϵ, as described in Theorem 3. Then, for any real number θ such that $|\theta\epsilon| \leq \theta^*$, we have*

$$E\left[\left(e^{\theta\epsilon\langle c^{(\ell)},\left(\bar{q}^{(\epsilon)}\right)^+\rangle} - 1\right)\left(e^{-\theta\epsilon\langle c^{(\ell)},\bar{u}^{(\epsilon)}\rangle} - 1\right)\right] \quad \textit{is } o\left(\epsilon^2\right).$$

We present the proof of Lemma 7 in Section C.1 of Appendix C.

Before continuing, we need to prove that the MGF of $\epsilon\langle c^{(\ell)}, \bar{q}^{(\epsilon)}\rangle$ exists in an interval around zero. The proof is presented in Section C.2 of Appendix C. Then, following the steps sketched in Step 1 in Section 3.3, we obtain (12).

When we applied the MGF method to the single-server queue and to the load-balancing system, we used the fact that the service rate vector is independent of the queue length vector to obtain (5) and (15), respectively. However, in the case of the generalized switch, this is no longer true. To overcome this difficulty, we use the following lemma.

Lemma 8. *Consider a generalized switch operating under the maxweight algorithm parametrized by ϵ, as described in Theorem 3. Then, for any $\theta \in \mathbb{R}$, we have*

$$E\left[\left(e^{\theta\epsilon\langle c^{(\ell)},\bar{q}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(\bar{B}-\langle c^{(\ell)},\bar{s}^{(\epsilon)}\rangle\right)} - 1\right)\right] \quad \textit{is } o\left(\epsilon^2\right).$$

We present the proof in Section C.3 of Appendix C. Working with the left-hand side of (12), we obtain

$$\begin{aligned}
&E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\left(1-e^{\theta\epsilon\langle c^{(\ell)},\bar{a}-\bar{s}\rangle}\right)\right]\\
&\overset{(a)}{=} E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\left(1-e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right)\right]+E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\left(e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}-\bar{s}\rangle\right)}-e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right)\right]\\
&\overset{(b)}{=} E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\right]\left(1-E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]\right)-E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]\left(1-E\left[e^{\theta\epsilon\left(\bar{B}-\langle c^{(\ell)},\bar{s}\rangle\right)}\right]\right)\\
&\quad+E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]E\left[\left(e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}-1\right)\left(e^{\theta\epsilon\left(\bar{B}-\langle c^{(\ell)},\bar{s}\rangle\right)}-1\right)\right]\\
&\overset{(c)}{=} E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\right]\left(1-E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]\right)-E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]\left(1-E\left[e^{\theta\epsilon\left(\bar{B}-\langle c^{(\ell)},\bar{s}\rangle\right)}\right]\right)+o\left(\epsilon^2\right),
\end{aligned}$$

where (a) holds after adding and subtracting $E[e^{\theta\epsilon(\langle c^{(\ell)},\bar{q}\rangle+\langle c^{(\ell)},\bar{a}\rangle-\bar{B})}]$ and reorganizing terms; (b) holds because $\bar{a}$ and $\bar{B}$ are independent of the queue lengths vector $\bar{q}$ and the potential service vector $\bar{s}$ and after adding and subtracting $E[e^{\theta\epsilon(\langle c^{(\ell)},\bar{a}\rangle-\bar{B})}]E[e^{\theta\epsilon(\bar{B}-\langle c^{(\ell)},\bar{s}\rangle)}-1]$; and (c) holds by Lemma 8 and because $\bar{a}$ and $\bar{B}$ are bounded. Reorganizing terms, we obtain

$$E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{q}\rangle}\right]=\frac{1-E\left[e^{-\theta\epsilon\langle c^{(\ell)},\bar{u}\rangle}\right]+E\left[e^{\theta\epsilon\langle c^{(\ell)},\bar{a}\rangle}\right]E\left[e^{-\theta\epsilon\langle c^{(\ell)},\bar{s}\rangle}-e^{-\theta\epsilon\bar{B}}\right]+o\left(\epsilon^2\right)}{1-E\left[e^{\theta\epsilon\left(\langle c^{(\ell)},\bar{a}\rangle-\bar{B}\right)}\right]}. \tag{23}$$

Step 2. Bound unused service and take heavy-traffic limit. The right-hand side of (23) yields a $\frac{0}{0}$ form in the limit as $\epsilon\downarrow 0$. Then we take the Taylor expansion of each of its terms, using Lemma 3. Similar to the case of the load-balancing system, in this step, we need to obtain bounds on $E[\langle c^{(\ell)},\bar{u}\rangle]$. In this case, we use the following lemma.

Lemma 9. *Consider a generalized switch parametrized by ϵ, as described in Theorem 3. Then*

$$E\left[\langle c,\bar{u}^{(\epsilon)}\rangle\right]+b^{(\ell)}-E\left[\langle c^{(\ell)},\bar{s}^{(\epsilon)}\rangle\right]=\epsilon.$$

Proof. We set to zero the drift of $V_1(q)=\langle c^{(\ell)},q\rangle$. We obtain

$$\begin{aligned}
0&=E\left[\langle c^{(\ell)},\bar{q}^+\rangle-\langle c^{(\ell)},\bar{q}\rangle\right]\\
&=E\left[\langle c^{(\ell)},\bar{q}+\bar{a}-\bar{s}+\bar{u}\rangle-\langle c^{(\ell)},\bar{q}\rangle\right]\\
&=E\left[\langle c^{(\ell)},\bar{a}\rangle\right]-E\left[\langle c^{(\ell)},\bar{s}\rangle\right]+E\left[\langle c^{(\ell)},\bar{u}\rangle\right].
\end{aligned} \tag{24}$$

Now observe that

$$\begin{aligned}
E\left[\langle c^{(\ell)},\bar{a}\rangle\right]&=\langle c^{(\ell)},r^{(\ell)}-\epsilon c^{(\ell)}\rangle\\
&=\langle c^{(\ell)},r^{(\ell)}\rangle-\epsilon\|c^{(\ell)}\|^2\\
&\overset{(a)}{=} b^{(\ell)}-\epsilon,
\end{aligned} \tag{25}$$

where (a) holds because $r^{(\ell)}\in\mathcal{F}^{(\ell)}$ and because $\|c^{(\ell)}\|=1$.

Then, using (25) in (24) and rearranging terms, we obtain the result. □

Now we expand each term on the right-hand side of (23). For the first term in the numerator, we have

$$\begin{aligned}
1-E\left[e^{-\theta\epsilon\langle c^{(\ell)},\bar{u}\rangle}\right]&=1-E\left[f_{\epsilon,-\langle c^{(\ell)},\bar{u}\rangle}(\theta)\right]\\
&=\theta\epsilon E\left[\langle c^{(\ell)},\bar{u}\rangle\right]-\frac{(\theta\epsilon)^2}{2}E\left[\left(\langle c^{(\ell)},\bar{u}\rangle\right)^2\right]+O\left(\epsilon^3\right).
\end{aligned} \tag{26}$$

In this case, the numerator has more terms than in the case of the single-server queue and the load-balancing system, so we will keep the first moment of the unused service in the equation in order to use Lemma 9. However, we still need to bound the second moment.

Claim 2. *Consider a generalized switch, as described in Theorem 3. Then*

$$\frac{(\theta\epsilon)^2}{2}E\left[\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle\right)^2\right] \text{ is } O\left(\epsilon^3\right).$$

We present a proof of Claim 2 in Section D.2 of Appendix D. Then, using Claim 2 in (26), we obtain

$$1 - E\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle}\right] = \theta\epsilon E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle\right] + O\left(\epsilon^3\right). \tag{27}$$

For the second term in the numerator, we have

$$\begin{aligned} E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle}\right]E\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\bar{B}}\right] &= E\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)}\right]E\left[e^{\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)} - 1\right] \\ &= E\left[f_{\epsilon,\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)}(\theta)\right]E\left[f_{\epsilon,\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)}(\theta) - 1\right]. \end{aligned} \tag{28}$$

Claim 3. *Consider a generalized switch as described in Theorem 3 and the notation introduced in Lemma 3. Then*

$$E\left[f_{\epsilon,\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)}(\theta)\right] = 1 + \theta\epsilon^2 + O\left(\epsilon^3\right)$$
$$\text{and}\quad E\left[f_{\epsilon,\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)}(\theta) - 1\right] = \theta\epsilon E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right] + O\left(\epsilon^3\right).$$

We prove the claim in Section D.3 of Appendix D. Using Claim 3 in (28), reorganizing terms, and using that $\bar{B}$ and $\bar{s}_i$ are bounded for all $i \in \{1, \ldots, n\}$, we obtain

$$E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle}\right]E\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\bar{B}}\right] = \theta\epsilon E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right] + O\left(\epsilon^3\right). \tag{29}$$

Then the numerator of (23) yields

$$\begin{aligned} &1 - E\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}^{(\epsilon)}\rangle}\right] + E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle}\right]E\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\bar{B}}\right] + o\left(\epsilon^2\right) \\ &\quad= \left(\theta\epsilon E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle\right] + \theta\epsilon E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right] + O\left(\epsilon^3\right)\right) + o\left(\epsilon^2\right) \\ &\quad\overset{(a)}{=} \theta\epsilon\left(E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle + \bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right]\right) + o\left(\epsilon^2\right) \\ &\quad\overset{(b)}{=} \theta\epsilon^2 + o\left(\epsilon^2\right), \end{aligned} \tag{30}$$

where (a) holds because $O(\epsilon^3)$ is $o(\epsilon^2)$ and (b) holds by Lemmas 6 and 9.

For the denominator, we obtain

$$\begin{aligned} &1 - E\left[e^{-\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}, \bar{\boldsymbol{a}}\rangle\right)}\right] \\ &\quad= 1 - E\left[f_{\epsilon,\left(\langle \boldsymbol{c}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)}(\theta)\right] \\ &\quad= -\theta\epsilon E\left[\langle \boldsymbol{c}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right] - \frac{(\theta\epsilon)^2}{2}E\left[\left(\bar{B} - \langle \boldsymbol{c}, \bar{\boldsymbol{a}}\rangle\right)^2\right] + O\left(\epsilon^3\right) \\ &\quad\overset{(a)}{=} \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle^2\right] + E\left[\bar{B}^2\right] - 2E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle\bar{B}\right]\right) + O\left(\epsilon^3\right) \\ &\quad\overset{(b)}{=} \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_i^{(\ell)}c_j^{(\ell)}\operatorname{Cov}\left[a_i^{(\epsilon)}, a_j^{(\epsilon)}\right] + \sigma_{B_\ell}^2 + \left(E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle\right] - E\left[\bar{B}\right]\right)^2\right) + O\left(\epsilon^3\right) \\ &\quad\overset{(c)}{=} \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_i^{(\ell)}c_j^{(\ell)}\operatorname{Cov}\left[a_i^{(\epsilon)}, a_j^{(\epsilon)}\right] + \sigma_{B_\ell}^2 + \epsilon^2\right) + O\left(\epsilon^3\right), \end{aligned} \tag{31}$$

where (*a*) holds by (25) and expanding the square, (*b*) holds by definition of variance and covariance because $\bar{\boldsymbol{a}}$ and $\bar{B}$ are independent and reorganizing terms, and (*c*) holds by (25).

Using (30) and (31) in (23), we obtain

$$E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\right] = \frac{\theta\epsilon^2 + o(\epsilon^2)}{\theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\sum_{i=1}^n\sum_{j=1}^n c_i^{(\ell)}c_j^{(\ell)}\mathrm{Cov}\left[a_i^{(\epsilon)}, a_j^{(\epsilon)}\right] + \sigma_{B_\ell}^2 + \epsilon^2\right) + O(\epsilon^3)}$$
$$= \frac{1 + o(1)}{1 - \frac{\theta}{2}\left(\sum_{i=1}^n\sum_{j=1}^n c_i^{(\ell)}c_j^{(\ell)}\mathrm{Cov}\left[a_i^{(\epsilon)}, a_j^{(\epsilon)}\right] + \sigma_{B_\ell}^2 + \epsilon^2\right) + O(\epsilon)}.$$

Then taking the heavy-traffic limit yields

$$\lim_{\epsilon\downarrow 0} E\left[e^{\theta\epsilon\langle \boldsymbol{c},\bar{\boldsymbol{q}}\rangle}\right] = \frac{1}{1 - \frac{\theta}{2}\left(\sum_{i=1}^n\sum_{j=1}^n c_i^{(\ell)}c_j^{(\ell)}\mathrm{Cov}\left[a_i, a_j\right] + \sigma_{B_\ell}^2\right)},$$

which is the MGF of an exponential random variable with mean $\left(\sum_{i=1}^n\sum_{j=1}^n c_i^{(\ell)}c_j^{(\ell)}\mathrm{Cov}[a_i, a_j]\right)/2$ This implies that $\bar{\boldsymbol{q}}_{\parallel}^{(\epsilon)} = \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle \boldsymbol{c}^{(\ell)} \Rightarrow \overline{\Upsilon}\boldsymbol{c}^{(\ell)}$, where $\overline{\Upsilon}$ is an exponential random variable with mean $\left(\sum_{i=1}^n\sum_{j=1}^n c_i^{(\ell)}c_j^{(\ell)}\mathrm{Cov}[a_i, a_j]\right)/2$.

Then we conclude that $\epsilon\bar{\boldsymbol{q}}^{(\epsilon)} = \epsilon\bar{\boldsymbol{q}}_{\parallel}^{(\epsilon)} + \epsilon\bar{\boldsymbol{q}}_{\perp}^{(\epsilon)}$ converges in distribution to $\overline{\Upsilon}\boldsymbol{c}^{(\ell)}$ as $\epsilon\downarrow 0$. □

6. Future Work

This paper develops the MGF method, which we believe can be used to study more general sets of queuing systems. We outline a few of such future directions in this section.

In this paper, we assumed that the number of arrivals and services in one time slot are bounded. We believe that this assumption is not required and that it is sufficient to assume that the first two moments of the arrival and service sequences exist. Relaxing these assumptions is an immediate future work. We will explore two paths for this generalization. One is the use of characteristic functions or one-sided Laplace transforms instead of MGF because they always exist for nonnegative random variables. The main challenge in this approach is to establish the SSC under unbounded arrivals and service sequences. In this paper, we used the SSC established by Eryilmaz and Srikant (2012), which is based on the results from Hajek (1982), where the existence of all the moments of the arrival and service processes is assumed. We will explore ways to relax this assumption. The second approach that we will pursue is the MGF truncation arguments, similar to the ones introduced by Braverman et al. (2018) for Markov decision processes. The main idea of their method is to take a second-order Taylor expansion of the value function in order to solve the Bellman equations. We believe that this can give us insight into work with the second-order Taylor expansion of the MGF.

Another question for future research is to use the MGF method to study the rate of convergence to the heavy-traffic limit. In addition to obtaining the results on the heavy-traffic-limiting behavior, the drift method also gives upper and lower bounds that are applicable in all traffic (Eryilmaz and Srikant 2012, Maguluri and Srikant 2016, Maguluri et al. 2018). These bounds give the rate of convergence to the heavy-traffic limit. Because the MGF method is a natural generalization of the drift method, it may be used to obtain results on rate of convergence too, which is a topic for future study.

The next set of future work is on developing the MGF method for use in systems that do not satisfy the CRP condition, and this will be the culmination of the present work because the main motivation in developing the MGF method is to study systems where the CRP condition is not met. We believe that the MGF method is a promising approach to obtaining the heavy-traffic distribution of the queue lengths when CRP condition does not hold, even though the drift method is known to fail in this case (Hurtado-Lange and Maguluri 2019), for the following reason. The queue length process is a multidimensional discrete time Markov chain (DTMC, or a continuous Markov chain in some cases). For a positive recurrent and irreducible DTMC, it is known that the stationary distribution exists and is unique. One first establishes positive recurrence of the DTMC using the Foster–Lyapunov theorem. This has an added benefit that one typically obtains as a consequence of a (possibly loose) upper bound on an expression of the form $E[\epsilon\sum_i \bar{q}_i]$. If P is the transition matrix, then the stationary distribution is a unique solution of the equation $\pi = \pi P$. Clearly, solving for the stationary distribution in general is hard. However, we know that it is unique and is characterized by this equation. If we take a two-sided Laplace transform of the equation $\pi = \pi P$, we obtain an equation that is the same as the one we obtain by setting the drift of the exponential test function to zero. Because the Laplace transform is invertible, solving this equation uniquely characterizes the stationary distribution through its MGF. However, as shown in

Section 3.2, even for the single-server queue, it is challenging to obtain a solution for this equation in all traffic (see (5)). Therefore, using the MGF approach, we seek to solve it in the heavy-traffic limit. To do this, we first need to prove tightness of the sequence of the stationary distributions as the heavy-traffic parameter ϵ goes to zero. Tightness follows directly from the bound on $E[\epsilon \sum_i \bar{q}_i]$ that we obtain from the Foster–Lyapunov theorem. Therefore, we expect that the MGF drift equation that we have in the heavy-traffic limit must have a unique solution. Typically, because the system is tractable in steady state, we expect to solve this equation explicitly to get the joint stationary distribution in steady state. Even in cases where this equation may not be solved explicitly, we may be able to obtain moments from this equation. For instance, we may be able to obtain the moment bounds computed by Maguluri and Srikant (2016), Maguluri et al. (2018), and Wang et al. (2018) from such an equation.

Two systems of special interest that do not satisfy the CRP condition are the bandwidth-sharing network operating under proportional scheduling and the input-queued cross-bar switch operating under maxweight. The bandwidth-sharing network (Massoulié and Roberts 2000) operating under the so-called proportional scheduling algorithm is a good model for studying flow-level dynamics in data centers. If the arrivals are Poisson and job sizes are exponential, it is known that the stationary distribution in heavy traffic is the product of exponentials (Kang et al. 2009, Ye and Yao 2012). The bandwidth-sharing network is one of the simplest systems that does not satisfy the CRP condition because of this product form structure. It is also known that the stationary distribution of the corresponding RBM in the diffusion limit is insensitive to the job size distribution as long as it belongs to the class of phase-type distributions, which are known to be dense in the space of distributions (Vlasiou et al. 2014). However, the interchange of limits step was not shown by Vlasiou et al. (2014), so their result does not show whether the stationary distribution of the original system in heavy traffic is also insensitive. Recently, the drift method was used to complete this limit-interchange step (Wang et al. 2018). We will use the MGF method to directly study the stationary distribution in heavy traffic under phase-type arrivals using the MGF method to show insensitivity and to show that the stationary distribution is indeed the product of exponentials.

The input-queued cross-bar switch is an idealized model of a data-center network. It can be modeled as an $n \times n$ matrix of queues where the rows represent the input ports and the columns represent the output ports. Therefore, the dimension of the state space is n^2. Maguluri and Srikant (2016) studied an input-queued cross-bar switch operating under maxweight and proved that SSC occurs on a $(2n-1)$-dimensional cone. Moreover, the expected sum of the scaled queue lengths in heavy traffic was obtained using the drift method, resolving an open conjecture. Characterizing the higher moments and the distribution (marginals and joint) of scaled queue lengths are still open questions. The MGF method is developed in this paper with the goal of answering these questions given the limitation of the drift method to solve these problems (Hurtado-Lange and Maguluri 2019).

7. Conclusion

In this paper, we introduced transform methods to compute the steady-state distribution of scaled queue lengths in heavy traffic. We focused on a two-sided Laplace transform, which is also known as MGF. We motivated the method with a single-server queue, and we applied it in queuing systems that satisfy the CRP condition, such as load-balancing systems and the generalized switch. The main idea in the MGF method is to set the drift on an exponential test function to zero. The key step is in getting a handle on the unused service, and this paper illustrates how the unused service is handled in two different types of queuing systems. Further developing the MGF method to study a system where the CRP condition is not satisfied, such as the bandwidth-sharing network and the input-queued switch, forms future work.

Acknowledgments

The authors thank Professor Jim Dai for meaningful feedback on the proof of Lemma 4. The authors acknowledge the support from the Institute for Data and Decision Analytics (iDDA) at the Chinese University of Hong Kong, Shenzhen, during the summer of 2018.

Appendix A. Proof of Lemma 3

Fix $\Theta > 0$ and $x \in \mathbb{R}$. Then, from Taylor approximation of $f_{\epsilon,x}(\theta) = e^{\theta\epsilon x}$ at $\theta = 0$, we have

$$e^{\theta\epsilon x} \leq 1 + \theta\epsilon x + \frac{(\theta\epsilon)^2}{2} x^2 + \frac{(\tilde{\theta}\epsilon)^3}{3!} x^3, \quad \forall \theta \in [-\Theta, \Theta], \forall x \in \mathbb{R},$$

where $\tilde{\theta}$ is a real number between zero and θ. Then, for all $0 \le x \le K$, we have

$$e^{\theta\epsilon x} \le 1 + \theta\epsilon x + \frac{(\theta\epsilon)^2}{2}x^2 + \frac{(\tilde{\theta}\epsilon)^3}{3!}K^3.$$

Because $\tilde{\theta}$ is between zero and θ and $|\theta| \le \Theta$, we have

$$\left|\frac{(\tilde{\theta}\epsilon)^3}{3!}K^3\right| = \frac{|\tilde{\theta}|^3\epsilon^3}{3!}K^3 \le \frac{(\Theta\epsilon)^3}{3!}K^3,$$

which is finite for every ϵ. Then

$$e^{\theta\epsilon x} \le 1 + \theta\epsilon x + \frac{(\theta\epsilon)^2}{2}x^2 + \frac{(\Theta\epsilon)^3}{3!}K^3.$$

Therefore,

$$\left|e^{\theta\epsilon x} - 1 - \theta\epsilon x - \frac{(\theta\epsilon)^2}{2}x^2\right| \le C_1\epsilon^3,$$

where $C_1 = (\Theta K)^3/3!$ is a finite constant.

Now, because $X^{(\epsilon)} \le K_{\max}$ with probability one, we have

$$E\left[e^{\theta\epsilon X^{(\epsilon)}}\right] \le 1 + \theta\epsilon E\left[X^{(\epsilon)}\right] + \frac{(\theta\epsilon)^2}{2}E\left[\left(X^{(\epsilon)}\right)^2\right] + \frac{\Theta\epsilon^3 K_{\max}}{3!},$$

which proves the lemma. □

Appendix B. Details of the Proofs in Section 4

In this section, we provide the details of the proofs of the lemmas stated in Section 4.

B.1. Proof of SSC in the Load-Balancing System Operating Under JSQ

In this section, we present an insight into the proof of SSC as developed in Eryilmaz and Srikant (2012). They prove the result for the case where the servers are independent, but it also holds in the case where they are not. We first state the result.

Proposition B.1. *Consider a load-balancing system as described in Corollary 1. Then, for each $m = 1, 2, \ldots$, there exists a finite constant M_m such that*

$$E\left[\|\bar{\mathbf{q}}_{\perp}^{(\epsilon)}\|^m\right] \le M_m.$$

This proof is based on a lemma that was first proved by Hajek (1982). The original statement is more general than what we need here, so we present the specific result that we will use, as stated by Eryilmaz and Srikant (2012).

Lemma B.1. *For an irreducible and aperiodic Markov chain $\{X(k) : k \ge 1\}$ over a countable state space $\mathcal{X}$, suppose that $Z : \mathcal{X} \longrightarrow \mathbb{R}_+$ is a nonnegative-valued Lyapunov function. The drift of Z at x is*

$$\Delta Z(x) \triangleq [Z(X(k+1)) - Z(X(k))]\mathbb{1}_{\{X(k)=x\}}.$$

Thus, $\Delta Z(x)$ is a random variable that measures the amount of change in the value of Z in one step, starting from state x. This drift is assumed to satisfy the following conditions:

a. *There exists $\eta > 0$ and $\kappa < \infty$ such that*

$$E\left[\Delta Z(x) \mid X(k) = x\right] \le -\eta, \quad \text{for all } x \in \mathcal{X} \text{ with } Z(x) \ge \kappa.$$

b. *There exists $D < \infty$ such that*

$$|\Delta Z(x)| \le D, \quad \text{with probability one for all } x \in \mathcal{X}.$$

Then there exist $\theta^* > 0$ and $C^* < \infty$ such that

$$\limsup_{k\to\infty} E\left[e^{\theta^* Z(X(k))}\right] \le C^*.$$

If we further assume that the Markov chain $\{X(k): k \geq 1\}$ is positive recurrent, then $Z(X(k))$ converges in distribution to a random variable $\overline{Z}$ for which

$$E\left[e^{\theta^* \overline{Z}}\right] \leq C^*.$$

Proof. Eryilmaz and Srikant (2012) use the Lyapunov function $Z(\boldsymbol{q}) = \|\boldsymbol{q}_\perp^{(\epsilon)}\|$, and they prove that

$$E\left[\Delta Z(\boldsymbol{q}) \,|\, \boldsymbol{q}(k) = \boldsymbol{q}\right] \leq -\delta + \frac{n(\max\{A_{\max}, S_{\max}\})^2 + 2nS_{\max}{}^2}{2\|\boldsymbol{q}_\perp^{(\epsilon)}\|},$$

where δ is a fixed constant in $(0, \mu_{\min})$. The proof is based on the fact that $\|\boldsymbol{x}\| = \sqrt{\|\boldsymbol{x}\|^2}$, that square root is a concave function, and that JSQ sends all arrivals to the shortest queue in each time slot. This verifies condition (a) of Lemma B.1.

To verify condition (b), Eryilmaz and Srikant (2012) prove that for all $\boldsymbol{q} \in \mathbb{R}_+^n$,

$$|\Delta Z(\boldsymbol{q})| \leq 2\sqrt{n} \max\{A_{\max}, S_{\max}\},$$

using triangle inequality and boundedness of the arrival and service processes.

Also, for $\epsilon > 0$, the Markov chain $\{\boldsymbol{q}(k): k \geq 1\}$ is positive recurrent. In addition, because projection is nonexpansive, we have $\|\boldsymbol{q}_\perp^{(\epsilon)}(k)\| \leq \|\boldsymbol{q}^{(\epsilon)}(k)\|$, which implies that $\{\boldsymbol{q}_\perp(k): k \geq 1\}$ is positive recurrent. Therefore, by Lemma B.1, there exists $\theta^* > 0$ and $C^* > 0$ such that

$$E\left[e^{\theta^* \|\bar{\boldsymbol{q}}_\perp^{(\epsilon)}\|}\right] \leq C^*.$$

Finally, because $\|\bar{\boldsymbol{q}}_\perp^{(\epsilon)}\| \geq 0$, and $f(x) = e^x$ is a nonnegative increasing function, we obtain that $E\left[e^{\theta \|\bar{\boldsymbol{q}}_\perp^{(\epsilon)}\|}\right] \leq C^*$ for all $\theta \in [-\theta^*, \theta^*]$. This implies that for each $m = 1, 2, \ldots,$

$$E\left[\|\bar{\boldsymbol{q}}_\perp^{(\epsilon)}\|^m\right] \leq M_m. \quad \square$$

B.2. Existence of MGF of $\epsilon \sum_{i=1}^n \bar{q}_i^{(\epsilon)}$ in the Load-Balancing System Operating Under JSQ

We first state the result formally.

Lemma B.2. *Consider a load-balancing system operating under JSQ, parametrized by $\epsilon \in (0, \mu_\Sigma)$, as described in Corollary 1. Then, for each $\epsilon \in (0, \mu_\Sigma)$, there exists $\Theta > 0$ such that $E[e^{\theta\epsilon \sum_{i=1}^n \bar{q}_i^{(\epsilon)}}] < \infty$ for all $\theta \in [-\Theta, \Theta]$.*

Proof. We omit the dependence on ϵ of the variables for ease of exposition. First, observe that if $\theta \leq 0$, then $E[e^{\theta\epsilon \sum_{i=1}^n \bar{q}_i}] < \infty$ trivially because $\bar{\boldsymbol{q}} \geq \boldsymbol{0}$ by definition of queue length.

In the rest of this proof, we assume that $\theta > 0$. Observe that the function $f(x) = e^{\theta\epsilon x}$ is convex. Then, by Jensen's inequality, we have that for all $\boldsymbol{q} \geq \boldsymbol{0}$,

$$e^{\frac{\theta\epsilon}{n} \sum_{i=1}^n q_i} \leq \frac{1}{n} \sum_{i=1}^n e^{\theta\epsilon q_i}.$$

Hence, it suffices to show that $\sum_{i=1}^n E[e^{\theta\epsilon \bar{q}_i}] < \infty$ for $\theta < \Theta$. We use the Foster–Lyapunov theorem (Hajek 2015, proposition 6.13) with Lyapunov function $V(\boldsymbol{q}) = \sum_{i=1}^n e^{\theta\epsilon q_i}$.

Using Lemma 1 for each of the n queues and rearranging terms, we obtain that for each $i \in [n]$ and $k \geq 1$,

$$e^{\theta\epsilon q_i(k+1)} = 1 - e^{-\theta\epsilon u_i(k)} + e^{\theta\epsilon\left(q_i(k) + a_i(k) - s_i(k)\right)}.$$

Then, using the notation $E_{\boldsymbol{q}}[\cdot] \triangleq E[\cdot \,|\, \boldsymbol{q}(k) = \boldsymbol{q}]$, we obtain

$$\begin{aligned} E_{\boldsymbol{q}}\left[V(\boldsymbol{q}(k+1)) - V(\boldsymbol{q}(k))\right] &= \sum_{i=1}^n E_{\boldsymbol{q}}\left[e^{\theta\epsilon q_i(k+1)} - e^{\theta\epsilon q_i(k)}\right] \\ &= \sum_{i=1}^n \left(1 - E_{\boldsymbol{q}}\left[e^{-\theta\epsilon u_i(k)}\right]\right) + \sum_{i=1}^n e^{\theta\epsilon q_i}\left(E_{\boldsymbol{q}}\left[e^{\theta\epsilon(a_i(k) - s_i(k))}\right] - 1\right). \end{aligned}$$

Observe that because $E_{\boldsymbol{q}}[e^{-\theta\epsilon u_i(k)}] \geq 0$, we have

$$\sum_{i=1}^n \left(1 - E_{\boldsymbol{q}}\left[e^{-\theta\epsilon u_i(k)}\right]\right) \leq n.$$

Then it suffices to show that for some Θ and some $\eta > 0$, we have

$$\sum_{i=1}^{n} e^{\theta\epsilon q_i}\left(E_{\boldsymbol{q}}\left[e^{\theta\epsilon(a_i(k)-s_i(k))}\right]-1\right) \le -\eta, \qquad \forall \theta \in (0,\Theta].$$

Given $\boldsymbol{q}(k)=\boldsymbol{q}$, let $i^* \in \operatorname{argmin}_{i\in\{1,\dots,n\}}\{q_i(k)\}$ be the queue where arrivals in time slot k are routed. Then

$$\begin{aligned}\sum_{i=1}^{n} e^{\theta\epsilon q_i}\left(E_{\boldsymbol{q}}\left[e^{\theta\epsilon(a_i(k)-s_i(k))}\right]-1\right) &= e^{\theta\epsilon q_{i^*}}\left(E\left[e^{\theta\epsilon\left(a(k)-s_{i^*}(k)\right)}\right]-1\right)+\sum_{\substack{i=1\\ i\neq i^*}}^{n} e^{\theta\epsilon q_i}\left(E\left[e^{-\theta\epsilon s_i(k)}\right]-1\right)\\ &= e^{\theta\epsilon q_{i^*}}\,\theta M'_{a-s_{i^*}}(\xi_{i^*})+\sum_{\substack{i=1\\ i\neq i^*}}^{n} e^{\theta\epsilon q_i}\left(-\theta M'_{s_i}(\xi_i)\right),\end{aligned}$$

where we used the notation $M_X(\theta)=E[e^{\theta\epsilon X}]$, and ξ_i, ξ_{i^*} are numbers between zero and θ for all $i \neq i^*$. The second equality holds by Taylor expansion up to the first order of $M_{a-s_{i^*}}(\theta)$ and $M_{s_i}(\theta)$ for all $i\neq i^*$, around $\theta=0$.

Also, observe that $M'_{a-s_{i^*}}(0)=\epsilon(\lambda-\mu_{i^*})$ and $M'_{s_i}(0)=\epsilon\mu_i$ for all $i\neq i^*$, and MGF is continuous at $\theta=0$ (Mood 1950, p. 78). Then, for each $i\in\{1,\dots,n\}$, there exists Θ_i such that

$$\begin{aligned} M'_{s_i}(\xi_i) &\ge \frac{\epsilon\mu_i}{2}, \quad \forall |\theta| < \Theta_i \quad \text{for each } i\neq i^*\\ \text{and}\quad \left|M'_{a-s_{i^*}}(\xi_{i^*})\right| &\le \left|\frac{\epsilon(\lambda-\mu_{i^*})}{2}\right|, \quad \forall|\theta|<\Theta_{i^*}.\end{aligned}$$

Let $\Theta=\min_{i=1,\dots,n}\Theta_i$. Then, for all $\theta\in(0,\Theta]$, we have

$$\begin{aligned} E_{\boldsymbol{q}}\left[V(\boldsymbol{q}(k+1))-V(\boldsymbol{q}(k))\right] &\le n + e^{\theta\epsilon q_{i^*}}\left(\frac{\theta\epsilon\left(\lambda-\mu_{i^*}\right)}{2}\right)-\sum_{\substack{i=1\\ i\neq i^*}}^{n} e^{\theta\epsilon q_i}\left(\frac{\theta\epsilon\mu_i}{2}\right)\\ &\overset{(a)}{=} +\frac{\theta\epsilon}{2}\sum_{i=1}^{n}\lambda_i\left(e^{\theta\epsilon q_{i^*}}-e^{\theta\epsilon q_i}\right)+\frac{\theta\epsilon}{2}\sum_{i=1}^{n} e^{\theta\epsilon q_i}(\lambda_i-\mu_i)\\ &\overset{(b)}{\le} n+\sum_{i=1}^{n} e^{\theta\epsilon q_i}\left(\frac{\theta\epsilon\left(\lambda_i-\mu_i\right)}{2}\right)\\ &\overset{(c)}{=} n-\frac{\theta\epsilon^2}{2n}\sum_{i=1}^{n} e^{\theta\epsilon q_i},\end{aligned}$$

where $\lambda_i \triangleq \mu_i - \frac{\epsilon}{n}$. Here (a) holds by adding and subtracting $\sum_{i=1}^{n} e^{\theta\epsilon q_i}((\theta\epsilon\lambda_i)/2)$, realizing that $\lambda=\sum_{i=1}^{n}\lambda_i$ and rearranging terms; (b) holds because $q_{i^*}\le q_i$ for all i by definition of i^*; and (c) holds because $\mu_i-\lambda_i=\epsilon/n$. □

B.3. Proof of Lemma 4

To prove Lemma 4, we use the following result.

Lemma B.3. *Consider the load-balancing system indexed by ϵ, as described in Theorem 2. Then, for any $\alpha\in\mathbb{R}$ and for all $k\ge 1$, we have*

$$\sum_{i=1}^{n} u_i^{(\epsilon)}(k)\left(e^{\frac{\alpha}{n}\sum_{j=1}^{n} q_j^{(\epsilon)}(k+1)}-1\right)=\sum_{i=1}^{n} u_i^{(\epsilon)}(k)\left(e^{-\alpha q_{\perp i}^{(\epsilon)}(k+1)}-1\right),$$

where $q_{\perp i}^{(\epsilon)}(k)$ is the ith element of $\boldsymbol{q}_{\perp}^{(\epsilon)}(k)$ for each $i\in\{1,\dots,n\}$.

Proof. If $\alpha=0$, the equation trivially holds. So now assume that $\alpha\neq 0$. Because $q_i(k+1)u_i(k)=0$ for all $i\in\{1,\dots,n\}$, we have

$$u_i(k)\left(e^{-\alpha q_i(k+1)}-1\right)=0, \quad \forall i\in\{1,\dots,n\}.$$

Then, summing over $i\in\{1,\dots,n\}$, we obtain

$$\sum_{i=1}^{n} u_i(k)\left(e^{-\alpha q_i(k+1)}-1\right)=0.$$

By definition of $\boldsymbol{q}_{\|}(k)$ and $\boldsymbol{q}_{\perp}(k)$, we have $\boldsymbol{q}(k)=\boldsymbol{q}_{\|}(k)+\boldsymbol{q}_{\perp}(k)$, so

$$\sum_{i=1}^{n} u_i(k)\left(e^{-\alpha\left(q_{\|i}(k+1)+q_{\perp i}(k+1)\right)}-1\right)=0.$$

However, $\boldsymbol{q}_{\|}(k+1) = ((\sum_{j=1}^n q_j(k+1))/n)\mathbf{1}$, so $q_{\|i}(k+1) = q_{\|1}(k+1)$ for all $i \in \{1,\dots,n\}$. Then, reorganizing terms, we obtain

$$\sum_{i=1}^n u_i(k)e^{-\alpha q_{\perp i}(k+1)} = e^{\alpha q_{\|1}(k+1)} \sum_{i=1}^n u_i(k).$$

By definition of $\boldsymbol{q}_{\|}(k)$, we obtain

$$\sum_{i=1}^n u_i(k)e^{-\alpha q_{\perp i}(k+1)} = e^{\frac{\alpha}{n}\sum_{j=1}^n q_j(k+1)} \sum_{i=1}^n u_i(k).$$

Finally, subtracting $\sum_{i=1}^n u_i(k)$ from both sides, we obtain

$$\sum_{i=1}^n u_i(k)\left(e^{\frac{\alpha}{n}\sum_{j=1}^n q_j(k+1)} - 1\right) = \sum_{i=1}^n u_i(k)\left(e^{-\alpha q_{\perp i}(k+1)} - 1\right). \quad \square$$

In the proof of Lemma 4, we use Lemma B.3 and the following facts:
1. The function $g(x) = (e^x - 1)/x$ is nonnegative and nondecreasing for all $x \in \mathbb{R}$.
2. Suppose that $0 \le x \le y$. Then, for all $\theta \in \mathbb{R}$, we have $e^{\theta x} - 1 \le (\theta x)((e^{\theta y} - 1)/(\theta y))$.
3. For all $x \in \mathbb{R}_+$, $(e^x - 1)/x < e^x$.

All these facts can be shown using calculus techniques, so we omit the proof. Now we prove Lemma 4.

Proof of Lemma 4. First, observe that if $\theta = 0$, the statement trivially holds. If $\theta \neq 0$, by properties of expectation and absolute value, we obtain

$$\begin{aligned}
&\left|E\left[\left(e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i^+} - 1\right)\left(e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i} - 1\right)\right]\right| \\
&\le E\left[\left|\left(e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i^+} - 1\right)\left(e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i} - 1\right)\right|\right] \\
&\overset{(a)}{=} |\theta|\epsilon E\left[\left|\left(\sum_{i=1}^n \bar{u}_i\right)\left(e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i^+} - 1\right)\left(\frac{e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i} - 1}{-\theta\epsilon\sum_{i=1}^n \bar{u}_i}\right)\right|\mathbb{1}_{\{\sum_{i=1}^n \bar{u}_i \neq 0\}}\right] \\
&\overset{(b)}{\le} |\theta|\epsilon\left(\frac{e^{|\theta|\epsilon n S_{\max}} - 1}{|\theta|\epsilon n S_{\max}}\right)E\left[\left|\sum_{i=1}^n \bar{u}_i\left(e^{\theta\epsilon\sum_{j=1}^n \bar{q}_j^+} - 1\right)\right|\right] \\
&\overset{(c)}{\le} |\theta|\epsilon\left(\frac{e^{|\theta|\epsilon n S_{\max}} - 1}{|\theta|\epsilon n S_{\max}}\right)E\left[\sum_{i=1}^n \bar{u}_i\left|e^{-\theta\epsilon n\bar{q}_{\perp i}} - 1\right|\right] \\
&\overset{(d)}{\le} |\theta|\epsilon\left(\frac{e^{|\theta|\epsilon S_{\max}} - 1}{|\theta|\epsilon S_{\max}}\right)E\left[\sum_{i=1}^n \bar{u}_i^p\right]^{\frac{1}{p}} E\left[\sum_{i=1}^n \left|e^{-\theta\epsilon n\bar{q}_{\perp i}} - 1\right|^{\frac{p}{p-1}}\right]^{\frac{p-1}{p}} \\
&\overset{(e)}{\le} |\theta|\epsilon^{1+\frac{1}{p}} S_{\max}^{\frac{p-1}{p}}\left(\frac{e^{|\theta|\epsilon S_{\max}} - 1}{|\theta|\epsilon S_{\max}}\right)E\left[\sum_{i=1}^n \left|e^{-\theta\epsilon n\bar{q}_{\perp i}} - 1\right|^{\frac{p}{p-1}}\right]^{\frac{p-1}{p}} \\
&= \theta^2\epsilon^{2+\frac{1}{p}} S_{\max}^{\frac{p-1}{p}} n\left(\frac{e^{|\theta|\epsilon S_{\max}} - 1}{|\theta|\epsilon S_{\max}}\right)\left(\sum_{i=1}^n E\left[\left|\frac{e^{-\theta\epsilon n\bar{q}_{\perp i}} - 1}{-\theta\epsilon n\bar{q}_{\perp i}}\right|^{\frac{p}{p-1}} |\bar{q}_{\perp i}|^{\frac{p}{p-1}}\mathbb{1}_{\{\bar{q}_{\perp i} \neq 0\}}\right]\right)^{\frac{p-1}{p}},
\end{aligned} \tag{B.1}$$

where $p > 1$. Here (a) holds because if $\sum_{i=1}^n \bar{u}_i = 0$, then $e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i} - 1 = 0$ and by multiplying and dividing everything by $|\theta\epsilon\sum_{i=1}^n \bar{u}_i|$; (b) holds by fact 1 stated previously because $\bar{u}_i \le S_{\max}$ for all $i \in \{1,\dots,n\}$ and because $0 \le \mathbb{1}_{\{\sum_{i=1}^n \bar{u}_i \neq 0\}} \le 1$; (c) holds by the triangle inequality and Lemma B.3; (d) holds by Hölder's inequality; and (e) holds because $\bar{u}_i \le S_{\max}$ for all $i \in \{1,\dots,n\}$, because $\sum_{i=1}^n E[\bar{u}_i] = \epsilon$, and because $x^{(1/p)}$ is an increasing function for $x \ge 0$.

By L'Hospital's rule, we have

$$\lim_{\epsilon\downarrow 0}\frac{e^{|\theta|\epsilon n S_{\max}} - 1}{|\theta|\epsilon n S_{\max}} = 1.$$

Then the last step is to prove that the last expression in (B.1) is $O(1)$. To do this, we show the following claim at the end of this section.

Claim B.1. *Consider a load-balancing system as described in Lemma 4. Then there exists $\theta_{\max} > 0$ finite such that for all $|\theta| < \theta_{\max}$, we have*

$$\left(\sum_{i=1}^n E\left[\left|\frac{e^{-\theta\epsilon n\bar{q}_{\perp i}} - 1}{-\theta\epsilon n\bar{q}_{\perp i}}\right|^{\frac{p}{p-1}} |\bar{q}_{\perp i}|^{\frac{p}{p-1}}\mathbb{1}_{\{\bar{q}_{\perp i} \neq 0\}}\right]\right)^{\frac{p-1}{p}} \text{ is } o(\epsilon^2).$$

An expression for $\theta_{\max}$ is provided in (B.2). Therefore,

$$E\left[\left(e^{\theta\epsilon\sum_{i=1}^n \bar{q}_i^+}-1\right)\left(e^{-\theta\epsilon\sum_{i=1}^n \bar{u}_i}-1\right)\right] \text{ is } o(\epsilon^2). \quad \square$$

Now we prove Claim B.1.

Proof. By Hölder's inequality, for each $i \in \{1,\ldots,n\}$,

$$E\left[\left|\frac{e^{-\theta\epsilon\bar{q}_{\perp i}}-1}{-\theta\epsilon\bar{q}_{\perp i}}\right|^{\frac{p}{p-1}}|\bar{q}_{\perp i}|^{\frac{p}{p-1}}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}}\right] \le E\left[\left|\frac{e^{-\theta\epsilon\bar{q}_{\perp i}}-1}{-\theta\epsilon\bar{q}_{\perp i}}\right|^{\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}}\right]^{\frac{\tilde{p}-1}{\tilde{p}}} E\left[|\bar{q}_{\perp i}|^{\left(\frac{p}{p-1}\right)\tilde{p}}\right]^{\frac{1}{\tilde{p}}},$$

where $\tilde{p}>1$. On the one hand, we can choose p large so that $\frac{p}{p-1}\approx 1$ and $\tilde{p}>1$ such that $(\frac{p}{p-1})\tilde{p}=2$. Then $E[|\bar{q}_{\perp i}|^{(\frac{p}{p-1})\tilde{p}}]^{\frac{1}{\tilde{p}}}$ is $o(\epsilon^2)$ by SSC.

Also, by SSC, we know that $\epsilon|\bar{q}_{\perp i}|$ converges to zero in the mean-square sense and therefore in distribution. Then, by the continuous mapping theorem (Gut 2012, theorem 10.4 in section 5), we have that

$$\left(\frac{e^{-\theta\epsilon|\bar{q}_{\perp i}|}-1}{-\theta\epsilon|\bar{q}_{\perp i}|}\right)^{\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)} \Longrightarrow 1.$$

It remains to prove that $(e^{-\theta\epsilon|q_{\perp i}|}-1)/(-\theta\epsilon|q_{\perp i}|)$ is bounded to conclude that its expected value also converges to one. In fact, we have

$$-\theta\epsilon|\bar{q}_{\perp i}| \le |\theta|\epsilon|\bar{q}_{\perp i}| \le |\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|$$

and $|\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|\ge 0$. Then, by the facts 1 and 3 stated previously, we obtain

$$0 \le \frac{e^{-\theta\epsilon|\bar{q}_{\perp i}|}-1}{-\theta\epsilon|\bar{q}_{\perp i}|}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}} \le \frac{e^{|\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|}-1}{|\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}} \le e^{|\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}} \le e^{|\theta|\epsilon\|\bar{\boldsymbol{q}}_\perp\|}.$$

Therefore,

$$\begin{aligned} E\left[\left(\frac{e^{-\theta\epsilon|\bar{q}_{\perp i}|}-1}{-\theta\epsilon|\bar{q}_{\perp i}|}\right)^{\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)}\mathbb{1}_{\{\bar{q}_{\perp i}\neq 0\}}\right] &\le E\left[e^{|\theta|\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)\epsilon\|\bar{\boldsymbol{q}}_\perp\|}\right] \\ &\overset{(a)}{\le} E\left[e^{|\theta|\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)\epsilon\|\bar{\boldsymbol{q}}\|}\right] \\ &\overset{(b)}{\le} E\left[e^{|\theta|\left(\frac{p}{p-1}\right)\left(\frac{\tilde{p}}{\tilde{p}-1}\right)\epsilon\sum_{i=1}^n \bar{q}_i}\right] \\ &\overset{(c)}{<} \infty, \end{aligned}$$

where (a) holds because projection is nonexpansive, (b) holds because norm-1 is greater than the Euclidean norm, and (c) holds by assumption of Theorem 2 for $|\theta|(p/(p-1))(\tilde{p}/(\tilde{p}-1))\le\Theta$. Then the claim holds with

$$\theta_{\max}=\Theta\left(\frac{\tilde{p}-1}{2}\right), \tag{B.2}$$

where we used that $(\frac{p}{p-1})\tilde{p}=2$. $\square$

C. Details of the Proofs in Section 5

In this section, we provide the details of the proofs of the lemmas stated in Section 5 that we use in the proof of Theorem 3.

C.1. Proof of Lemma 7

To prove Lemma 7, we use the following lemma, which is similar to Lemma B.3.

Lemma C.1. *Consider a generalized switch parametrized by ϵ, as described in Theorem 3. Then, for any $\alpha\in\mathbb{R}$ and for all $k\ge 1$, we have*

$$\sum_{i=1}^n c_i^{(\ell)}u_i^{(\epsilon)}(k)e^{-\frac{\alpha}{c_i^{(\ell)}}\bar{q}_{\perp i}^{(\epsilon)}(k+1)} = \langle \boldsymbol{c}^{(\ell)},\boldsymbol{u}^{(\epsilon)}(k)\rangle e^{\alpha\langle \boldsymbol{c}^{(\ell)},\boldsymbol{q}^{(\epsilon)}(k+1)\rangle}.$$

Proof. First, observe that if $\alpha = 0$, the lemma trivially holds. Now we prove the lemma for $\alpha \neq 0$. From (3), we know that $q_i(k+1)u_i(k) = 0$ for all $i \in \{1, \dots, n\}$. Then, for all $\beta \in \mathbb{R}$, we have

$$u_i\left(e^{-\beta q_i(k+1)} - 1\right) = 0, \quad \forall i \in \{1, \dots, n\},$$

and this equation implies that

$$c_i^{(\ell)} u_i\left(e^{-\beta q_i(k+1)} - 1\right) = 0, \quad \forall i \in \{1, \dots, n\}.$$

Without loss of generality, we assume that $c_i^{(\ell)} > 0$ for all $i \in \{1, \dots, n\}$ because otherwise the last equation holds trivially. Let $\alpha \in \mathbb{R}$, and for each $i \in \{1, \dots, n\}$, let $\alpha_i \in \mathbb{R}$ be such that $\alpha = \alpha_i c_i^{(\ell)}$ for all $i \in \{1, \dots, n\}$. Then

$$c_i^{(\ell)} u_i\left(e^{-\alpha_i q_i(k+1)} - 1\right) = 0, \quad \forall i \in \{1, \dots, n\}.$$

Summing over all $i \in \{1, \dots, n\}$, we obtain

$$\begin{aligned}
0 &= \sum_{i=1}^n c_i^{(\ell)} u_i(k)\left(e^{-\alpha_i q_i(k+1)} - 1\right) \\
&= \sum_{i=1}^n c_i^{(\ell)} u_i(k)\left(e^{-\alpha_i q_{\|i}(k+1) - \alpha_i q_{\perp i}(k+1)} - 1\right) \\
&\overset{(a)}{=} \sum_{i=1}^n c_i^{(\ell)} u_i(k)\left(e^{-\alpha_i \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle c_i^{(\ell)} - \alpha_i q_{\perp i}(k+1)} - 1\right) \\
&\overset{(b)}{=} \sum_{i=1}^n c_i^{(\ell)} u_i(k)\left(e^{-\alpha \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle - \frac{\alpha}{c_i^{(\ell)}} q_{\perp i}(k+1)} - 1\right) \\
&\overset{(c)}{=} e^{-\alpha \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle} \sum_{i=1}^n c_i^{(\ell)} u_i(k) e^{-\frac{\alpha}{c_i^{(\ell)}} q_{\perp i}(k+1)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{u}(k)\rangle,
\end{aligned}$$

where (a) holds by definition of $\boldsymbol{q}_{\|}(k)$, (b) holds by definition of α, and (c) holds by expanding the product and reorganizing terms. Therefore, we have

$$\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{u}(k)\rangle = e^{-\alpha \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle} \sum_{i=1}^n c_i^{(\ell)} u_i(k) e^{-\frac{\alpha}{c_i^{(\ell)}} q_{\perp i}(k+1)}.$$

Multiplying both sides by $e^{\alpha \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle}$, we obtain

$$\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{u}(k)\rangle e^{\alpha \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}(k+1)\rangle} = \sum_{i=1}^n c_i^{(\ell)} u_i(k) e^{-\frac{\alpha}{c_i^{(\ell)}} q_{\perp i}(k+1)},$$

which proves the lemma. □

Now we prove Lemma 7.

Proof of Lemma 7. First, observe that if $\theta = 0$, the lemma holds trivially. Now assume that $\theta \neq 0$. Because $\boldsymbol{c}^{(\ell)} \geq 0$ and $\bar{u}_i \leq \bar{s}_i \leq S_{\max}$ for all $i \in \{1, \dots, n\}$, we have

$$0 \leq \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle \leq S_{\max}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{1}\rangle.$$

Then, from facts 1 and 2 stated in Section B.3 of Appendix B, we have

$$\left|e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle}\right| \leq \left|\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle\right| \left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{1}\rangle} - 1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{1}\rangle}\right). \tag{C.1}$$

Now, by properties of expected value, we have

$$\begin{aligned}
&\left|E\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle}-1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle}-1\right)\right]\right| \\
&\quad\le E\left[\left|e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle}-1\right|\left|e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle}-1\right|\right] \\
&\quad\overset{(a)}{\le}|\theta\epsilon|\left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}-1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}\right)E\left[\left|\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle}-1\right)\right|\right] \\
&\quad\overset{(b)}{=}|\theta\epsilon|\left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}-1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}\right)E\left[\left|\sum_{i=1}^{n}c_i^{(\ell)}\bar{u}_i\left(e^{-\left(\frac{\theta\epsilon}{c_i^{(\ell)}}\right)\bar{q}^{+}_{\perp i}}\right)\right|\right] \\
&\quad\overset{(c)}{\le}|\theta\epsilon|\left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}-1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}\right)E\left[\sum_{i=1}^{n}c_i\bar{u}_i\left|e^{-\left(\frac{\theta\epsilon}{c_i^{(\ell)}}\right)\bar{q}^{+}_{\perp i}}-1\right|\right] \\
&\quad\overset{(d)}{\le}|\theta\epsilon|\left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}-1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}\right)E\left[\sum_{i=1}^{n}\left(c_i^{(\ell)}\bar{u}_i\right)^{p}\right]^{\frac{1}{p}}E\left[\sum_{i=1}^{n}\left|e^{-\left(\frac{\theta\epsilon}{c_i^{(\ell)}}\right)\bar{q}^{+}_{\perp i}}-1\right|^{\frac{p}{p-1}}\right]^{\frac{p-1}{p}},
\end{aligned}$$

where $p>1$. Here (a) holds by (C.1); (b) holds by Lemma C.1 with $\alpha=\theta\epsilon$; (c) holds by triangle inequality; and (d) holds by Hölder's inequality.

However,

$$E\left[\sum_{i=1}^{n}\left(c_i^{(\ell)}\bar{u}_i\right)^{p}\right]\le (c_{\max}S_{\max})^{p-1}E\left[\sum_{i=1}^{n}c_i^{(\ell)}\bar{u}_i\right]\le (c_{\max}S_{\max})^{p-1}\epsilon,$$

where $c_{\max}=\max_i c_i^{(\ell)}$, and the last equality holds for the following reason. By Lemma 9, we have

$$E\left[\langle \boldsymbol{c}^{(\ell)},\boldsymbol{u}\rangle\right]=\epsilon-\boldsymbol{b}^{(\ell)}+E\left[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right].$$

Also, by definition of the capacity region in (19) and because $\bar{\boldsymbol{s}}$ depends on the channel state, we have that $E[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle]\in\mathscr{C}$. Then

$$-\boldsymbol{b}^{(\ell)}+E\left[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right]\le 0. \tag{C.2}$$

Therefore,

$$\begin{aligned}
&\left|E\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle}-1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle}-1\right)\right]\right| \\
&\quad\le|\theta|\epsilon^{1+\frac{1}{p}}(c_{\max}S_{\max})^{\frac{p-1}{p}}\left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}-1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)},\mathbf{1}\rangle}\right)E\left[\sum_{i=1}^{n}\left|e^{-\left(\frac{\theta\epsilon}{c_i^{(\ell)}}\right)\bar{q}^{+}_{\perp i}}-1\right|^{\frac{p}{p-1}}\right]^{\frac{p-1}{p}}.
\end{aligned}$$

The rest of the argument is similar to the last steps in the proof of Lemma 4. However, in this case, we do not need to use the existence of the MGF of $\epsilon\sum_{i=1}^{n}\bar{q}_i$ because we know that $E[e^{\theta\epsilon\|q_{\perp}\|}]$ is bounded for $\theta\epsilon\le\theta^*$ from SSC. □

C.2. Existence of MGF of $\epsilon\|\bar{\boldsymbol{q}}\|$ in the Generalized Switch

We prove the following lemma.

Lemma C.2. *Consider a generalized switch parametrized by ϵ, as described in Theorem 3. Then, for each $\epsilon>0$, there exists $\Theta>0$ such that $E[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}]<\infty$ for all $\theta\in[-\Theta,\Theta]$.*

Proof. First, observe that if $\theta=0$, the lemma holds trivially. Therefore, in this proof, we assume that $\theta\ne0$. We use the Foster–Lyapunov theorem (Hajek 2015, proposition 6.13) with Lyapunov function $V(\boldsymbol{q})=e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\boldsymbol{q}\rangle}$. In this proof, we use the notation

$$E_{\boldsymbol{q}}[\cdot]\triangleq E\left[\cdot|\overline{\boldsymbol{q}}=\boldsymbol{q}\right]\quad\text{and}\quad E_t[\cdot]\triangleq E[\cdot|T(k)=t]$$

The drift of $V(\bar{\boldsymbol{q}})$ conditioned on $\bar{\boldsymbol{q}} = \boldsymbol{q}$ is

$$
\begin{aligned}
&E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle} - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\right] \\
&\overset{(a)}{=} E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}-\bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle} + 1 - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle} + o\left(\epsilon^2\right)\right] \\
&= E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}\rangle} e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle} + 1 - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle} + o\left(\epsilon^2\right)\right] \\
&\overset{(b)}{=} E_{\boldsymbol{q}}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}\rangle - \bar{B}\right)} - e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle} + 1 - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle} + o\left(\epsilon^2\right)\right] \\
&\quad + E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}-\bar{\boldsymbol{s}}\rangle} - e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}\rangle - \bar{B}\right)}\right] \\
&\overset{(c)}{=} E_{\boldsymbol{q}}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}\rangle - \bar{B}\right)} - e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle} + 1 - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle} + o\left(\epsilon^2\right)\right] \\
&\quad + E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle}\right] E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\left(e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\bar{B}}\right),\right]
\end{aligned}
$$

where (a) holds by expanding the product and rearranging terms in Lemma 7, (b) holds after adding and subtracting $E_{\boldsymbol{q}}[e^{\theta\epsilon(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}+\bar{\boldsymbol{a}}\rangle - \bar{B})}]$ and reorganizing terms, and (c) holds because the arrival process is independent of the queue length and services processes.

However,

$$
\begin{aligned}
E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\left(e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\bar{B}}\right)\right] &= E_{\boldsymbol{q}}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle - \bar{B}\right)}\left(e^{\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)} - 1\right)\right] \\
&= E_{\boldsymbol{q}}\left[E_t\left[e^{-\theta\epsilon\bar{B}}\right] E_t\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\left(e^{\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)} - 1\right)\right]\right] \\
&= E_{\boldsymbol{q}}\left[E_t\left[e^{-\theta\epsilon\bar{B}}\right] E_t\left[e^{\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)} - 1\right]\right] + o\left(\epsilon^2\right),
\end{aligned}
$$

where the last equality holds by Lemma 7 and because the random variable $\bar{B}$ is bounded (because it takes finitely many values).

Rearranging terms, we obtain

$$
E_{\boldsymbol{q}}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}^{+}\rangle} - e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{q}}\rangle}\right] = 1 - E_{\boldsymbol{q}}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle}\right] + E\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle}\right] E_{\boldsymbol{q}}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle} - e^{\theta\epsilon\bar{B}}\right] + o\left(\epsilon^2\right), \tag{C.3}
$$

$$
+ e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\boldsymbol{q}\rangle} E\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle - \bar{B}\right)} - 1\right]. \tag{C.4}
$$

Observe that the right-hand side of (C.3) is bounded because $\bar{u}_i \le \bar{s}_i \le S_{\max}$ and $\bar{a}_i \le A_{\max}$ with probability one for all $i \in \{1, \ldots, n\}$. Also, $\bar{B}$ is bounded because it takes a finite number of values.

Then it suffices to show that for $\delta > 0$, there exists $\Theta > 0$ such that

$$
E\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle - \bar{B}\right)} - 1\right] < -\delta, \qquad \forall \theta \in [-\Theta, 0) \cup (0, \Theta].
$$

This result can be easily using the continuity of MGF at $\theta = 0$ and the Taylor expansion of $E[e^{\theta\epsilon(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle - \bar{B})}]$ with respect to θ up to first order around $\theta = 0$. We omit the details for brevity. □

C.3. Proof of Lemma 8

First, observe that if $\theta = 0$, the proof holds trivially. Now assume that $\theta \neq 0$.

In this proof, we use a geometric vision of the maxweight algorithm. Before presenting the technical details, we present an intuitive overview of the proof. Recall that given the channel state, the maxweight algorithm maximizes $\langle \boldsymbol{q}(k), \boldsymbol{x}\rangle$ over the set of feasible service rate vectors. Then maxweight solves an optimization problem with linear objective function. Equivalently, maxweight finds a vector $\boldsymbol{x}^*$ that is an optimal solution of

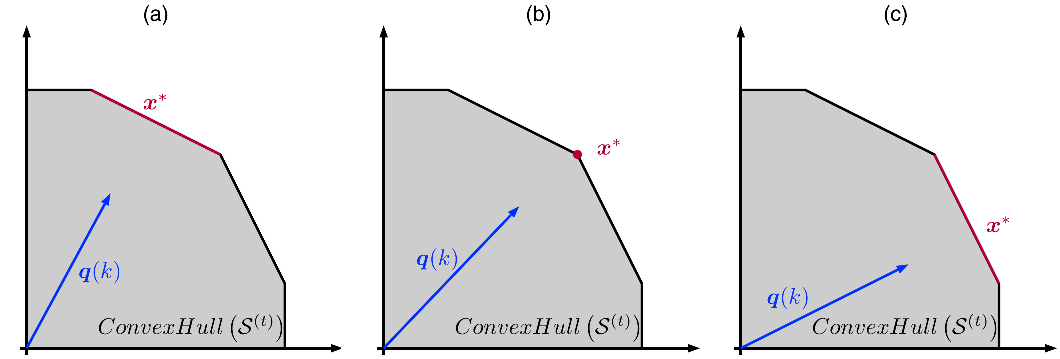
$$
\begin{aligned}
\max \quad & \left\langle \boldsymbol{q}(k), \boldsymbol{x}\right\rangle \\
\text{subject to} \quad & \boldsymbol{x} \in \mathit{ConvexHull}\left(\mathcal{S}^{(t)}\right)
\end{aligned} \tag{C.5}
$$

and sets $\boldsymbol{s}(k)$ as one of these optimal solutions. To make the optimization problem linear, we use $\mathit{ConvexHull}(\mathcal{S}^{(t)})$ as the feasible region instead of $\mathcal{S}^{(t)}$. However, this does not change the problem because the objective function is linear, and therefore, an optimal solution of (C.5) is at an extreme point, that is, at a point in $\mathcal{S}^{(t)}$.

The gradient of the objective function is $\boldsymbol{q}(k)$. Then, depending on its direction, the optimal solution(s) $\boldsymbol{x}^*$ will belong to a different facet or vertex of $\mathit{ConvexHull}(\mathcal{S}^{(t)})$. In Figure C.1, we present pictorial examples in which we show the optimal solution(s) when the vector of queue lengths goes in three different directions.

Recall that $\boldsymbol{q}_{\|}(k)$ goes in the same direction as $\boldsymbol{c}^{(\ell)}$. Also, if ϵ is small, we expect that $\boldsymbol{q}(k) \approx \boldsymbol{q}_{\|}(k)$ by SSC. Then, if ϵ is small, we expect that any optimal solution $\boldsymbol{x}^*$ to the linear program (C.5) satisfies $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}^*\rangle = b^{(t,\ell)}$ with high probability.

Figure C.1. Example of Optimal Solutions Depending on the Queue Lengths Vector.



Example 1: Multiple solutions, since $\boldsymbol{q}(k)$ is perpendicular to the second facet from left to right.

Example 2: Unique solution.

Example 3: Another example of multiple solutions.

Now we present the technical details. We start with a definition. Let $t \in \mathcal{T}$, and suppose that the channel state is $\bar{T} = t$. Then let $\nu^{(t)} \in (0, \pi/2]$ be an angle such that $\langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}} \rangle = b^{(t,\ell)}$ if $\|\boldsymbol{q}_{\|}\|/\|\boldsymbol{q}\| \geq \cos(\nu^{(t)})$. Let $\nu_{\bar{\boldsymbol{q}}}$ be the angle between $\bar{\boldsymbol{q}}_{\|}$ and $\bar{\boldsymbol{q}}$, and define $\nu_{\min} \triangleq \min_{t\in\mathcal{T}} \nu^{(t)}$. Therefore, because $\bar{\mathbf{s}}$ is scheduled using the maxweight algorithm, if the channel state is t, we have

$$b^{(t,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}} \rangle \quad \text{implies} \quad \nu^{(t)} < \nu_{\bar{\boldsymbol{q}}}. \tag{C.6}$$

In this proof, we use the notation $E_t[\cdot] = E[\,\cdot\,|T = t]$. By definition of conditional expectations, we have

$$E\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}^{(\epsilon)}\rangle\right)} - 1\right)\right] = \sum_{t\in\mathcal{T}} \psi_t E_t\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}^{(\epsilon)}\rangle\right)} - 1\right),$$

where

$$\begin{aligned}
&E_t\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}^{(\epsilon)}\rangle\right)} - 1\right)\right] \\
&\quad \overset{(a)}{=} E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\|}\|} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)\mathbb{1}_{\{b^{(t,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\}}\right] \\
&\quad \overset{(b)}{\leq} E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\|}\|} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)\mathbb{1}_{\{\nu_{\bar{\boldsymbol{q}}} > \nu^{(t)}\}}\right] \\
&\quad = E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\perp}\|\cot\left(\nu_{\bar{\boldsymbol{q}}}\right)} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)\mathbb{1}_{\{\nu_{\bar{\boldsymbol{q}}} > \nu^{(t)}\}}\right] \\
&\quad \overset{(c)}{\leq} E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\perp}\|\cot\left(\nu^{(t)}\right)} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)\mathbb{1}_{\{\nu_{\bar{\boldsymbol{q}}} > \nu^{(t)}\}}\right] \\
&\quad \overset{(d)}{=} E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\perp}\|\cot\left(\nu^{(t)}\right)} - 1\right)\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)\right] \\
&\quad \overset{(e)}{\leq} E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\perp}\|\cot\left(\nu^{(t)}\right)} - 1\right)^{p}\right]^{\frac{1}{p}} E_t\left[\left(e^{\theta\epsilon\left(b^{(t,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \bar{\mathbf{s}}\rangle\right)} - 1\right)^{\frac{p}{p-1}}\right]^{\frac{p-1}{p}},
\end{aligned}$$

where $p > 1$. Here (a) holds by definition of the indicator function and because $\bar{\boldsymbol{q}}_{\|} = \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{q}} \rangle \boldsymbol{c}^{(\ell)}$ by the definition of projection, (b) holds by (C.6), (c) and (d) hold because $\cot(\nu)$ is decreasing for $\nu \in (0, \pi/2]$, (d) holds by (C.6) and by definition of the indicator function, and (e) holds by Hölder's inequality.

Using an argument similar to the one at the end of Lemma 4, it can be proved that

$$0 \leq E_t\left[\left(e^{\theta\epsilon\|\bar{\boldsymbol{q}}_{\perp}\|\cot\left(\nu^{(t)}\right)} - 1\right)^{p}\right]^{\frac{1}{p}} \text{ is } O(\epsilon)$$

converges to a constant as $\epsilon \downarrow 0$. By contrast,

$$
\begin{aligned}
&E_t\left[\left(e^{\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)}-1\right)^{\frac{p}{p-1}}\right]\\
&\quad= E_t\left[\left(e^{\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)}-1\right)^{\frac{p}{p-1}}\mathbb{1}_{\{b^{(t,\ell)}\neq\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\}}\right]\\
&\quad= E\left[\left(\frac{e^{\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)}-1}{\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)}\right)^{\frac{p}{p-1}}\left(\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)\right)^{\frac{p}{p-1}}\mathbb{1}_{\{b^{(t,\ell)}\neq\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\}}\right]\\
&\quad\leq \left(\frac{e^{\theta\epsilon\left(\bar{B}_{\max}-\langle \boldsymbol{c}^{(\ell)},S_{\max}\mathbf{1}\rangle\right)}-1}{\theta\epsilon\left(\bar{B}_{\max}-\langle \boldsymbol{c}^{(\ell)},S_{\max}\mathbf{1}\rangle\right)}\right)^{\frac{p}{p-1}}\left(\theta\epsilon\left(\bar{B}_{\max}-\langle \boldsymbol{c}^{(\ell)},S_{\max}\mathbf{1}\rangle\right)\right)^{\frac{p}{p-1}}P\left[b^{(t,\ell)}\neq\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right],
\end{aligned}
$$

where $\bar{B}_{\max} = \max_{t\in\mathcal{T}} b^{(t,\ell)}$. Eryilmaz and Srikant (2012) prove that $P\left[b^{(t,\ell)}\neq\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right] = K\epsilon$ for a finite constant K, and their proof also holds here. Therefore,

$$
E\left[\left(e^{\theta\epsilon\left(b^{(t,\ell)}-\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle\right)}-1\right)^{\frac{p}{p-1}}\right] \text{ is } O\left(\epsilon^{1+\frac{p}{p-1}}\right). \quad \square
$$

D. Proof of the Claims in Sections 4.2 and 5.2

In this appendix, we show the proof of all the claims that we made in the proofs of our theorems.

D.1. Proof of Claim 1

We have

$$
\begin{aligned}
0 \leq \frac{(\theta\epsilon)^2}{2}E\left[\left(\sum_{i=1}^n \bar{u}_i\right)^2\right] &\overset{(a)}{\leq} \epsilon^2\left(\frac{nS_{\max}\theta^2}{2}\right)E\left[\sum_{i=1}^n \bar{u}_i\right],\\
&\overset{(b)}{=} \epsilon^3\left(\frac{nS_{\max}\theta^2}{2}\right)
\end{aligned}
$$

where (a) holds because, by definition of unused service, we have $\bar{u}_i \leq \bar{s}_i \leq S_{\max}$ and all terms are nonnegative, and (b) holds by Lemma 5.

Therefore,

$$
\frac{(\theta\epsilon)^2}{2}E\left[\left(\sum_{i=1}^n \bar{u}_i\right)^2\right] \text{ is } O\left(\epsilon^3\right). \quad \square
$$

D.2. Proof of Claim 2

We have

$$
\begin{aligned}
0 \leq \frac{(\theta\epsilon)^2}{2}E\left[\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle\right)^2\right] &\overset{(a)}{\leq} \epsilon^2\left(\frac{\langle \boldsymbol{c}^{(\ell)},S_{\max}\mathbf{1}\rangle\theta^2}{2}\right)E\left[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle\right]\\
&\overset{(b)}{\leq} \epsilon^3\left(\frac{\langle \boldsymbol{c}^{(\ell)},S_{\max}\mathbf{1}\rangle\theta^2}{2}\right),
\end{aligned}
$$

where (a) holds because $\bar{u}_i \leq \bar{s}_i \leq S_{\max}$ and $\boldsymbol{c}^{(\ell)} \geq \boldsymbol{0}$, and (b) holds by Lemma 9 because $E[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{s}}\rangle - \bar{B}] \leq 0$.

Therefore,

$$
\frac{(\theta\epsilon)^2}{2}E\left[\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{u}}\rangle\right)^2\right] \text{ is } O\left(\epsilon^3\right). \quad \square
$$

D.3. Proof of Claim 3

For the first expression, from Lemma 3, we have

$$
\begin{aligned}
E\left[f_{\epsilon,\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle-\bar{B}\right)}(\theta)\right] &= 1+\theta\epsilon E\left[\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle-\bar{B}\right]+\frac{(\theta\epsilon)^2}{2}E\left[\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle-\bar{B}\right)^2\right]+O\left(\epsilon^3\right)\\
&= 1+\theta\epsilon^2+\frac{(\theta\epsilon)^2}{2}E\left[\left(\langle \boldsymbol{c}^{(\ell)},\bar{\boldsymbol{a}}\rangle-\bar{B}\right)^2\right]+O\left(\epsilon^3\right),
\end{aligned}
$$

where the last equality holds by (25). Also,

$$0 \le \frac{(\theta\epsilon)^2}{2} E\left[\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)^2\right] \overset{(a)}{\le} \epsilon^2 \left(\frac{\left(\langle \boldsymbol{c}^{(\ell)}, A_{\max}\mathbf{1}\rangle + B_{\max}\right)\theta^2}{2}\right) E\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right]$$
$$\overset{(b)}{=} \epsilon^3 \left(\frac{\left(\langle \boldsymbol{c}^{(\ell)}, A_{\max}\mathbf{1}\rangle + B_{\max}\right)\theta^2}{2}\right),$$

where (a) holds because $\bar{a}_i \le A_{\max}$ with probability one for all $i \in \{1, \ldots, n\}$, $\boldsymbol{c}^{(\ell)} \ge \boldsymbol{0}$, $\bar{B}$ is bounded by a constant that we denote $B_{\max}$ and because all quantities are nonnegative, and (b) holds by (25). Then

$$\frac{(\theta\epsilon)^2}{2} E\left[\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)^2\right] \text{ is } O(\epsilon^3).$$

Therefore,

$$E\left[f_{\epsilon,\left(\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{a}}\rangle - \bar{B}\right)}(\theta)\right] = 1 + \theta\epsilon^2 + O(\epsilon^3).$$

This proves the first equation of the claim.

For the second expression, using Lemma 3, we obtain

$$E\left[f_{\epsilon,\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)}(\theta)\right] - 1 = \theta\epsilon E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right] + \frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)^2\right] + O(\epsilon^3).$$

However,

$$0 \le \frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)^2\right] \overset{(a)}{\le} \epsilon^2 \left(\frac{\left(B_{\max} + \langle \boldsymbol{c}^{(\ell)}, S_{\max}\mathbf{1}\rangle\right)\theta^2}{2}\right) E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right]$$
$$\overset{(b)}{\le} \epsilon^3 \left(\frac{\left(B_{\max} + \langle \boldsymbol{c}^{(\ell)}, S_{\max}\mathbf{1}\rangle\right)\theta^2}{2}\right),$$

where (a) holds because $\bar{s}_i \le S_{\max}$ with probability one for all $i \in \{1, \ldots, n\}$, $\boldsymbol{c}^{(\ell)} \ge \boldsymbol{0}$, and $\bar{B} \le B_{\max}$, and all quantities are nonnegative (see (C.2) to see why $E[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle] \ge 0$), and (b) holds by Lemma 9 and because $E[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{u}}\rangle] \ge 0$, given that $\bar{\boldsymbol{u}} \ge \boldsymbol{0}$ and $\boldsymbol{c}^{(\ell)} \ge \boldsymbol{0}$.

Then

$$\frac{(\theta\epsilon)^2}{2} E\left[\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)^2\right] \text{ is } O(\epsilon^3).$$

Therefore,

$$E\left[f_{\epsilon,\left(\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right)}(\theta)\right] - 1 = \theta\epsilon E\left[\bar{B} - \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle\right] + O(\epsilon^3). \quad \square$$

Endnote

[1] In fact, the exponential moment bound is not part of the SSC statement of Eryilmaz and Srikant (2012), but their proof of their proposition 2 implies it.

References

Bertsimas D, Gamarnik D, Tsitsiklis JN (2001) Performance of multiclass markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Probability* 11(4):1384–1428.

Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Math. Oper. Res.* 45(3):1069–1103.

Braverman A, Dai J (2017) Stein's method for steady-state diffusion approximations of M/Ph/n+M systems. *Ann. Appl. Probability* 27(1):550–581.

Braverman A, Dai J, Feng J (2017a) Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stochastic Systems* 6(2):301–366.

Braverman A, Dai J, Miyazawa M (2017b) Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stochastic Systems* 7(1):143–196.

Braverman A, Gurvich I, Huang J (2020) On the Taylor expansion of value functions. *Oper. Res.* 68(2):631–654.

Chen H, Ye H (2012) Asymptotic optimality of balanced routing. *Oper. Res.* 60(1):163–179.

Dai J (2018) Steady-state approximations: Achievement lecture. *ACM SIGMETRICS Performance Evaluation Rev.* 46(1):1.

Dai J, Lin W (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probability* 18(6):2239–2299.

Ephremides A, Varaiya P, Walrand J (1980) A simple dynamic routing problem. *IEEE Trans. Automatic Control* 25(4):690–693.

Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3–4):311–359.

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

Foschini G, Salz J (1978) A basic dynamic routing problem and diffusion. *IEEE Trans. Communications* 26(3):320–327.

Gamarnik D, Zeevi A (2006) Validity of heavy traffic steady-state approximations in Generalized Jackson Networks. *Ann. Appl. Probability* 16(1):56–90.

Gaunt R, Walton N (2020) Stein's method for the single server queue in heavy traffic. *Statist. Probability Lett.* 156:108566.

Gupta G, Shroff N (2010) Delay analysis for wireless networks with single hop traffic and general interference constraints. *IEEE/ACM Trans. Networks* 18(2):393–405.

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probability* 24(6):2527–2559.

Gurvich I, Huang J, Mandelbaum A (2013) Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Oper. Res.* 39(2):325–373.

Gut A (2012) *Probability: A Graduate Course*, vol. 75 (Springer Science & Business Media, New York).

Hajek B (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Probability* 14(3):502–525.

Hajek B (2015) *Random Processes for Engineers* (Cambridge University Press, Cambridge, UK).

Harrison J (1988) Brownian models of queueing networks with heterogeneous customer populations Fleming W, Lions PL, eds. *Stochastic Differential Systems, Stochastic Control Theory and Applications* (Springer, New York), 147–186.

Harrison J (1998) Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Ann. Appl. Probability* 8(3):822–848.

Harrison J, López M (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.* 33(4):339–368.

Harrison P, Patel N (1992) *Performance Modeling of Communication Networks and Computer Architectures* (Addison Wesley Longman, Upper Saddle River, NJ).

Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Oper. Res.* 66(4):1168–1188.

Hurtado-Lange D, Maguluri ST (2019) Heavy-traffic analysis of queueing systems with no complete resource pooling. *ACM SIGMETRICS Performance Evaluation Rev.* 46(2):81–82.

Kang W, Kelly F, Lee N, Williams R (2009) State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probability* 19(5):1719–1780.

Kingman J (1961) The single server queue in heavy traffic. *Math. Proc. Cambridge Philosophy Soc.* 57(4):902–904.

Kingman J (1962a) On queues in heavy traffic. *J. Roy. Statist. Soc. B* 24(2):383–392.

Kingman J (1962b) Some inequalities for the queue GI/G/1. *Biometrika* 49(3):315–324.

Köllerström J (1974) Heavy traffic theory for queues with several servers. i. *J. Appl. Probability* 11(3):544–552.

Lehoczky J (1996) Real-time queueing theory. *17th IEEE Real-Time Systems Symposium* (Institute of Electrical and Electronics Engineers, Washington, DC), 186–195.

Lehoczky J (1997) Using real-time queueing theory to control lateness in real-time systems. *Performance Evaluation Rev.* 25(1):158–168.

Li B, Kong X, Wang L (2018) Optimal load-balancing for high-density wireless networks with flow-level dynamics. *Proc. 18th ACM Internat. Sympos. Mobile Ad Hoc Networking Comput.* (ACM, New York), 316–317.

Lindley D (1952) The theory of queues with a single server. *Math. Proc. Cambridge Philosophical Soc.* 48(2):277–289.

Liu X, Ying L (2019) A simple steady-state analysis of load balancing algorithms in the sub-Halfin-Whitt regime. *Performance Evaluation Rev.* 46(2):15–17.

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

Lukacs E (1970) *Characteristic Functions* (Griffin, London).

Maguluri ST, Srikant R (2016) Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stochastic Systems* 6(1):211–250.

Maguluri ST, Burle S, Srikant R (2018) Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems* 88(3–4):279–309.

Maguluri ST, Srikant R, Ying L (2014) Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation* 81:20–39.

Marshall K (1968) Some inequalities in queuing. *Oper. Res.* 16(3):651–668.

Massoulié L, Roberts J (2000) Bandwidth sharing and admission control for elastic traffic. *Telecomm. Systems* 15(1–2):185–201.

McKeown N, Anantharam V, Walrand J (1996) Achieving 100% throughput in an input queued switch. *Proc. IEEE INFOCOM* 47(8):296–302.

Meyn S (2008) Stability and asymptotic optimality of generalized MaxWeight policies. *SIAM J. Control Optim.* 47(6):3259–3294.

Mitzenmacher M (1996) *Load Balancing and Density Dependent Jump Markov Processes* (IEEE, Piscataway, NJ).

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

Mood A (1950) *Introduction to the Theory of Statistics* (McGraw-Hill, New York).

Skorokhod A (1961) Stochastic equations for diffusion processes in a bounded region. *Theory Probability Appl.* 6(3):264–274.

Srikant R, Ying L (2014) *Communication Networks: An Optimization, Control and Stochastic Networks Perspective* (Cambridge University Press, Cambridge, UK).

Stolyar A (2004) Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probability* 14(1):1–53.

Stolyar A (2017) Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers. *Queueing Systems* 85(1–2):31–65.

Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automation Control* 37(12):1936–1948.

van der Boor M, Borst S, van Leeuwaarden J, Mukherjee D (2018) Scalable load balancing in networked systems: A survey of recent advances. Preprint, submitted June 14, https://arxiv.org/abs/1806.05444.

Vlasiou M, Zhang J, Zwart B (2014) Insensitivity of proportional fairness in critically loaded bandwidth sharing networks. Preprint, submitted June 18, https://arxiv.org/abs/1411.4841.

Vvedenskaya N, Dobrushin R, Karpelevich F (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems Inform. Transmission* 32(1):15–27.

Wang CH, Maguluri ST, Javidi T (2017) Heavy traffic queue length behavior in switches with reconfiguration delay. *INFOCOM 2017: IEEE Conf. Computer Commun.* (IEEE, New York), 1–9.

Wang W, Maguluri ST, Srikant R, Ying L (2018) Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *SIGMETRICS Performance Evaluation Rev.* 45(3):232–245.

Weber R (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Probability* 15(2):406–413.

Williams R (1998) Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* 30:27–88.

Williams R (2000) On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Commun.* 28(5.1):49–71.

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

Ye HQ, Yao D (2012) A stochastic network under proportional fair resource control—diffusion limit with multiple bottlenecks. *Oper. Res.* 60(3):716–738.

Ying L (2016) On the approximation error of mean-field models. *Proc. 2016 ACM SIGMETRICS Internat. Conf. Measurement Model. Comput. Sci.* (ACM, New York), 285–297.

Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Measurement Analysis Comput. Syst.* 1(1):12:1–12:27.

Ying L, Srikant R, Kang X (2017) The power of slightly more than one sample in randomized load balancing. *Math. Oper. Res.* 42(3):692–722.

Zhou X, Tan J, Shroff N (2018) Flexible load balancing with multi-dimensional state-space collapse: Throughput and heavy-traffic delay optimality. *Performance Evaluation* 127:176–193.