Transform Methods for Heavy-Traffic Analysis

Daniela Hurtado-Lange*  Siva Theja Maguluri
d.hurtado@gatech.edu siva.theja@gatech.edu

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

July 10, 2019

Abstract

The drift method was recently developed to study queueing systems in steady-state. It was successfully used to obtain bounds on the moments of the scaled queue lengths, that are asymptotically tight in heavy-traffic, in a wide variety of systems including generalized switches [11], input-queued switches [42, 41], bandwidth sharing networks [62], etc. 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 MGF method to obtain the stationary distribution of queue lengths in heavy-traffic in queueing systems that satisfy the Complete Resource Pooling condition. In particular, we study load balancing systems and generalized switches under general settings.

1 Introduction

Exact analysis of queueing systems that arise in study of Stochastic Processing Networks is usually intractable, so it is common to analyze them in various asymptotic regimes to get insights on system behaviour. 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 main advantages of the heavy-traffic limit is that many queueing systems in this limit behave as if they live in a much lower dimensional subspace of the state space. 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. Typically when the CRP condition is satisfied, the state space collapses into a one-dimensional subspace and the system behaves as if there is a single bottleneck, i.e. like a single server queue [24, 9].

Over the past decades, several queueing 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 [32], 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 [25, 22, 64, 23, 54, 14]. 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 space. Since the queue lengths cannot be negative, they ‘reflect’ at the origin, so this lower dimensional Brownian Motion process is called a Reflected Brownian Motion 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 same as the heavy-traffic limiting stationary distribution of the

*This work was partially funded by the CONICYT PFCHA/DOCTORADO BECAS CHILE/2018 - 72190413
original (unscaled) queueing system. However, this must be formally established by proving that the limit to the steady-state and the limit to the heavy-traffic (equivalently, 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. Even though this method has been successfully used to study a wide variety of problems that satisfy the CRP condition, it is challenging to study systems when the CRP condition is not satisfied.

In addition, three different ‘direct methods’ were developed to study the queueing systems in heavy-traffic without considering the scaled process and the diffusion limits [8]. Therefore, none of these direct methods require the interchange of limits step. They are the drift method [11, 42, 41, 62, 70], the Stein’s method [17, 5, 3] and the BAR method [6]. We briefly describe each of these methods below.

The main idea in the drift method is to carefully choose a test function, and set its drift to zero in steady-state, i.e. to equate the expected value of the test function in steady-state to the same value at the following time step. Since 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 [20] and [1]. 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 $m^{th}$ 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 [19] 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 queueing systems that satisfy the CRP condition [11], and include a load balancing system and an ad hoc wireless network in presence of interference and fading (time-varying channel conditions). The drift method was later successfully applied 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 [42, 41] and the bandwidth sharing network [62]. 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 [28]. Therefore, a new method is needed.

In this paper we develop the 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 main idea in the MGF method is to use an exponential test function, and set its drift to zero in steady-state. 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 the 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 $\theta$, 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 that 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 yields the stationary distribution in one step.

While the drift method is based on setting 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 [6], 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 the current 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, while 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. On the other hand, the Stein’s method is focused on computing rates of convergence to the limiting distribution. Stein’s method for studying queueing systems was first introduced by [17]. Erlang-A and Erlang-C queueing models were studied using Stein’s method by [5], and [3]. 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 mean field regime [68, 69], in Halfin-Whitt regime in [2], and in sub-Halfin-Whitt regimes in [38]. Universal approximations for queues with abandonment were obtained using Stein’s method in [27]. More recently, a single server queue in heavy-traffic was studied using Stein’s method in [15]. [18] 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 queueing systems that satisfy the CRP condition. In order to motivate 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 heavy-traffic scaled queue length in heavy-traffic limit converges to an exponential distribution. This is of course a classic result first proved by [32] using the diffusion limit method, and later by [11] 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 queueing systems under some general conditions and we exemplify with specific systems and resource allocation policies that satisfy those conditions. In Section 4 we study load balancing systems and identify that the Join the Shortest Queue (JSQ) [13, 65] and power-of-two choices [59, 50, 51] routing policies satisfy the assumptions. In Section 5 we study generalized switches [54] under the CRP condition and note that MaxWeight scheduling algorithm [56] satisfies our assumptions. We also show that ad hoc wireless networks operating under MaxWeight scheduling algorithm satisfy our assumptions. All these systems exhibit a one-dimensional SSC due to the CRP condition, and we show that the stationary distribution of this one-dimensional component is exponential. In addition to Moment Generating Functions, 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 also present a brief discussion about the pros and cons of different approaches.

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 above 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, since the test function is essentially the MGF. Even though most of the results that we present have already been established in the literature using diffusion limit and drift methods, 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, we show that even under correlated arrivals or services, the heavy-traffic scaled stationary distribution continues to be exponential (Theorem 4.2.3 and 5.2.3, 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 is on queueing 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 [42, 41]. 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, where it was first proposed by [11] to prove known results in systems under the CRP condition. The drift method was later generalized to study the input-queued switch when CRP condition is not satisfied [42, 41].
1.1 Notation

In this section we introduce the notation that we will use along the 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 $X$.

We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{R}^n$ to denote the set $n$-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 $i \in \{1, \ldots, n\}$. We use $\mathbf{1}$ to denote a vector of ones, i.e., if $\mathbf{x} = \mathbf{1}$, then $x_i = 1$ for all $i \in \{1, \ldots, n\}$. The dot product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$ and the Euclidian norm is denoted by $||\mathbf{x}||$, i.e., $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

The indicator function of $A$ is $\mathbb{1}_{(A)}$, i.e., $\mathbb{1}_{(A)}$ is one if $A$ is true and 0 otherwise. Convergence in distribution is denoted by $\Rightarrow$.

We say $f(x)$ is $O(g(x))$ if $\lim_{x \to 0} \left| \frac{f(x)}{g(x)} \right|$ is finite and we say that $f(x)$ is $o(g(x))$ if $\lim_{x \to 0} \frac{f(x)}{g(x)} = 0$.

2 Related Work

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

Moment Generating Functions have been used in the literature to study queueing systems such as the classical analysis of $M/G/1$ queue [26]. However, here we use the MGF to study heavy-traffic scaled queue lengths, since 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 [33] and [31] to study heavy-traffic queueing systems, and moment generating functions were used by [34, 35]. 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 [31] using Characteristic Functions and tools from complex analysis. [33] also used Characteristic Functions to study single server queue. The diffusion limit method to study queueing systems was developed by studying the single server queue [32]. The well known Kingman bound for the expected waiting time in a single server queue was developed in the 60’s [30], and later [46] used similar arguments to compute bounds on the second moment. These formed the basis for the drift method that was developed by [11]. The single server queue was also presented as an illustrative example of the BAR method [6]. Most of these papers study the delay in $G/G/1$ queue in continuous time which evolves according to the Lindley’s equation [37]. Similar to [11], in this paper we study the queue length in a discrete time. The queue lengths process evolves according to equation (3), which is equivalent to Lindley’s equation for the waiting time of $k^{th}$ 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 70’s. It was shown that the JSQ policy minimizes the mean delay among the class of policies that do not know the job durations [10, 63, 65]. Heavy-traffic optimality of the JSQ policy in a system with two servers was established by [13] using the diffusion limit method, where they also introduced the notion of SSC. Since then, the load balancing system has been extensively studied both to improve performance and to decrease the complexity of the algorithms [7, 36, 2, 70, 68, 69, 12, 39, 55, 67]. One lower complexity algorithm that has received attention is the power-of-two choices algorithm [59, 50, 51, 7]. An exhaustive survey of literature on load balancing is presented by [57]. The most relevant work for our purposes is the study of the JSQ policy under the drift method by [11] and that of the power-of-two choices algorithm by [43].

MaxWeight algorithm was first proposed by [56] in the context of scheduling for down-links 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 [48], cloud computing [43], was generalized into the back-pressure algorithm [56] in networks, and was extensively studied by [54], [16], and [49]. The generalized switch model subsumes many of these systems, and has been studied under the CRP condition using the diffusion limit method [54], and the drift method [11]. We use the notion of SSC as developed by [11]. [9] generalizes the results in [54] to stochastic processing networks, 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. We start by defining a general model, then we motivate the method with a single server queue. 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 queueing systems.

3.1 A general queueing model

We first introduce a general queueing model for stochastic processing network that includes the single server queue, the load balancing system and the generalized switch as a special case. We provide the details for each system in the corresponding section.

We consider a single hop queueing 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 \( i \in \{1, \ldots, n\} \) we let \( q_i(k) \) be the number of jobs in the \( i \)-th queue at the beginning of time slot \( k \), i.e. the number of jobs waiting to be served and the job that is being served (if any). Let \( a_i(k) \) be the number of arrivals to the \( i \)-th queue in time slot \( k \) and \( s_i(k) \) be the potential number of jobs that can be served from queue \( i \) in time slot \( k \). We say \( s_i(k) \) is potential service because, if there are not enough jobs in line, the less than \( s_i(k) \) jobs are processed. In this paper, for ease of exposition, we assume that \( a_i(k) \) and \( s_i(k) \) are upper bounded by absolute constants. Most of the results can be generalized to the case of arrival and service distributions with unbounded support after making appropriate assumptions on existence of their moments. The difference between the potential service and the actual service is the unused service \( u_i(k) \). We let \( q(k), a(k), s(k), u(k) \) be \( n \) dimensional vectors with elements \( q_i(k), a_i(k), s_i(k) \) and \( u_i(k) \), respectively, for \( i \in \{1, \ldots, n\} \). In the case of a single server queue, we omit the subindex, since \( n = 1 \). In some queueing systems, the control problem is to decide the vector \( a(k) \) in each time slot and, in others, the vector \( s(k) \). The load balancing system is an example of a queueing system where \( a(k) \) must be selected in each time slot, and the generalized switch of a system where \( s(k) \) must be selected in each time slot. We give more details about this selection processes in Section 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 \( q(k) \), the control problem is solved. Then, arrivals occur and, at the end of each time slot, the servers process jobs. Therefore, the dynamics of the queues are as follows

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

If for each \( i \in \{1, \ldots, n\} \), \( a_i(k) \) and \( s_i(k) \) depend only on \( q(k) \) or if they are independent of \( q(k) \), then Equation (1) implies that the process \( \{q(k) : k \geq 1\} \) defines a Markov chain.

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

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

Observe that Equation (2) implies

\[
g_i(k+1)u_i(k) = 0 \quad \forall i \in \{1, \ldots, n\}, \forall k \geq 1 \tag{3}
\]

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

In this paper we add a line on top of the variables and vectors to denote steady-state. Specifically, we let \( q \triangleq a(q) \), \( s \triangleq s(q) \) and \( u \triangleq u(q) \) be steady-state vectors to which \( \{q(k) : k \geq 1\} \), \( \{a(k) : k \geq 1\} \), \( \{s(k) : k \geq 1\} \) and \( \{u(k) : k \geq 1\} \). Let \( q^+ \triangleq q + u - s + \tau \) denote the queue length at time \( k+1 \) in terms of the queue length, arrival and service at time \( k \), assuming 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 distribution of scaled queue lengths in heavy-traffic, so before introducing it we specify what we mean by heavy-traffic and how we parametrize the queueing systems to obtain the limit. Take \( \epsilon > 0 \) and consider a set of queueing systems parametrized by \( \epsilon \). The parametrization is such that the vector of arrival rates is at distance \( \epsilon \) from a specific point \( r \) in the boundary of the capacity region \( C \), and the heavy-traffic limit is the limit as \( \epsilon \downarrow 0 \). In this paper we add a superscript \( \epsilon \) when we refer to the parametrized queueing system. More details on the parametrization of each queueing system will be provided once the models are completely specified, i.e. in Section 4 for the load balancing system and in Section 5 for the generalized switch.

### 3.2 Motivation for the MGF method: Single server queue

In this section we provide a proof of the well-known result that the scaled queue length of a single server queue has exponential distribution in heavy-traffic. One proof of this statement was provided by [32], where he used diffusion limits approach. We do not provide all the details of our proofs, since the single server queue is a special case of the load balancing system \( (n = 1) \) and this system is studied in details in Section 4.

We consider a single server queue operating in discrete time and let \( q(k) \) denote the number of packets at the beginning of time slot \( k \). Arrivals in each time slot form a sequence of i.i.d. random variables and the number of arrivals in time slot \( k \) is denoted by the random variable \( a(k) \). Potential service is also assumed to be a sequence of i.i.d. random variables, and the number of jobs that can be served in time slot \( k \) is denoted by \( s(k) \). Let \( u(k) \) be the unused service in time slot \( k \). The arrival and service sequences are assumed to be independent of each other. We assume there exist finite constants \( A_{max} \) and \( S_{max} \) such that \( a(k) \leq A_{max} \) and \( s(k) \leq S_{max} \) with probability 1, for all \( k = 1, 2, \ldots \). Under these assumptions, \( a(k) \) and \( s(k) \) are bounded for all \( k \geq 1 \), so their MGF’s \( E[e^{\theta a(1)}] \) and \( E[e^{\theta s(1)}] \) exist for all \( \theta \in \mathbb{R} \).

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

Let \( \bar{q}^{(\epsilon)} \) and \( \bar{s} \) be steady-state random variables to which the processes \( \{a^{(\epsilon)}(k) : k \geq 1\} \) and \( \{s(k) : k \geq 1\} \) converge in distribution as \( \epsilon \uparrow \infty \). Let \( \lambda^{(\epsilon)} = E[\bar{a}^{(\epsilon)}] \), \( (\bar{a}^{(\epsilon)})^2 = Var[\bar{a}^{(\epsilon)}] \), \( \mu = E[\bar{s}] \) and \( \sigma_s^2 = Var[\bar{s}] \). Let \( u^{(\epsilon)} \triangleq u(q^{(\epsilon)}, u^{(\epsilon)}, \bar{s}) \) be the unused service in steady-state.

In the rest of this section we prove Theorem 3.2.1. This is a well-known result and there are proofs using diffusion limits [32] and the drift method [11] in the literature. We present an alternate proof which is simpler than the two proofs mentioned above, and will serve as a template for the MGF method.

**Theorem 3.2.1.** Let \( \epsilon \in (0, \mu) \) and consider a single server queue parametrized by \( \epsilon \) as described above. Let \( q^{(\epsilon)} \) be a steady-state random variable to which \( \{q(k) : k \geq 1\} \) converges in distribution and let \( \sigma_a = \lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} \). Then, \( \epsilon q^{(\epsilon)} \Rightarrow \bar{Y} \) as \( \epsilon \downarrow 0 \), where \( \bar{Y} \) is an exponential random variable with mean \( \frac{1}{2\mu^2} \).

Before presenting the proof, we prove two key lemmas.
Lemma 3.2.2. Consider a single server queue parametrized by $\epsilon$ as described above. Then,

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

Proof. (of Lemma 3.2.2)

We set to zero the drift of $V_1(q) = q$, and we obtain

$$0 = E\left((\pi^{(\epsilon)})^+ - \pi^{(\epsilon)}\right) = E\left((\pi^{(\epsilon)} + \pi^{(\epsilon)} - \pi + \pi^{(\epsilon)} - \pi^{(\epsilon)})\right).$$

Rearranging terms we obtain

$$E\left[\pi^{(\epsilon)}\right] = E\left[\pi - \pi^{(\epsilon)}\right] = \mu - (\mu - \epsilon) = \epsilon.$$

The next lemma is an exponential version of Equation (3).

Lemma 3.2.3. Consider a single server queue parametrized by $\epsilon$ as described above. Then, for all $\alpha, \beta \in \mathbb{R}$ and each $k \geq 1$ we have

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

Proof. (of Lemma 3.2.3)

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

Now we prove Theorem 3.2.1.

Proof. (of Theorem 3.2.1)

It is well known that the Markov chain $\{q^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent for $\epsilon > 0$, so we omit the proof. For instance, the reader can find a proof using Foster-Lyapunov theorem in Theorem 3.4.2 of [53]. Then, $\pi^{(\epsilon)}$ is well defined.

If we expand the product in Lemma 3.2.3 and rearrange terms we obtain

$$e^{\theta q^{(\epsilon)}(k+1)} = 1 - e^{-\theta u^{(\epsilon)}(k)} + e^{\theta a^{(\epsilon)}(k) + a^{(\epsilon)}(k) - s(k)}$$

Observe that Equation (4) holds for all $k \geq 1$. In particular, it holds in steady-state. Also, it can be shown that $E\left[e^{\theta \pi^{(\epsilon)}}\right] < \infty$ in an interval around 0. We omit the proof because in Lemma B.3.1 we provide a proof for the load balancing system, which is a more general case. Therefore, we can set to zero the drift of $V_1(q) = e^{\theta q}$. We obtain

$$E\left[e^{\theta \pi^{(\epsilon)}}\right] = E\left[e^{\theta (\pi^{(\epsilon)})^+}\right] = 1 - E\left[e^{-\theta \pi^{(\epsilon)}}\right] + E\left[e^{\theta (\pi^{(\epsilon)} + \pi^{(\epsilon)} - \pi)}\right],$$

where the last equality holds by Equation (4). Since $\pi^{(\epsilon)}$ and $\pi$ are independent of the queue length, rearranging terms we obtain

$$E\left[e^{\theta \pi^{(\epsilon)}}\right] = \frac{1 - E\left[e^{-\theta \pi^{(\epsilon)}}\right]}{1 - E\left[e^{\theta (\pi^{(\epsilon)} - \pi)}\right]}.$$
The right hand side yields a $\frac{\theta^2}{2}$ form in the limit as $\epsilon \downarrow 0$. Then, we take Taylor series of each term. We use Lemma 3.3.2, which is presented in Section 3.3.

For the numerator we obtain

$$1 - E \left[ e^{-\theta \epsilon \pi(\epsilon)} \right] = 1 - E \left[ f_{\pi(\epsilon)}(\theta) \right] = \theta E \left[ \pi(\epsilon) \right] - \frac{(\theta \epsilon)^2}{2} E \left[ (\pi(\epsilon))^2 \right] + O(\epsilon^3)$$

where the last equality holds by Lemma 3.2.2 and because $E \left[ (\pi(\epsilon))^2 \right]$ is $O(\epsilon)$. Details of this argument will be provided in Section 4 for the load balancing system (see Claim 4.3.3), because the single server queue is a particular case.

For the denominator we obtain

$$1 - E \left[ e^{\theta \epsilon (\pi(\epsilon) - \bar{\pi})} \right] = 1 - E \left[ f_{\pi(\epsilon)}(\theta) \right] = -\theta E \left[ \pi(\epsilon) - \bar{\pi} \right] - \frac{(\theta \epsilon)^2}{2} E \left[ (\pi(\epsilon) - \bar{\pi})^2 \right] + O(\epsilon^3)$$

where the last step holds because $E \left[ \pi(\epsilon) \right] = \mu - \epsilon$ and by definition of variance.

If we replace these results in Equation (5) and cancel out $\theta \epsilon^2$ from numerator and denominator we obtain

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

Therefore, taking the heavy-traffic limit we obtain

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

Since the right hand side is the MGF of an exponential random variable with mean $\frac{\sigma_a^2 + \sigma_s^2}{\theta}$, Equation (6) implies $\pi(\epsilon)$ converges to such an exponential random variable in distribution (Theorem 9.5 in Section 5, [19]).

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

### 3.3 General framework

In the last subsection we proved a well-known result for the single server queue, using the MGF method. In this subsection we describe the method in detail for more general queueing systems that satisfy the CRP condition. Before presenting the framework, we present a formal definition of the CRP condition. We use the definition provided in [54] and [45].

**Definition 3.3.1.** Consider a set of queueing systems parametrized by $\epsilon$ as described in Section 3.1, with capacity region $C$. Suppose that in heavy-traffic, i.e. as $\epsilon \downarrow 0$, the vector of arrival rates approaches a point $\mathbf{r}$ in the boundary of $C$. We say that the queueing system satisfies the Complete Resource Pooling (CRP) condition if the outer normal vector to $C$ at $\mathbf{r}$ is unique.

The CRP condition implies that, in the heavy-traffic limit, the $n$-dimensional queueing system behaves as a one-dimensional queueing system, i.e. as a single server queue [24, 9]. Therefore, the MGF method is essentially similar to the proof of Theorem 3.2.1 after one establishes SSC.

We now present the general framework as a sequence of 5 steps.
Step 1. Positive recurrence.

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

Then, there exists a steady-state random vector $\overrightarrow{q}^{(\epsilon)}$ to which the queue lengths process $\{q^{(\epsilon)}(k) : k \geq 1\}$ converges in distribution as $k \uparrow \infty$. For ease of exposition, we let $\overrightarrow{\sigma}^{(\epsilon)}$, $\overrightarrow{\pi}^{(\epsilon)}$, $\overrightarrow{w}^{(\epsilon)}$ be the random vectors of arrivals, potential services and unused services in steady-state respectively in the system parameterized by $\epsilon$.

Step 2. State Space Collapse.

Prove SSC into a one-dimensional subspace.

Let $\mathcal{K}$ be the subspace where the state space collapses in heavy-traffic. For any $n$ dimensional vector $\overrightarrow{x}$, let $\overrightarrow{x}_\parallel$ be its projection on $\mathcal{K}$ and let $\overrightarrow{x}_\perp \triangleq \overrightarrow{x} - \overrightarrow{x}_\parallel$. In this step it should be proved that $E \left[ \|\overrightarrow{q}_\perp^{(\epsilon)}\|_m \right]$ is $o \left( \frac{1}{\epsilon^m} \right)$ for all $m = 1, 2, \ldots$, which is equivalent to proving that $\epsilon^m E \left[ \|\overrightarrow{q}_\perp^{(\epsilon)}\|_m \right]$ is $o(1)$ for all $m = 1, 2, \ldots$.

The queueing systems that we study in this paper actually exhibit stronger form of SSC, where $E \left[ \|\overrightarrow{q}_\perp^{(\epsilon)}\|_m \right]$ is $O(1)$ for all $m = 1, 2, \ldots$. However, a weaker form of SSC is studied in [61] and [60].

From this step, we conclude that

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

i.e. $\|\overrightarrow{q}_\perp^{(\epsilon)}\|$ converges to zero in the mean squares sense and, therefore, in probability.

Step 3. Set to zero the drift of a linear test function.

Set to zero the drift of $V_1(\overrightarrow{q}) = \langle \overrightarrow{c}, \overrightarrow{q} \rangle$ in steady-state, where $\overrightarrow{c}$ is the unit vector which is normal vector to the capacity region $\mathcal{C}$ at the point $\overrightarrow{r}$, i.e. $\overrightarrow{c}$ is the direction in which SSC occurs and so $\mathcal{K}$ is just the linear subspace spanned by $\overrightarrow{c}$. Note that $V_1(\overrightarrow{q})$ is a linear function, but is related to the norm $\|\overrightarrow{q}\|$ by $|V_1(\overrightarrow{q})| = \|\overrightarrow{q}\|$.

From this step we obtain an expression for $E \left[ \langle \overrightarrow{c}, \overrightarrow{u}^{(\epsilon)} \rangle \right]$, that will be used in Step 5 to compute the heavy-traffic limit.

Step 4. Prove an exponential version of Equation (3).

Prove that

$$E \left[ \left( e^{\theta \epsilon \langle \overrightarrow{c}, (\overrightarrow{q}^{(\epsilon)})^+ \rangle} - 1 \right) \left( e^{-\theta \epsilon \langle \overrightarrow{c}, \overrightarrow{w}^{(\epsilon)} \rangle} - 1 \right) \right] \text{ is } o(\epsilon^2). \quad (7)$$

This step is key in the MGF method because it allows us to obtain an explicit expression for the MGF of $\epsilon \langle \overrightarrow{c}, (\overrightarrow{q}^{(\epsilon)})^+ \rangle$ in few steps, as will be shown later in this paper. To prove Equation (7) it is essential to use SSC. Observe that this expression is valid for all traffic.

Step 5. Set the drift of $V_2(\overrightarrow{q}) = e^{\theta \epsilon \langle \overrightarrow{c}, \overrightarrow{q} \rangle}$ to zero in steady-state and compute heavy-traffic limit of the MGF of $\epsilon \langle \overrightarrow{c}, \overrightarrow{q} \rangle$.

We can only set to zero the drift of $V_2(\overrightarrow{q})$ in steady-state if the MGF of $e^{\theta \epsilon \langle \overrightarrow{c}, \overrightarrow{q} \rangle}$ exists in an interval around 0, and this must be proved. An alternative approach (where this difficulty does not arise), is to use characteristic functions, because they always exist. However, working with characteristic functions involve the use of complex analysis. Another way to overcome this difficulty is to use one-sided Laplace transform, i.e. to consider $\theta < 0$. One-sided Laplace transform of $\langle \overrightarrow{c}, \overrightarrow{q} \rangle$ always exists because $\overrightarrow{c}$ and $\overrightarrow{q}$ are nonnegative.
Nonnegativity of $c$ usually follows from the shape of the capacity region, and since it is coordinate convex. In this paper, we focus on the two-sided Laplace transform. Additionally, since our arrival and service random variables are bounded, all their moments exist.

When we set the drift of $V_2(q) = e^{\theta c \mathbb{E}(q)}$ to zero, and use Equation (7), we obtain an expression for $E \left[ e^{\theta c \mathbb{E}(q)} \right]$ that depends on the existence of the other variables that affect the dynamics of the queues. This expression yields a $\frac{\partial}{\partial \theta}$ form when we take the limit as $\epsilon \downarrow 0$. We evaluate this limit by expanding the MGF’s in Taylor expansion at $\theta = 0$. We use the following lemma for this purpose.

**Lemma 3.3.2.** Let $X^{(\epsilon)}$ be set of random variables indexed by $\epsilon > 0$. Assume $X^{(\epsilon)}$ is bounded for all $\epsilon$, i.e. there exists a constant $K_{\text{max}}$ that does not depend on $\epsilon$ such that $X \leq K_{\text{max}}$ with probability 1. Define $f_{\epsilon,X}(\theta) = e^{\theta c X}$. Then,
\[
\left| E \left[ f_{\epsilon,X}^{(\epsilon)}(\theta) \right] - 1 - \theta c E \left[ X^{(\epsilon)} \right] - \frac{\left( \theta c \right)^2}{2} E \left[ (X^{(\epsilon)})^2 \right] \right| \leq C \epsilon^3,
\]
where $C$ is a bounded constant. With a slight abuse of notation, we write the inequality above as follows
\[
E \left[ f_{\epsilon,X}(\theta) \right] = 1 + \theta c E \left[ X \right] + \frac{\left( \theta c \right)^2}{2} E \left[ X^2 \right] + O(\epsilon^3).\tag{8}
\]

The proof of Lemma 3.3.2 is straightforward, and we present it in Appendix A.

**Remark 3.3.3.** Since we are working with a bounded random variable, the proof that we presented of Lemma 3.3.2 was straightforward. However, in general, one needs an assumption on the existence of the MGF. 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 3.3.2. For instance, Theorem 2.3.3. in [40] can be used when one is working with characteristic functions.

From this step, we obtain that the MGF of $\epsilon(c, q^{(\epsilon)})$ converges to random variable $Y$ with exponential distribution. Then, $c q^{(\epsilon)} = \epsilon(c, q^{(\epsilon)}) c \Rightarrow Y c$ as $\epsilon \downarrow 0$ because $c$ is a fixed vector. Since we know from SSC in Step 2 that $c q^{(\epsilon)} \rightarrow 0$ in probability, using Slutsky’s theorem (Theorem 11.4 in Section 5, [19]), we have that $c q^{(\epsilon)} = \epsilon q^{(\epsilon)} + \epsilon q^{(\epsilon)} \Rightarrow Y c$ as $\epsilon \downarrow 0$.

## 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 $E[s_i(1)] = \mu_i$, and let $\mu_\Sigma = \sum_{i=1}^{n} \mu_i$. 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 $\text{Cov}[s_i, s_j]$ be the covariance between $s_i(1)$ and $s_j(1)$. We assume there exists a finite constant $S_{\text{max}}$ such that $s_i(k) \leq S_{\text{max}}$ with probability 1 for all $i \in \{1, \ldots, n\}$ and all $k \geq 1$. 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 the routing policy is fixed for all $k \geq 1$, but we do not assume any particular
policy, to avoid losing generality. After routing, \( a_i(k) \) is the number of arrivals that were routed to the \( i \)th queue in time slot \( k \), for \( i \in \{1, \ldots, n\} \). Let \( \lambda = E[a(1)] \) and \( \sigma_a^2 = \text{Var}[a(1)] \). We assume there exists a finite constant \( A_{\text{max}} \) such that \( a(k) \leq A_{\text{max}} \) with probability 1 for all \( k \geq 1 \), and that the arrival process is independent of the queue lengths and of the service processes. The dynamics of the queues are according to Equation (2).

It is well known that the capacity region of the load balancing problem is \( C = \{ \lambda \in \mathbb{R}_+ : \lambda < \mu \Sigma \} \). A proof can be found in Appendix A of [11].

To study the heavy-traffic limit of this queueing system, we parametrize the arrival process as follows. For \( \epsilon \in (0, \mu \Sigma) \), we consider a load balancing system as described above, where the arrival process \( \{a^{(\epsilon)}(k) : k \geq 1\} \) is such that \( E[a^{(\epsilon)}(1)] = \mu \Sigma - \epsilon \) and \( \text{Var}[a^{(\epsilon)}(1)] = \sigma_a^{(\epsilon)} \). In other words, the arrival rates approach the point \( r = \mu \Sigma \) in the boundary of \( C \). Observe that the capacity region \( C \) of the load balancing system is one-dimensional, so the normal vector at \( r = \mu \Sigma \) is unique. Therefore, this queueing system satisfies the CRP condition as defined in Definition 3.3.1.

### 4.2 MGF method applied to load balancing systems

In this subsection we state the main theorem of this section and provide some examples, and in the next subsection we will 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 4.2.1.** A routing algorithm \( \mathcal{A} \) is throughput optimal for the load balancing system described in Section 4.1 if the Markov chain \( \{q^{(\epsilon)}(k) : k \geq 1\} \) operating under \( \mathcal{A} \) is positive recurrent for all \( \epsilon \in (0, \mu \Sigma) \).

**Definition 4.2.2.** Consider a routing algorithm \( \mathcal{A} \). Let

\[ K = \{x \in \mathbb{R}^n : x_i = x_j \quad \forall i, j \in \{1, \ldots, n\} \} \]

and for any vector \( y \in \mathbb{R}^n \), let \( y_\parallel \) be the projection of \( y \) on \( K \) and let \( y_\perp \triangleq y - y_\parallel. \) We say that the algorithm \( \mathcal{A} \) satisfies SSC if the load balancing system described in Section 4.1 operating under \( \mathcal{A} \) satisfies the following property.

\[ \epsilon E \left[ \left\| q^{(\epsilon)}(k) \right\|_m^m \right] = o \left( \frac{1}{\epsilon^m} \right) \quad \forall m = 1, 2, \ldots, \]

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

Now we formally present the result.

**Theorem 4.2.3.** Let \( \epsilon \in (0, \mu \Sigma) \) and consider a set of load balancing systems parametrized by \( \epsilon \), 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 \( \overline{q}^{(\epsilon)} \) be a steady-state random vector to which the queue lengths process \( \{q^{(\epsilon)}(k) : k \geq 1\} \) converges in distribution. Further, assume \( \lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a \) and let \( \overline{q} \) be a steady-state vector to which the process \( \{s(k) : k \geq 1\} \) converges in distribution. Then \( \epsilon \overline{q}^{(\epsilon)} \Rightarrow \overline{Y} \) as \( \epsilon \downarrow 0 \), where \( \overline{Y} \) is an exponential random variable with mean \( \frac{1}{2n} \left( \sigma_a^2 + \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[\overline{s}_i, \overline{s}_j] \right) \).

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

**Definition 4.2.4.** Consider a load balancing system as described in Section 4.1. Then, for each \( k \geq 1 \), given the vector of queue lengths \( q^{(\epsilon)}(k) \), the routing policy selects \( i^*(k) \) and sends arrivals according to the following formula.

\[ a_i(k) = \begin{cases} a(k), & \text{if } i = i^*(k) \\ 0, & \text{otherwise} \end{cases} \]
The routing policy Join the Shortest Queue (JSQ) sends all arrivals in time slot \( k \) to the queue with the least number of packets, breaking ties at random. Formally, under JSQ routing policy:

\[
i^*(k) \in \arg\min_{i \in \{1, \ldots, n\}} \left\{ q^{(c)}_i(k) \right\},
\]

breaking ties at random.

The routing policy power-of-two choices selects two queues uniformly at random, say \( i_1, i_2 \in \{1, \ldots, n\} \) 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 power-of-two choices, if queues \( i_1 \) and \( i_2 \) are selected,

\[
i^*(k) \in \arg\min_{i \in \{i_1, i_2\}} \left\{ q^{(c)}_i(k) \right\},
\]

breaking ties at random.

In the following two corollaries we show that these routing policies satisfy the assumptions of Theorem 4.2.3 and, therefore, the distribution of the scaled queue lengths of a load balancing system operating under these policies is exponential in heavy-traffic.

**Corollary 4.2.5.** Consider a set of load balancing systems parametrized by \( \epsilon \) as described in Section 4.1, operating under JSQ routing policy. Then, \( \epsilon \bar{\mathbf{q}}^{(c)} \Rightarrow \bar{\Upsilon}_1 \mathbf{1} \) as \( \epsilon \downarrow 0 \), where \( \bar{\Upsilon}_1 \) is an exponential random variable with mean

\[
\frac{1}{2n} \left( \sigma_a^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} [s_i, s_j] \right).
\]

A particular case of the queueing system described in Corollary 4.2.5 is the load balancing system operating under JSQ with independent servers. In this case, \( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} [s_i, s_j] \) reduces to the sum of variances of the servers. This is one of the queueing systems studied by [11].

**Proof.** (of Corollary 4.2.5) We only need to check throughput optimality of JSQ and that it satisfies SSC, as defined in Definition 4.2.2.

[11] 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 [11]. We provide a sketch of SSC as proved in [11] in Appendix B.1.

**Corollary 4.2.6.** Consider a set of load balancing systems parametrized by \( \epsilon \) as described in Section 4.1, operating under Power-of-two choices and where all the servers are identical and independent. Let \( \sigma_s^2 \) be the variance of \( s_1(k) \). Then, \( \epsilon \bar{\mathbf{q}}^{(c)} \Rightarrow \bar{\Upsilon}_2 \mathbf{1} \) as \( \epsilon \downarrow 0 \), where \( \bar{\Upsilon}_2 \) is an exponential random variable with mean

\[
\frac{1}{2n} \left( \sigma_a^2 + n \sigma_s^2 \right).
\]

**Proof.** (of Corollary 4.2.6) Similarly to the proof of Corollary 4.2.5, we check throughput optimality and that SSC is satisfied.

In Section 4.3 of [44] the authors prove SSC. Throughput optimality can be proved using Foster-Lyapunov theorem and the calculations that [44] develop in the proof of SSC. We omit the proof in this paper, since our goal is to introduce the MGF method.

The assumption of independent service process is not required for this corollary. However, in order to relax this assumption, one has to prove SSC, which is straight forward extension of SSC in [44] along the lines of the proof for JSQ in Appendix B.1. However, the assumption of identical service is essential for the power-of-two algorithm to be throughput optimal. The case when the servers are not identical was studied in [7] 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 \( \mu_i \), for all \( i \in \{1, \ldots, n\} \). Then, the arrivals are sent to the server with the shortest queue among the \( d \) selected servers. They prove
that this queueing 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 defined in Definition 4.2.2 is established. This is straightforward extension, and we do not present the details here because the focus is on illustrating the MGF approach.

In this subsection we presented the main theorem of this section, and we presented 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 exponential distribution, we only need to check these two assumptions.

4.3 Proof of Theorem 4.2.3

In the rest of this section we prove Theorem 4.2.3 using the MGF method. For ease of exposition, we omit the dependence on $\epsilon$ of the variables in our proofs. Before presenting the proofs we prove two lemmas.

Let $\bar{a}(\epsilon) \triangleq a^{(\epsilon)}(\bar{q})$ and $\bar{s}$ be steady-state vectors to which $\{a^{(\epsilon)}(k) : k \geq 1\}$ and $\{s(k) : k \geq 1\}$ converge in distribution. We let $\bar{u}(\epsilon) \triangleq u(\bar{q}(\epsilon), \bar{a}(\epsilon), \bar{s})$ denote the unused service in steady-state.

**Lemma 4.3.1.** Consider a load balancing system as described in Section 4.1, operating under any throughput optimal routing policy. Then,

$$E \left[ \sum_{i=1}^{n} \frac{\pi_i^{(\epsilon)}}{\pi_i} \right] = \epsilon.$$

**Proof.** We set to zero the drift of $V_1(q) = \langle c, q \rangle$ in steady-state. In this case, from the definition of $K$ in Definition 4.2.2 we have $c = \frac{1}{\sqrt{n}} 1$. Then, we obtain

$$0 = E \left[ V_1 \left( \left( \frac{\bar{q}(\epsilon)}{\bar{q}(\epsilon)} \right) \right) - V_1 \left( \frac{\bar{q}(\epsilon)}{\bar{q}(\epsilon)} \right) \right]$$

$$= \frac{1}{\sqrt{n}} E \left[ \sum_{i=1}^{n} \left( \frac{\bar{q}_i^{(\epsilon)}}{\bar{q}_i^{(\epsilon)}} \right) - \sum_{i=1}^{n} \bar{q}_i^{(\epsilon)} \right]$$

$$= \frac{1}{\sqrt{n}} E \left[ \sum_{i=1}^{n} \left( \frac{\bar{q}_i^{(\epsilon)}}{\bar{q}_i^{(\epsilon)}} + \bar{a}_i^{(\epsilon)} - \bar{s}_i + \bar{u}_i^{(\epsilon)} \right) - \sum_{i=1}^{n} \bar{q}_i^{(\epsilon)} \right]$$

$$= \frac{1}{\sqrt{n}} E \left[ \bar{a}(\epsilon) - \sum_{i=1}^{n} \bar{s}_i + \sum_{i=1}^{n} \bar{u}_i^{(\epsilon)} \right]$$

where $(a)$ holds by definition of $\bar{q}^{+}$; and $(b)$ holds because $\bar{a}$ is the total number of arrivals to the queueing system in one time slot and for each $i \in \{1, \ldots, n\}$, $\bar{a}_i$ is the number of arrivals routed to the $i$th queue. Rearranging terms and canceling $\frac{1}{\sqrt{n}}$, we obtain

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

$$= \sum_{i=1}^{n} \mu_i - (\mu_\Sigma - \epsilon)$$

$$= \epsilon,$$

where $(a)$ holds by definition of $\bar{a}(\epsilon)$; and $(b)$ holds by definition of $\mu_\Sigma$. $\square$
The following Lemma generalizes Lemma 3.2.3 for the load balancing system.

**Lemma 4.3.2.** Consider the load balancing system parametrized by $\epsilon$ described in Theorem 4.2.3. Then, for any real number $\theta$ we have

$$E \left[ \left( e^{\theta \epsilon \sum_{i=1}^{n} (\pi_i^{(\epsilon)})^+} - 1 \right) \left( e^{-\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} - 1 \right) \right] \text{ is } o(\epsilon^2)$$

We present the proof of Lemma 4.3.2 in Appendix B.2.

Now we prove Theorem 4.2.3.

**Proof:** Proof of Theorem 4.2.3.

We use the MGF method.

**Step 1. Positive recurrence.**

One hypothesis of the theorem is that the routing policy is throughput optimal. Therefore, for all $\epsilon \in (0, \mu_\Sigma)$, the Markov chain $\{q^{(\epsilon)}(k) : k \geq 1\}$ is positive recurrent.

**Step 2. State Space Collapse.**

It is also one of the assumptions of Theorem 4.2.3 that SSC holds, so there is nothing to prove.

**Step 3. Use linear test function.**

We set to zero the drift of $V_1(q) = \langle c, q \rangle$, where $c = \frac{1}{\sqrt{n}} \mathbf{1}$. This is exactly what we did in the proof of Lemma 4.3.1. Then, this step yields Lemma 4.3.1.

**Step 4. Prove exponential version of Equation (3)**

Since $\langle c, q \rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} n_i$, this step is equivalent to Lemma 4.3.2 using $\frac{\theta}{\sqrt{n}}$ instead of $\theta$. For ease of exposition, we work with $\theta$ in the rest of this proof.

**Step 5. Set to zero the drift of $V_2(q) = e^{\theta \epsilon \langle c, q \rangle}$ in steady-state and compute heavy-traffic limit of the MGF of $\epsilon(c, q)$.**

Before setting to zero the drift of $V_2(q)$ we need to prove that the MGF of $\epsilon(c, \overline{q}^{(\epsilon)})$ exists in an interval around 0. The proof is presented in Appendix B.3.

Now we can set to zero the drift of $V_2(q)$. Note that $P[A^{(\epsilon)} - \sum_{i=1}^{n} \bar{\xi}_i \neq 0] > 0$ whenever $\epsilon > 0$. If we expand the product in the expression of Lemma 4.3.2 we obtain

$$E \left[ e^{\theta \epsilon \sum_{i=1}^{n} (\pi_i^{(\epsilon)})^+} \right] = 1 - E \left[ e^{-\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right] + o(\epsilon^2).$$

Recall $\sum_{i=1}^{n} \pi_i^{(\epsilon)} = \pi^{(\epsilon)}$ and $\pi^{(\epsilon)}, \bar{\pi}$ are independent of $\overline{q}^{(\epsilon)}$. Also, since $E \left[ e^{\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right] < \infty$, we have

$$E \left[ e^{\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right] = E \left[ e^{\theta \epsilon \sum_{i=1}^{n} (\pi_i^{(\epsilon)})^+} \right].$$

Therefore, reorganizing terms we obtain

$$E \left[ e^{\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right] = \frac{1 - E \left[ e^{-\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right] + o(\epsilon^2)}{1 - E \left[ e^{\theta \epsilon \sum_{i=1}^{n} \pi_i^{(\epsilon)}} \right]} \quad (9)$$
Now we take the limit. To do that, we take Taylor series of each term in the right hand side of Equation (9). We use Lemma 3.3.2.

For the numerator, we obtain

\[
1 - E \left[ e^{-\theta \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = 1 - E \left[ f_{e^{-\theta \Sigma_{i=1}^{n} \bar{\pi}_i}} (\theta) \right] = \theta \epsilon E \left[ \frac{\theta \epsilon}{2} \right] + O(\epsilon^3)
\]

where the last equality holds by Lemma 4.3.1.

**Claim 4.3.3.** Consider a load balancing system as described in Theorem 4.2.3. Then,

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

We prove the claim in Appendix D.1. Using the claim in Equation (10) we obtain

\[
1 - E \left[ e^{-\theta \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = \theta \epsilon^2 + O(\epsilon^3),
\]

(11)

For the denominator, we obtain

\[
1 - E \left[ e^{\theta \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = 1 - E \left[ f_{e^{\theta \Sigma_{i=1}^{n} \bar{\pi}_i}} (\theta) \right] = - \theta \epsilon \left( \bar{\pi} - \sum_{i=1}^{n} \bar{\pi}_i \right) - \frac{(\theta \epsilon)^2}{2} E \left[ \left( \bar{\pi} - \sum_{i=1}^{n} \bar{\pi}_i \right)^2 \right] + O(\epsilon^3)
\]

where the last step holds because \( E[\bar{\pi}] = \mu_\Sigma - \epsilon \), \( E[\Sigma_{i=1}^{n} \bar{\pi}_i] = \mu_\Sigma \) and by definition of covariance.

Using Equations (11) and (12) in (9), and since \( O(\epsilon^3) \) is \( o(\epsilon^2) \), we obtain

\[
E \left[ e^{\theta \epsilon \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = \frac{\theta \epsilon^2 + o(\epsilon^2)}{\theta \epsilon^2 - \frac{(\theta \epsilon)^2}{2} \left( \sigma^{(e)}_a \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\bar{\pi}_i, \bar{\pi}_j] + \epsilon^2} + O(\epsilon^3)
\]

Factorizing \( \theta \epsilon^2 \) from the numerator and denominator, we obtain

\[
E \left[ e^{\theta \epsilon \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = \frac{1 + \frac{1}{\theta} o(\epsilon^2)}{1 - \frac{1}{\theta} \left( \sigma^{(e)}_a \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\bar{\pi}_i, \bar{\pi}_j] + O(\epsilon)}
\]

Therefore, taking the limit we obtain

\[
\lim_{\epsilon \to 0} E \left[ e^{\theta \epsilon \Sigma_{i=1}^{n} \bar{\pi}_i} \right] = \frac{1}{1 - \theta \left( \sigma^{(e)}_a \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\bar{\pi}_i, \bar{\pi}_j] \right)}.
\]

15
which is the MGF of an exponential random variable with mean \( \frac{1}{2} \left( \sigma_a^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\eta_i, \eta_j] \right) \). Then,

\[
\epsilon(c, \eta(c))c = \epsilon \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \right) \Rightarrow \tilde{\Upsilon} \mathbf{1} \quad \text{as } \epsilon \downarrow 0,
\]

where \( \tilde{\Upsilon} \) is an exponential random variable with mean

\[
\frac{1}{2 \sqrt{n}} \left( \sigma_a^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[\eta_i, \eta_j] \right).
\]

Therefore, we conclude that \( \epsilon c^{(c)} = \epsilon c^{(1)} + \epsilon c^{(2)} \Rightarrow \tilde{\Upsilon} \mathbf{1} \) as \( \epsilon \downarrow 0 \). This proves Theorem 4.2.3.

\[\square\]

## 5 Generalized switch

In this section we apply the MGF method in the context of a generalized switch. We compute the distribution of the scaled vector of queue lengths in heavy-traffic when SSC occurs into a one-dimensional subspace. The generalized switch is a model that was first introduced by [54], and it represents a generalization of a variety of queueing systems, such as the input-queued switch [48], cloud computing [43], down-links in wireless base stations [56], etc.

### 5.1 Generalized switch model

Consider a system with \( n \) separate queues, as described in Section 3. For each \( i \in \{1, \ldots, n\} \), 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 \( i, j \in \{1, \ldots, n\} \), let \( \lambda_i = E[a_i(1)] \) and \( \text{Cov}[a_i, a_j] \) be the covariance between \( a_i(k) \) and \( a_j(k) \) for all \( k \geq 1 \). We assume there exists a finite constant \( A_{\max} \) such that \( a_i(k) \leq A_{\max} \) with probability 1 for all \( i \in \{1, \ldots, n\} \) and for all \( k \geq 1 \). For each \( k \geq 1 \), let \( \mathbf{a}(k) \) be an \( n \)-dimensional vector with elements \( a_i(k) \) for \( i \in \{1, \ldots, n\} \), and \( \mathbf{\lambda} \) be an \( n \)-dimensional vector with elements \( \lambda_i, i \in \{1, \ldots, n\} \).

The servers interfere with each other, so the set of active servers must satisfy interference constraints in each time slot. Additionally, there are conditions of the environment that affect these constraints. We group all these conditions in a variable that we call 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 the arrival processes. We assume that the state space for the channel state is a finite set \( \mathcal{T} \) and we let \( \psi_t \triangleq P[T(k) = t] \) for each \( t \in \mathcal{T} \). If the channel state is \( T(k) = t \), we let \( \mathcal{S}^{(t)} \) be the set of feasible service rate vectors and we assume it is a finite set for each \( t \in \mathcal{T} \). We assume that if \( \mathbf{x} \in \mathcal{S}^{(t)} \) for some \( t \in \mathcal{T} \), then \( \mathbf{x} - x_i e^i \in \mathcal{S}^{(t)} \) as well, for all \( i \in \{1, \ldots, n\} \). Here, for each \( i \in \{1, \ldots, n\} \), \( x_i \) is the \( i \)th element of \( \mathbf{x} \) and \( e^i \) is the \( i \)th canonical vector, i.e. an \( n \)-dimensional vectors with a 1 in the \( i \)th elements and zeroes in all other entries. In other words, the sets \( \mathcal{S}^{(t)} \) contain the projection on the coordinate axes of all its vectors.

In this queueing system the control problem which is a scheduling problem, is to select \( s(k) \) in each time slot after realizing the channel state. Let \( s(k) \) be the solution of the scheduling problem in time slot \( k \). Since \( \mathcal{S}^{(t)} \) 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 \( i \in \{1, \ldots, n\} \) and all \( k \geq 1 \).

In each time slot, the order of events is as follows. First, the channel state and the queue lengths are observed. Second, the scheduling problem is solved. Third, arrivals occur, and at the end of each time slot, the servers process jobs. Therefore, the dynamics of the queues are described by Equation (2). Observe that the service rates vector in each time slot is now a function of the queue lengths and the channel state, and it is no longer a random variable.

It is known [11] that the capacity region of this queueing system is

\[
\mathcal{C} = \sum_{t \in \mathcal{T}} \psi_t \text{ConvexHull} \left\{ \mathcal{S}^{(t)} \right\}
\]

(13)
Since for each $t \in \mathcal{T}$, the set $\mathcal{S}^{(t)}$ is finite and $\mathcal{T}$ is finite too, then the set $\mathcal{C}$ is the convex hull of finitely many points and, therefore, it is a polytope. In order to exploit this structure, we describe it as the intersection of a finite number of half-spaces. Let $L$ be the minimal number of hyperplanes that are required to describe $\mathcal{C}$, and for each $\ell \in \{1, \ldots, L\}$ let $c^{(\ell)} \in \mathbb{R}^n$ and $b^{(\ell)} \in \mathbb{R}$ be the parameters that define each facet of the polytope. Then,

$$\mathcal{C} = \left\{ x \in \mathbb{R}^n : \langle c^{(\ell)}, x \rangle \leq b^{(\ell)} \quad \text{for all } \ell \in \{1, \ldots, L\} \right\}. \quad (14)$$

Without loss of generality we assume $c^{(\ell)} \geq 0$, $\|c^{(\ell)}\| = 1$ and $b^{(\ell)} > 0$ for all $\ell \in \{1, \ldots, L\}$, 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 $\ell \in \{1, \ldots, L\}$ we let $\mathcal{F}^{(\ell)} = \{ r \in \mathcal{C} : \langle c^{(\ell)}, r \rangle = b^{(\ell)} \}$ be the $\ell^{\text{th}}$ facet of the polytope $\mathcal{C}$.

Observe that, by the definition of the capacity region $\mathcal{C}$ given in Equation (13), for each $k \geq 1$, the service rates vector $s(k)$ does not necessarily belong to $\mathcal{C}$. Therefore, for each $\ell \in \{1, \ldots, L\}$ and each $t \in \mathcal{T}$ we define the maximum $c^{(\ell)}$-weighted service rate available in channel state $t$ [11] as follows

$$b^{(t,\ell)} = \max_{s \in \mathcal{S}^{(t)}} \langle c^{(\ell)}, s \rangle. \quad (15)$$

In other words, given that the channel state is $t$, $b^{(t,\ell)}$ is a real number such that the hyperplane $\mathcal{H}^{(t,\ell)} = \{ x \in \mathbb{R}^n : \langle c^{(\ell)}, x \rangle = b^{(t,\ell)} \}$ is tangent to the boundary of ConvexHull $\{ \mathcal{S}^{(t)} \}$. Let $\{ B_t(k) : k \geq 1 \}$ be a sequence of i.i.d. random variables such that $P[B_t(k) = b^{(t,\ell)}] = \psi_t$ and $\sigma_{B_t}^2 = \text{Var}[B_t(k)]$.

To do heavy-traffic analysis, we fix a facet $\mathcal{F}^{(\ell)}$ and we study a set of generalized switches where $\lambda$ approaches a fixed point in the interior of $\mathcal{F}^{(\ell)}$. Formally, we fix $r^{(\ell)}$ in the interior of $\mathcal{F}^{(\ell)}$ and we let $\epsilon > 0$. Then, the system parametrized by $\epsilon$ is such that $E[a^{(\epsilon)}_i(k)] = r^{(\ell)} - \epsilon c^{(\ell)}$ and $\text{Var}[a^{(\epsilon)}_i(k)] = \left( \sigma_{a_i}^{(\epsilon)} \right)^2$ for each $i \in \{1, \ldots, n\}$. In this case, since the point $r = r^{(\ell)}$ of the boundary of the capacity region $\mathcal{C}$ is in the interior of a facet, there is a unique outer normal vector to $\mathcal{C}$ at $r$. In fact, this normal vector is $c^{(\ell)}$. Therefore, the CRP condition as defined in Definition 3.3.1 is satisfied. Observe that if $r$ 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 $\mathcal{C}$ at $r$. In fact, these outer normal vectors form a cone.

### 5.2 MGF method applied to generalized switches

In this subsection we state the main theorem of this section and we provide some examples. In the next subsection we prove the theorem. Before presenting the statement we introduce the following definitions.

**Definition 5.2.1.** A scheduling algorithm $\mathcal{A}$ is throughput optimal for the generalized switch described in Section 5.1 if for every $\epsilon \in (0,1)$ the Markov chain $\{ q^{(\epsilon)}(k) : k \geq 1 \}$ is positive recurrent.

**Definition 5.2.2.** Consider a scheduling algorithm $\mathcal{A}$. Let

$$\mathcal{K} = \left\{ x \in \mathbb{R}^n : x = \alpha c^{(\ell)} \text{ for some } \alpha \geq 0 \right\}$$

and for any vector $y \in \mathbb{R}^n$, let $y \overset{\triangle}{=} \langle c^{(\ell)}, y \rangle c^{(\ell)}$ and $y_\bot \overset{\triangle}{=} y - y \|$. We say that the algorithm $\mathcal{A}$ satisfies SSC if the generalized switch described in Section 5.1 operating under $\mathcal{A}$ satisfies the following property

$$\epsilon E \left[ \left\| \overline{q}^{(\epsilon)} \| \right|^m \right] = o \left( \frac{1}{\epsilon^m} \right) \quad \forall m = 1, 2, \ldots ,$$

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

We now present the main result of this section.
Theorem 5.2.3. Let $\epsilon \in (0, 1)$. Given the $t^{th}$ facet of $C$, $F^{(t)}$, and a vector $r^{(t)}$ in the relative interior of $F^{(t)}$, consider a set of generalized switches parametrized by $\epsilon$, as described in Section 5.1. Suppose that the scheduling algorithm is throughput optimal and satisfies SSC. For each $\epsilon$, let $\mathbf{q}^{(t)}$ be a steady-state vector to which the queue lengths process \( \{q^{(t)}(k) : k \geq 1\} \) converges in distribution. Further, let $\mathbf{e}^{(t)}$ be a steady-state vector to which \( \{e^{(t)}(k) : k \geq 1\} \) converges in distribution and let \( \lim_{\epsilon \downarrow 0} \text{Cov} \left[ \mathbf{q}^{(t)}, \mathbf{e}^{(t)} \right] = \text{Cov} \left[ \mathbf{q}, \mathbf{e} \right] \) for all $i, j \in \{1, \ldots, n\}$. Then, $\mathbf{q}^{(t)} \Rightarrow \mathbf{T} \mathbf{e}^{(t)}$ as $\epsilon \downarrow 0$, where $\mathbf{T}$ is an exponential random variable with mean

\[
\frac{1}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} c_i^{(t)} c_j^{(t)} \text{Cov} \left[ q_i, q_j \right] + \sigma_{Bt}^2 \right),
\]

where $c_i^{(t)}$ is the $i^{th}$ element of $c^{(t)}$, for each $i \in \{1, \ldots, n\}$.

Before presenting examples of queueing systems for which the assumptions of Theorem 5.2.3 are satisfied, we define the MaxWeight scheduling algorithm.

Definition 5.2.4. Consider a generalized switch as described in Section 5.1. Then, for each $k \geq 1$, given the vector of queue lengths $q^{(t)}(k)$ and the channel state $T(k) = t$, MaxWeight algorithm selects $s(k)$ that satisfies

\[
s(k) \in \arg \max_{a \in S^{(t)}} \langle x, q^{(t)}(k) \rangle,
\]

breaking ties at random.

In the next corollaries we present examples of queueing systems that satisfy the assumptions of Theorem 5.2.3.

Corollary 5.2.5. Consider a set of generalized switches parametrized by $\epsilon$ as described in Section 5.1, operating under MaxWeight algorithm. Then, $\mathbf{q}^{(t)} \Rightarrow \mathbf{T}_1 \mathbf{e}^{(t)}$ as $\epsilon \downarrow 0$, where $\mathbf{T}_1$ is an exponential random variable with mean

\[
\frac{1}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} c_i^{(t)} c_j^{(t)} \text{Cov} \left[ q_i, q_j \right] + \sigma_{Bt}^2 \right).
\]

Proof. (of Corollary 5.2.5) It suffices to check that MaxWeight is throughput optimal and that it satisfies SSC. A proof of throughput optimality can be found in Proposition 2 of [54] and SSC is proved in Proposition 6 of [11].

Therefore, the assumptions of Theorem 5.2.3 are satisfied, which proves the Corollary.

Corollary 5.2.6. Consider a set of generalized switches parametrized by $\epsilon$, as described in Section 5.1, operating under MaxWeight algorithm. Suppose that $T$ has only one element, i.e. the channel state is fixed over time. Then, $\mathbf{q}^{(t)} \Rightarrow \mathbf{T}_2 \mathbf{e}^{(t)}$, where $\mathbf{T}_2$ is an exponential random variable with mean

\[
\frac{1}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} c_i^{(t)} c_j^{(t)} \text{Cov} \left[ q_i, q_j \right] \right).
\]

The queueing system described in Corollary 5.2.6 is also known as ad hoc wireless network. Observe that this system is a particular case of the generalized switch operating under MaxWeight described in Corollary 5.2.5. However, in an ad hoc wireless network we have $\sigma_{Bt}^2 = 0$ because the channel state is not a random variable anymore. Then, the proof of Corollary 5.2.5 is valid for Corollary 5.2.6. The input-queued switch or a cross bar switch [53, 42, 41] is yet another system that is well studied. When only one port of the switch is saturated, it satisfies the CRP condition [54], and forms a special case of Corollary 5.2.6.

5.3 Proof of Theorem 5.2.3

In the rest of this section we prove Theorem 5.2.3 using the MGF method. Before presenting the proofs, we introduce three lemmas that we will use.

Let $\mathbf{T}$ and $\mathbf{F}$ be steady-state random variables to which \( \{T(k) : k \geq 1\} \) and \( \{B_t(k) : k \geq 1\} \) converge in distribution, respectively. Let $\mathbf{q}^{(t)} \overset{\Delta}{=} s(\mathbf{q}^{(t)}, \mathbf{T})$ and $\mathbf{u}^{(t)} \overset{\Delta}{=} u(\mathbf{q}^{(t)}, \mathbf{u}^{(t)}, \mathbf{F})$ be the service rates and unused service vectors in steady-state, respectively.

For ease of exposition, we omit the dependence on $\epsilon$ of the variables in our proofs.
**Lemma 5.3.1.** Consider a generalized switch as described in Section 5.1, operating under a throughput optimal scheduling policy. Then,

\[ E \left[ \langle c, \mathbf{u}(\epsilon) \rangle \right] + E \left[ \overline{B} \right] - E \left[ \langle c^{(\ell)}(\epsilon), \overline{s}(\epsilon) \rangle \right] = \epsilon. \]

**Proof. (of Lemma 5.3.1)**

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

\[
0 = E \left[ \langle c^{(\ell)}, q \rangle \right] - \langle c^{(\ell)}, \overline{q} \rangle \\
= E \left[ \langle c^{(\ell)}, q + a - q + \overline{a} \rangle - \langle c^{(\ell)}, \overline{q} \rangle \right] \\
= E \left[ \langle c^{(\ell)}, \overline{q} \rangle - E \left[ \langle c^{(\ell)}, \overline{s}(\epsilon) \rangle \right] + E \left[ \langle c^{(\ell)}, \overline{u}(\epsilon) \rangle \right]. \tag{16}
\]

Now, observe that by definition of the random variable \( B \), we have

\[
E \left[ B \right] = \sum_{t \in T} \psi_t b^{(t,\ell)}(a) = \sum_{t \in T} \psi_t \max_{s \in S(t)} \langle c^{(\ell)}, s \rangle \\
= \max_{s \in C} \langle c^{(\ell)}, s \rangle = b^{(\ell)} \tag{17}
\]

where \( (a) \) holds by the definition of \( b^{(t,\ell)} \) given in Equation (15); \( (b) \) holds by definition of the capacity region \( C \) given in Equation (13); and \( (c) \) holds by definition of the \( \ell^{th} \) facet and because the objective function in the maximization problem is linear.

Therefore,

\[
E \left[ \langle c^{(\ell)}, \overline{u} \rangle \right] = \langle c^{(\ell)}, \overline{r}^{(\ell)}(\epsilon) - \epsilon c^{(\ell)} \rangle \\
= \langle c^{(\ell)}, \overline{r}^{(\ell)} \rangle - \epsilon \left\| c^{(\ell)} \right\|^2 \\
= b^{(\ell)} - \epsilon \\
= E \left[ \overline{B} \right] - \epsilon, \tag{18}
\]

where \( (a) \) holds because \( \overline{r}^{(\ell)} \in \overline{F}^{(\ell)} \); and \( (b) \) holds by Equation (17).

Then, using Equation (18) in (16) and rearranging terms we obtain the result. \( \square \)

The next lemma generalized Lemma 3.2.3 for the case of a generalized switch.

**Lemma 5.3.2.** Consider a generalized switch parametrized by \( \epsilon \) as described in Theorem 5.2.3. Then, for any real number \( \theta \) we have

\[
E \left[ e^{\theta \epsilon \langle c^{(\ell)}, (\overline{s}(\epsilon))^+ \rangle} - 1 \right] \left( e^{-\theta \epsilon \langle c^{(\ell)}, \overline{s}(\epsilon) \rangle} - 1 \right) \text{ is } o(\epsilon^2)
\]

We present the proof of Lemma 5.3.2 in Appendix C.1.

When we applied the MGF method to the single server queue and to the load balancing system, we used that the service rates vector is independent of the queue lengths vector. However, in the case of the generalized switch this is no longer true. To overcome this difficulty we use the following lemma.
Lemma 5.3.3. Consider a generalized switch parametrized by $\epsilon$, as described in Theorem 5.2.3. Then, for any $\theta \in \mathbb{R}$ we have
\[
E \left[ \left( e^{\theta c(c^l,\pi^v)} - 1 \right) \left( e^{\theta a(c^l,T-c^l,\pi^v)} - 1 \right) \right] \text{ is } o(\epsilon^2).
\]

We present the proof in Appendix C.2.

Now we prove Theorem 5.2.3.

Proof. (of Theorem 5.2.3)

We use the MGF method.

Step 1. Positive recurrence.

One of the hypothesis of the theorem is that the scheduling policy is throughput optimal. Therefore, the Markov chain $\{q^l(k) : k \geq 1\}$ is positive recurrent for $\epsilon > 0$.

Step 2. State Space Collapse.

It is also one of the assumptions that SSC holds, so there is nothing to prove.

Step 3. Use linear test function.

In this case $c = c^l$, so we set to zero the drift of $V_1(q) = \langle c^l, q \rangle$. This is what we did in Lemma 5.3.1. Therefore, from this step we obtain Lemma 5.3.1.

Step 4. Prove exponential version of Equation (3)

This step is exactly what we proved in Lemma 5.3.2.

Step 5. Set to zero the drift of $V_2(q) = e^{c\langle c, q \rangle}$ in steady-state and compute heavy-traffic limit of the MGF of $e^{c\langle c, q \rangle}$.

Similarly to Step 5 in the proof of Theorem 4.2.3, before setting to zero the drift of $V_2(q)$ we need to prove that the MGF of $e^{c\langle c^l, \overline{T}^v \rangle}$ exists in an interval around 0. The proof is presented in Appendix C.3.

Now we set to zero the drift of $V_2(q)$. Note that $P \left[ \langle c^l, \pi^v \rangle - \overline{T}^v \neq 0 \right] > 0 \text{ whenever } \epsilon > 0$. If we expand the product in the expression of Lemma 5.3.2 we obtain
\[
E \left[ \left( e^{\theta c(c^l,\pi^v)} - 1 \right) \left( e^{-\theta c(c^l,\pi^v)} - 1 \right) \right] = E \left[ e^{\theta c(c^l,\pi^v) + \pi^v - \pi^v} - e^{\theta c(c^l,\pi^v)} + e^{-\theta c(c^l,\pi^v)} + 1 \right] + o(\epsilon^2)
\]
\[
\overset{(a)}{=} E \left[ e^{\theta c(c^l,\pi^v) + \pi^v} - e^{\theta c(c^l,\pi^v)} + e^{-\theta c(c^l,\pi^v)} + 1 \right] + o(\epsilon^2)
\]
\[
\overset{(b)}{=} E \left[ e^{\theta c(c^l,\pi^v)} \right] E \left[ e^{\theta c(c^l,\pi^v) - \overline{T}^v} - 1 \right] + 1 - E \left[ e^{-\theta c(c^l,\pi^v)} \right] + o(\epsilon^2)
\]
\[
\overset{(c)}{=} E \left[ e^{\theta c(c^l,\pi^v)} \right] E \left[ e^{\theta c(c^l,\pi^v) - \overline{T}^v} - 1 \right] + 1 - E \left[ e^{-\theta c(c^l,\pi^v)} \right] + o(\epsilon^2)
\]
where (a) holds by adding and subtracting \( E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)}) \pm \overline{e}^{(i)} B} \right] \) and reorganizing terms; (b) holds because \( \overline{e}^{(i)} \) is independent of \( \overline{e}^{(i)} \), \( \overline{f}^{(i)} \) and \( B \) and by Lemma C.3.1; (c) holds by Lemma 5.3.3 after reorganizing terms and because \( \overline{e}^{(i)} \) is bounded.

Reorganizing terms we obtain

\[
E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)})} \right] = \frac{1 - E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} \right] + E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)})} \right] E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} - e^{-\theta e B} \right] + o(\epsilon^2)}{1 - E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)}) - \epsilon B} \right]}. \tag{19}
\]

We take Taylor expansion of the terms on the right hand side of Equation (19). We use Lemma 3.3.2. For the first term in the numerator, we have

\[
1 - E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} \right] = 1 - E \left[ f_{c_i, -e^{(i)} \overline{e}^{(i)}}(\theta) \right] = \theta e \left[ (c^{(i)}, \overline{e}^{(i)}) \right] - \frac{(\theta e)^2}{2} E \left[ (e^{(i)}, \overline{e}^{(i)})^2 \right] + O(\epsilon^3). \tag{20}
\]

**Claim 5.3.4.** Consider a generalized switch as described in Theorem 5.2.3. Then,

\[
\frac{(\theta e)^2}{2} E \left[ (e^{(i)}, \overline{e}^{(i)})^2 \right] \text{ is } O(\epsilon^3)
\]

We present a proof of Claim 5.3.4 in Appendix D.2.

Then, using Claim 5.3.4 in Equation (20) we obtain

\[
1 - E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} \right] = \theta e \left[ (c^{(i)}, \overline{e}^{(i)}) \right] + O(\epsilon^3). \tag{21}
\]

For the second term in the numerator, we have

\[
E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)})} \right] E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)}) - \epsilon B} \right] = \frac{1 + \theta e^2 + O(\epsilon^3)}{1 - E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)}) - \epsilon B} \right]}
\]

\[
= E \left[ f_{c_i, (c^{(i)}, \overline{e}^{(i)})}(\theta) \right] E \left[ f_{c_i, (c^{(i)} - c^{(i)}, \overline{e}^{(i)} - \epsilon B)}(\theta) - 1 \right] \tag{22}
\]

**Claim 5.3.5.** Consider a generalized switch as described in Theorem 5.2.3 and the notation introduced in Lemma 3.3.2. Then,

\[
E \left[ f_{c_i, (c^{(i)}, \overline{e}^{(i)} - \epsilon B)}(\theta) \right] = 1 + \theta e^2 + O(\epsilon^3)
\]

and \( E \left[ f_{c_i, (c^{(i)} - c^{(i)}, \overline{e}^{(i)} - \epsilon B)}(\theta) - 1 \right] = \theta e \left[ \overline{e}^{(i)} - (c^{(i)}, \overline{e}^{(i)}) \right] + O(\epsilon^3) \)

We prove the claim in Appendix D.3. Using Claim 5.3.5 in Equation (22), reorganizing terms and using that \( B \) and \( \overline{e}_i \) are bounded for all \( i \in \{1, \ldots, n\} \), we obtain

\[
E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)})} \right] E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)}) - \epsilon B} \right] = \theta e \left[ B - (c^{(i)}, \overline{e}^{(i)}) \right] + O(\epsilon^3) \tag{23}
\]

Then, the numerator of Equation (19) yields

\[
1 - E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} \right] + E \left[ e^{\theta e(c^{(i)}, \overline{e}^{(i)})} \right] E \left[ e^{-\theta e(c^{(i)}, \overline{e}^{(i)})} - e^{-\theta e B} \right] + o(\epsilon^2) = \left( \theta e \left[ (c^{(i)}, \overline{e}^{(i)}) \right] + \theta e \left[ B - (c^{(i)}, \overline{e}^{(i)}) \right] + O(\epsilon^3) \right) + O(\epsilon^2)
\]

\[
= (a) \theta e \left( E \left[ (c^{(i)}, \overline{e}^{(i)}) + B - (c^{(i)}, \overline{e}^{(i)}) \right] \right) + O(\epsilon^2)
\]

\[
(b) \theta e^2 + O(\epsilon^2), \tag{24}
\]

21
where (a) holds because $O(\varepsilon^3)$ is $o(\varepsilon^2)$; and (b) holds by Lemma 5.3.1.

For the denominator, we obtain

$$1 - E \left[ e^{-\theta \epsilon (\mathbf{B} - \langle c, \mathbf{\bar{a}} \rangle)} \right] = 1 - E \left[ f_{\epsilon,\langle c, \mathbf{\bar{a}} \rangle}(\theta) \right]$$

$$= - \theta \epsilon E \left[ \langle c, \mathbf{\bar{a}} \rangle - \mathbf{B} \right] - \frac{(\theta \epsilon)^2}{2} E \left[ (\mathbf{B} - \langle c, \mathbf{\bar{a}} \rangle)^2 \right] + O(\varepsilon^3)$$

(a) holds by Equation (18) and expanding the product; (b) holds by definition of variance and covariance and because $\mathbf{a}$ and $\mathbf{\bar{a}}$ are independent; and (c) holds by Equation (18).

Using Equations (24) and (25) in (19) we obtain

$$E \left[ e^{\theta \epsilon (c^{(i)} \mathbf{\bar{a}})} \right] = \frac{\theta \epsilon^2 + o(\varepsilon^2)}{\theta \epsilon^2 - \frac{(\theta \epsilon)^2}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} \left[ \mathbf{\nu}_i^{(c)}, \mathbf{\nu}_j^{(c)} \right] + \sigma^2_{B_t} + \varepsilon^2 \right) + O(\varepsilon^3)}$$

$$= \frac{1}{1 - \frac{\theta}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} \left[ \mathbf{\varpi}_i^{(c)}, \mathbf{\varpi}_j^{(c)} \right] + \sigma^2_{B_t} + \varepsilon^2 \right) + O(\varepsilon)}$$

Then, taking the heavy-traffic limit yields

$$\lim_{\varepsilon \downarrow 0} E \left[ e^{\theta \epsilon (c^{(i)} \mathbf{\bar{a}})} \right] = \frac{1}{1 - \theta \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} \left[ \mathbf{\varpi}_i, \mathbf{\varpi}_j \right] + \sigma^2_{B_t} \right)}$$

which is the MGF of an exponential random variable with mean $\frac{1}{\theta} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} \left[ \mathbf{\varpi}_i, \mathbf{\varpi}_j \right] + \sigma^2_{B_t} \right)$. This implies that $q^{(c)} = \langle c^{(c)}, \mathbf{\bar{q}}^{(c)} \rangle^{c^{(c)}} \Rightarrow \mathbf{\Upsilon} c^{(t)}$, where $\mathbf{\Upsilon}$ is an exponential random variable with mean $\frac{1}{\theta} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov} \left[ \mathbf{\varpi}_i, \mathbf{\varpi}_j \right] + \sigma^2_{B_t} \right)$.

Then, we conclude that $\mathbf{\bar{q}}^{(c)} = \mathbf{\bar{q}}^{(c)}_{\parallel} + \mathbf{\bar{q}}^{(c)}_{\perp}$ converges in distribution to $\mathbf{\Upsilon} c^{(t)}$ as $\epsilon \downarrow 0$. This proves Theorem 5.2.3.

6 Future work

The current paper develops the MGF method, which we believe can be used to study more general set of queueing systems. We outline a few 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 it is sufficient to assume that the first two moments of the arrival and service processes 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, since they always exist for non-negative random variables. The main challenge in this approach is to establish the SSC under unbounded arrivals and service sequences. In the current paper, we used the SSC established by [11], which is based on the results from [20], 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 [4] for Markov Decision Processes. The main idea of their method is to take second order Taylor expansion of the value function in order to solve the Bellman’s equations. We believe this can give us insight to 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 [11, 42, 41]. These bounds give the rate of convergence to the heavy-traffic limit. Since 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 its 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 when the CRP condition is not met. We believe that the MGF method is a promising approach to obtain 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 because of the following reason. The queue lengths process is a multi-dimensional 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 Foster-Lyapunov Theorem. This has an added benefit that one typically obtains as a consequence a (possibly loose) upper bound on an expression of them form $E[\epsilon \sum_{i} \eta_i]$. If $P$ is the transition matrix, then the stationary distribution is a unique solution of the equation, $\pi = P\pi$. 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 two-sided Laplace transform of the equation $\pi = P\pi$ we obtain an equation, which is same as the one we obtain by setting the drift of the exponential test function to zero. Since 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 Equation (5)). Therefore, using the MGF approach, we seek to solve it in the heavy-traffic limit. To do this, one first needs to prove tightness of the sequence of the stationary distributions as the heavy-traffic parameter $\epsilon$ goes to zero. Tightness follows directly from the bound on $E[\epsilon \sum_{i} \eta_i]$ that one obtained 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 since 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 when this equation may not be solved explicitly, one may be able to obtain moments from this equation. For instance, one may be able to obtain the moment bounds computed by [42], [41] and [61] 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 crossbar switch operating under MaxWeight. The bandwidth-sharing network [47] 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 product of exponentials [29, 66]. 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 [58]. However, the interchange
of limits step was not shown by [58], so their result does not show if 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 [61]. 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 \( n^2 \). [42] studied an input-queued cross-bar switch operating under MaxWeight and proved that SSC occurs onto 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 [28].

7 Conclusion

In this paper we introduced the MGF method to compute the steady-state distribution of the scaled queue lengths in heavy-traffic. We motivated the method with a single server queue and we applied it in queueing 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 the paper illustrates how the unused service is handled in two different types of queueing systems. Further developing the MGF method to study system when the CRP condition is not satisfied such as the bandwidth sharing network and the input-queued switch forms future work.
Appendix

A Proof of Lemma 3.3.2

Proof. (of Lemma 3.3.2)

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

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

where $\hat{\theta}$ is a real number between 0 and $\theta$. Then, for all $0 \leq x \leq K$ we have

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

Since $\hat{\theta}$ is between 0 and $\theta$, and $|\theta| \leq \Theta$ we have

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

which is finite for every $\epsilon$. Then,

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

Therefore,

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

where $C_1 = \frac{\Theta^3 K^3}{3!}$ is a finite constant.

Now, since $X^{(\epsilon)}(\epsilon) \leq K_{\max}$ with probability 1, we have

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

which proves the lemma.

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, that we use in the proof of Theorem 4.2.3.

B.1 Proof of SSC in the load balancing system operating under JSQ

In this section we present an insight of the proof of SSC as developed in [11]. 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.1.1. Consider a load balancing system as described in Corollary 4.2.5. Then, for each $m = 1, 2, \ldots$ there exists a finite constant $M_m$ such that

$$E \left[ \| \mathbf{q}^{(\epsilon)} \|^m \right] \leq M_m.$$
This proof is based on a lemma that was first proved by [20]. The original statement is more general than what we need here, so we present the specific result that we will use, as stated by [11].

**Lemma B.1.2.** For an irreducible and aperiodic Markov Chain \(\{X(k) : k \geq 1\}\) over a countable state space \(\mathcal{X}\), suppose \(Z : \mathcal{X} \to \mathbb{R}_+\) is a non-negative valued Lyapunov function. We define the drift of \(Z\) at \(x\) as

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

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:

(C1) There exists \(\eta > 0\) and \(\kappa < \infty\) such that

\[E[\Delta Z(x) | X(k) = x] \leq -\eta\quad\text{for all } x \in \mathcal{X} \text{ with } Z(x) \geq \kappa\]

(C2) There exists \(D < \infty\) such that

\[|\Delta Z(x)| \leq D\quad\text{with probability 1 for all } x \in \mathcal{X}\]

Then, there exist \(\theta^* > 0\) and \(C^* < \infty\) such that

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

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 \(\bar{Z}\) for which

\[E\left[e^{\theta \bar{Z}}\right] \leq C^*
\]

**Proof.** (of Proposition B.1.1) [11] use the Lyapunov function \(Z(q) = \|q^{(\epsilon)}_\perp\|\) and they prove that

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

where \(\delta\) is a fixed constant in \((0, \mu_{\min})\). This verifies condition (C1) of Lemma B.1.2.

To verify condition (C2), they prove that for all \(q \in \mathbb{R}^n_+\)

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

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

Also, for \(\epsilon > 0\) the Markov Chain \(\{q(k) : k \geq 1\}\) is positive recurrent. Therefore, by Lemma B.1.2 there exists \(\theta^* > 0\) and \(C^* > 0\) such that

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

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

\[E\left[\|q^{(\epsilon)}_\perp\|^m\right] \leq M_m\]

\(\square\)
B.2 Proof of Lemma 4.3.2

To prove Lemma 4.3.2 we use the following result.

**Lemma B.2.1.** Consider the load balancing system indexed by $\epsilon$ described in Theorem 4.2.3. Then, for any $\alpha \in \mathbb{R}$ and for all $k \geq 1$ we have

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

where $q_{\perp}^{(c)}(k)$ is the $i$th element of $q_{\perp}^{(c)}(k)$, for each $i \in \{1, \ldots, n\}$.

**Proof.** (of Lemma B.2.1)

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

\[
u_i(k)(e^{-\alpha q_{\perp}^{(C)}(k+1)} - 1) = 0 \quad \forall i \in \{1, \ldots, n\}.
\]

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

\[
\sum_{i=1}^{n} \nu_i(k) \left( e^{-\alpha q_{\perp}^{(C)}(k+1)} - 1 \right) = 0.
\]

By definition of $q_{\parallel}(k)$ and $q_{\perp}(k)$ we have $q(k) = q_{\parallel}(k) + q_{\perp}(k)$, so

\[
\sum_{i=1}^{n} \nu_i(k)(e^{-\alpha (q_{\parallel}(k+1)) - 1}) = 0.
\]

But $q_{\parallel}(k+1) = \left( \frac{1}{n} \sum_{j=1}^{n} q_j(k+1) \right) 1$ so $q_{\parallel}(k+1) = q_{\parallel}(k+1)$ for all $i \in \{1, \ldots, n\}$. Then, reorganizing terms we obtain

\[
\sum_{i=1}^{n} \nu_i(k)e^{-\alpha q_{\perp}^{(C)}(k+1)} = e^{\alpha q_{\parallel}(k+1)} \sum_{i=1}^{n} \nu_i(k).
\]

By definition of $q_{\parallel}(k)$ we obtain

\[
\sum_{i=1}^{n} \nu_i(k)e^{-\alpha q_{\perp}^{(C)}(k+1)} = e^{\alpha \sum_{j=1}^{n} q_j(k+1)} \sum_{i=1}^{n} \nu_i(k).
\]

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

\[
\sum_{i=1}^{n} \nu_i(k) \left( e^{\alpha \sum_{j=1}^{n} q_j(k+1)} - 1 \right) = \sum_{i=1}^{n} \nu_i(k) \left( e^{-\alpha q_{\perp}^{(C)}(k+1)} - 1 \right).
\]

In the proof of Lemma 4.3.2 we use Lemma B.2.1 and the following facts:

(i) The function $g(x) = \frac{e^x - 1}{x}$ is non-negative and non-decreasing for all $x \in \mathbb{R}$

(ii) Suppose $0 \leq x \leq y$. Then, for all $\theta \in \mathbb{R}$ we have $e^{\theta y} - 1 \leq (\theta x) \left( \frac{e^{\theta y} - 1}{\theta y} \right)$
(iii) For all \( x \in \mathbb{R}_+ \), \( \frac{e^x - 1}{x} < e^x \)

All these facts can be shown using just calculus techniques, so we omit the proof. Now we prove Lemma 4.3.2.

**Proof. (of Lemma 4.3.2)**

By properties of expectation and absolute value, we have

\[
\begin{align*}
|E\left[\left(e^{\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\left(e^{-\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\right]| \\
\leq E\left[\left|\sum_{i=1}^{n} \bar{t}_i\right|^2 \left|\left(e^{\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\left(e^{-\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\right|\right] \\
= |\theta| \epsilon \left|\sum_{i=1}^{n} \bar{t}_i\right|^2 E\left[\left(e^{\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\left(e^{-\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\right] \\
\leq |\theta| \epsilon \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) E\left[\left(e^{\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\left(e^{-\theta \sum_{i=1}^{n} \bar{t}_i} - 1\right)\right] \\
\leq |\theta| \epsilon \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) \sqrt{E\left[\sum_{i=1}^{n} (\bar{t}_i)^2\right]} \cdot \sqrt{E\left[\sum_{i=1}^{n} \left|\theta \epsilon \bar{t}_i\right|^2\right]} \\
= \theta^2 \epsilon^3 \sqrt{S_{\text{max}}} \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) \sqrt{\sum_{i=1}^{n} \left|\theta \epsilon \bar{t}_i\right|^2} \\
= \theta^2 \epsilon^3 \sqrt{S_{\text{max}}} \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) \sqrt{\sum_{i=1}^{n} \left|\theta \epsilon \bar{t}_i\right|^2} \\
\leq \theta^2 \epsilon^3 \sqrt{S_{\text{max}}} \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) \sqrt{M_4} \\
\leq \theta^2 \epsilon^3 \sqrt{S_{\text{max}}} \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) \sqrt{E\left[\sum_{i=1}^{n} \left|\theta \epsilon \bar{t}_i\right|^2\right]} \\
\leq \theta^2 \epsilon^3 \sqrt{S_{\text{max}}} \left(e^{\theta |\epsilon| n S_{\text{max}} - 1} - 1\right) M_4,
\end{align*}
\]

where (a) holds by the fact (i) stated above and because \( \bar{t}_i \leq S_{\text{max}} \) for all \( i \in \{1, \ldots, n\} \), (b) holds by triangle inequality and Lemma B.2.1, (c) holds by Cauchy-Schwarz inequality, (d) holds because \( \bar{t}_i \leq S_{\text{max}} \) for all \( i \in \{1, \ldots, n\} \), \( \sum_{i=1}^{n} E [\bar{t}_i] = \epsilon \) and \( x \) is an increasing function.

By L'Hospital’s rule we have

\[
\lim_{\epsilon \downarrow 0} \frac{e^{\theta |\epsilon| n S_{\text{max}} - 1}}{|\theta| n S_{\text{max}}} = 1
\]

On the other hand, by Cauchy-Schwarz inequality, for each \( i \in \{1, \ldots, n\} \)

\[
E\left[\left(e^{-\theta \bar{t}_i} - 1 \right)^2 \left(\bar{t}_i\right)^2\right] \leq \left( E\left[\left(e^{-\theta \bar{t}_i} - 1 \right)^4 \right]\right)^{1/2} \left( E\left[\left(\bar{t}_i\right)^4\right]\right)^{1/2} \\
\leq \left( E\left[\left(e^{-\theta \bar{t}_i} - 1 \right)^4 \right]\right)^{1/2} \left(M_4\right)^{1/2},
\]

where the last inequality holds by SSC as assumed in Theorem 4.2.3 and because for all \( i \in \{1, \ldots, n\} \), \( (\bar{t}_i)^4 \leq ||\bar{t}_i||^4 \).

On the other hand, observe

\[
-\theta \epsilon ||\bar{t}_i|| \leq |\theta| |\epsilon| ||\bar{t}_i|| \leq |\theta| |\epsilon| ||\bar{t}_i||
\]

28
and $|\theta|\|\mathbf{q}_\perp\| \geq 0$. Then, by the facts (i) and (iii) stated above we obtain

\[
0 \leq e^{-\theta e\|\mathbf{q}_\perp\|} - 1 \leq \frac{e^{\|\theta\|\|\mathbf{q}_\perp\|} - 1}{\|\theta\|\|\mathbf{q}_\perp\|} \leq e^{\|\theta\|\|\mathbf{q}_\perp\|}
\]

Also, by SSC we have

\[
0 \leq \lim_{\epsilon \downarrow 0} E \left[ \epsilon\|\mathbf{q}_\perp\| \right] \leq \lim_{\epsilon \downarrow 0} \epsilon^2 M_2 = 0
\]

Then, $\|\mathbf{q}_\perp\|$ converges to zero in the mean-square sense and, therefore, in distribution. Therefore, the MGF of $\|\mathbf{q}_\perp\|$ converges to the MGF of 0 (Theorem 9.5 in Section 5, [19]), i.e.

\[
\lim_{\epsilon \downarrow 0} E \left[ e^{\epsilon\|\mathbf{q}_\perp\|} \right] = 1
\]

Therefore,

\[
E \left[ \left( e^{\theta \sum_{i=1}^{n} \overline{q}_i^+} - 1 \right) \left( e^{-\theta \epsilon \sum_{i=1}^{n} \overline{q}_i} - 1 \right) \right] \text{ is } o(\epsilon^2)
\]

\[
\square
\]

### B.3 Existence of MGF of $\epsilon(c, \mathbf{q}^{(\epsilon)})$ in the load balancing system

First, we state the result formally.

**Lemma B.3.1.** Consider a load balancing system parametrized by $\epsilon$ as described in Theorem 4.2.3. Then, for each $\epsilon \in (0, \mu_\Sigma)$ there exists $\Theta > 0$ such that $E \left[ e^{\theta \epsilon \sum_{i=1}^{n} \overline{q}_i} \right] < \infty$ for all $\theta \in [-\Theta, \Theta]$.

**Proof. (of Lemma B.3.1)**

First observe that if $\theta = 0$, then $E \left[ e^{\theta \epsilon \sum_{i=1}^{n} q_i} \right] = 1$, so it is finite. Now assume $\theta \neq 0$. We use Foster-Lyapunov theorem (Proposition 6.13 in [21]) with Lyapunov function $V(q) = e^{\theta \epsilon \sum_{i=1}^{n} q_i}$. In this proof we use the following notation $E_q \left[ \cdot \right] = E \left[ \cdot \mid q(k) = q \right]$.

The drift of $V(q(k))$ conditioned on $q(k) = q$ is

\[
E_q \left[ V(q(k+1)) - V(q(k)) \right] = E_q \left[ e^{\theta \epsilon \sum_{i=1}^{n} q_i(k+1)} - e^{\theta \epsilon \sum_{i=1}^{n} q_i(k)} \right]
\]

\[
= E_q \left[ e^{-\theta \epsilon \sum_{i=1}^{n} u_i(k)} + e^{\theta \epsilon \sum_{i=1}^{n} (q_i(k)+a_i(k)-s_i(k))} \right] - E_q \left[ e^{\theta \epsilon \sum_{i=1}^{n} q_i(k)} \right] + o(\epsilon^2)
\]

\[
= E_q \left[ 1 - e^{-\theta \epsilon \sum_{i=1}^{n} u_i(k)} \right] + e^{\theta \epsilon \sum_{i=1}^{n} q_i(k)} E \left[ e^{\theta \epsilon (a(k) - \sum_{i=1}^{n} s_i(k))} - 1 \right] + o(\epsilon^2),
\]

where (a) holds by Lemma 4.3.2, after expanding the product and reorganizing terms; and (b) holds because $a(k) = \sum_{i=1}^{n} u_i(k)$ and $\sum_{i=1}^{n} s_i(k)$ are independent of $q(k)$.

Observe that the term $o(\epsilon^2)$ represents a constant that does not depend on the queue lengths, and $1 - E_q \left[ e^{-\theta \epsilon \sum_{i=1}^{n} u_i(k)} \right]$ is bounded above with probability 1 because, by definition of unused service, we have

\[
0 \leq \sum_{i=1}^{n} u_i(k) \leq \sum_{i=1}^{n} s_i(k) \leq nS_{\text{max}}.
\]

Therefore, it suffices to show the existence of $\Theta > 0$ and $\delta > 0$ such that

\[
E \left[ e^{\theta \epsilon (a(k) - \sum_{i=1}^{n} s_i(k))} \right] - 1 < -\delta \quad \forall \theta \in [-\Theta, 0) \cup (0, \Theta]
\] (26)
Let $M(\theta) = E \left[ e^{\theta (a(k) - \sum_{i=1}^{n} s_i(k))} \right]$. We take Taylor approximation of $M(\theta)$ around $\theta = 0$. In this argument we need to interchange derivative and expected value. Notice $e^{\theta (a(k) - \sum_{i=1}^{n} s_i(k))}$ is a non-negative measurable function and its derivative with respect to $\theta$ exists. Also, $e^{\theta (a(k) - \sum_{i=1}^{n} s_i(k))} \leq e^{\theta |A_{\max} + n S_{\max}|}$ with probability 1 and the upper bound is an integrable function. Therefore, by Dominated Convergence Theorem, we can interchange derivative (limit) and expectation (integral). We obtain

$$M(\theta) = M(0) + \theta M'(\zeta) = 1 + \theta M'(\zeta),$$

where $\zeta$ is a real number between 0 and $\theta$. But the derivative of $M(\theta)$ at $\theta = 0$ is

$$M'(0) = \epsilon E \left[ a(k) - \sum_{i=1}^{n} s_i(k) \right] e^{\theta a(k) - \sum_{i=1}^{n} s_i(k)} \bigg|_{\theta=0}$$

$$= \epsilon E \left[ a(k) - \sum_{i=1}^{n} s_i(k) \right]$$

$$= \epsilon (\mu \Sigma - \epsilon - \mu \Sigma)$$

$$= -\epsilon^2.$$

Since $M'(\theta)$ is continuous (page 78 in [52]), there exists $\Theta > 0$ such that

$$M'(0) = -\epsilon^2 \leq M'(\theta) < -\frac{\epsilon^2}{2} \quad \forall \theta \in [-\Theta, \Theta].$$

Therefore, we can take $\delta = \frac{\epsilon^2}{2}$ in Equation (26), which proves the lemma.

\[ \square \]

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 5.2.3.

C.1 Proof of Lemma 5.3.2

To prove Lemma 5.3.2 we use the following lemma, which is similar to Lemma B.2.1.

**Lemma C.1.1.** Consider a generalized switch parametrized by $\epsilon$, as described in Theorem 5.2.3. Then, for any $\alpha \in \mathbb{R}$ and for all $k \geq 1$ we have

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

**Proof.** (of Lemma C.1.1)

First observe that if $\alpha = 0$ the lemma trivially holds. Now we prove the lemma for $\alpha \neq 0$.

From Equation (3) we know that $q_{i}(k+1)u_{i}(k) = 0$ for all $i \in \{1, \ldots, n\}$. Then, for all $\beta \in \mathbb{R}$ we have

$$u_{i} \left( e^{-\beta q_{i}(k+1)} - 1 \right) = 0 \quad \forall i \in \{1, \ldots, n\},$$

and this equation implies

$$c_{i}^{(\epsilon)} u_{i} \left( e^{-\beta q_{i}(k+1)} - 1 \right) = 0 \quad \forall i \in \{1, \ldots, n\}.$$
Without loss of generality, we assume $c_i^{(t)} > 0$ for all $i \in \{1, \ldots, n\}$ because otherwise the last equation holds trivially. Let $\alpha \in \mathbb{R}$ and for each $i \in \{1, \ldots, n\}$ let $\alpha_i \in \mathbb{R}$ be such that $\alpha = \alpha_i c_i$ for all $i \in \{1, \ldots, n\}$. Then,

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

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

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

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

$$\overset{(a)}{=} \sum_{i=1}^{n} c_i u_i(k) \left( e^{-\alpha (c_i^{(t)} q(k+1)) c_i - \alpha S_{q_i^{(k+1)}}} - 1 \right)$$

$$\overset{(b)}{=} \sum_{i=1}^{n} c_i^{(t)} u_i(k) \left( e^{-\alpha (c_i^{(t)} q(k+1)) + \frac{\alpha}{c_i} S_{q_i^{(k+1)}}} - 1 \right)$$

$$\overset{(c)}{=} e^{-\alpha (c_i^{(t)} q(k+1))} \sum_{i=1}^{n} c_i^{(t)} u_i(k) e^{-\frac{\alpha}{c_i} S_{q_i^{(k+1)}}} - \langle c^{(t)}, u(k) \rangle$$

where (a) holds by definition of $q_i^{(k)}$; (b) holds by definition of $\alpha$; and (c) holds by expanding the product and reorganizing terms. Therefore, we have

$$\langle c^{(t)}, u(k) \rangle = e^{-\alpha (c_i^{(t)} q(k+1))} \sum_{i=1}^{n} c_i^{(t)} u_i(k) e^{-\frac{\alpha}{c_i} S_{q_i^{(k+1)}}}.$$

Multiplying both sides by $e^{\alpha (c_i^{(t)} q(k+1))}$ we obtain

$$\langle c^{(t)}, u(k) \rangle e^{\alpha (c_i^{(t)} q(k+1))} = \sum_{i=1}^{n} c_i u_i(k) e^{-\frac{\alpha}{c_i} S_{q_i^{(k+1)}}},$$

which proves the lemma.

Now we prove Lemma 5.3.2.

Proof. (of Lemma 5.3.2)

Since $c^{(t)} \geq 0$ and $\underline{w} \leq \underline{x} \leq S_{\max}$ for all $i \in \{1, \ldots, n\}$, we have

$$0 \leq \langle c^{(t)}, \underline{w} \rangle \leq \langle c^{(t)}, S_{\max} \underline{1} \rangle.$$

Then, from facts (i) and (ii) stated in Appendix B.2 we have

$$\left| e^{-\theta \epsilon \langle c^{(t)}, \underline{w} \rangle} \right| \leq \left| \theta \epsilon \langle c^{(t)}, \underline{w} \rangle \right| \left( e^{-\theta \epsilon S_{\max} \langle c^{(t)}, \underline{1} \rangle} - 1 \right).$$

(27)
Now, by properties of expected value, we have

\[
| E \left[ \left( e^{\theta e(c^{(j)}\hat{\mathbf{x}}) - 1} \right) \left( e^{-\theta e(c^{(j)}\bar{\mathbf{x}}) - 1} \right) \right] \\
\leq E \left[ \left( e^{\theta e(c^{(j)}\hat{\mathbf{x}}) - 1} \right) \left( e^{-\theta e(c^{(j)}\bar{\mathbf{x}}) - 1} \right) \right] \\
\overset{(a)}{\leq} \left| \theta e \right| \left( \frac{e^{-\theta e S_{\max}(c^{(j)}, 1) - 1}}{-\theta e S_{\max}(c^{(j)}, 1)} \right) E \left[ \left( e^{\theta e(c^{(j)}\hat{\mathbf{x}}) - 1} \right) \right] \\
\overset{(b)}{=} \left| \theta e \right| \left( \frac{e^{-\theta e S_{\max}(c^{(j)}, 1) - 1}}{-\theta e S_{\max}(c^{(j)}, 1)} \right) E \left[ \sum_{i=1}^{n} c_i^{(j)} u_i \left( -e^{\frac{\theta e}{c_i^{(j)}}} \bar{\mathbf{q}}_{i}^{+} \right) \right] \\
\overset{(c)}{\leq} \left| \theta e \right| \left( \frac{e^{-\theta e S_{\max}(c^{(j)}, 1) - 1}}{-\theta e S_{\max}(c^{(j)}, 1)} \right) E \left[ \sum_{i=1}^{n} c_i^{(j)} u_i \left( -e^{\frac{\theta e}{c_i^{(j)}}} \bar{\mathbf{q}}_{i}^{+} - 1 \right) \right] \\
\overset{(d)}{\leq} \left| \theta e \right| \left( \frac{e^{-\theta e S_{\max}(c^{(j)}, 1) - 1}}{-\theta e S_{\max}(c^{(j)}, 1)} \right) \sqrt{E \left[ \sum_{i=1}^{n} \left( c_i^{(j)} u_i \right)^2 \right] E \left[ \sum_{i=1}^{n} \left( -e^{\frac{\theta e}{c_i^{(j)}}} \bar{\mathbf{q}}_{i}^{+} - 1 \right)^2 \right]}
\]

where (a) holds by Equation (27); (b) holds by Lemma C.1.1 with \( \alpha = \theta e \); (c) holds by triangle inequality; and (d) holds by Cauchy-Schwarz inequality.

But

\[
E \left[ \sum_{i=1}^{n} \left( c_i^{(j)} u_i \right)^2 \right] \leq c_{\max} S_{\max} E \left[ \sum_{i=1}^{n} c_i^{(j)} u_i \right] \leq c_{\max} S_{\max} e
\]

where \( c_{\max} = \max_i c_i^{(j)} \) and the last equality holds because, by Lemma 5.3.1 we have

\[
E \left[ (c^{(j)}, u) \right] = e - E \left[ B \right] + E \left[ (c^{(j)}, \bar{\mathbf{x}}) \right],
\]

and by definition of the capacity region in (13) and because \( \bar{\mathbf{x}} \) depends on the channel state, we have that \( E \left[ (c^{(j)}, \bar{\mathbf{x}}) \right] \in \mathcal{C} \). Then, since \( E \left[ B \right] = b^{(j)} \) we have

\[
-E \left[ B \right] + E \left[ (c^{(j)}, \bar{\mathbf{x}}) \right] \leq 0.
\]

Therefore,

\[
\left| E \left[ \left( e^{\theta e(c^{(j)}\hat{\mathbf{x}}) - 1} \right) \left( e^{-\theta e(c^{(j)}\bar{\mathbf{x}}) - 1} \right) \right] \right| \\
\leq \left| \theta e \right| c_{\max} S_{\max} \left( \frac{e^{-\theta e S_{\max}(c^{(j)}, 1) - 1}}{-\theta e S_{\max}(c^{(j)}, 1)} \right) \sqrt{E \left[ \sum_{i=1}^{n} \left( -e^{\frac{\theta e}{c_i^{(j)}}} \bar{\mathbf{q}}_{i}^{+} - 1 \right)^2 \right]}.
\]

The rest of the argument is equivalent to the last steps in the proof of Lemma 4.3.2, so we omit it. \( \Box \)

### C.2 Proof of Lemma 5.3.3

**Proof.** (of Lemma 5.3.3)

32
In this proof we use geometric properties of SSC. We start with a definition. For each \( t \in T \), suppose that

the channel state is \( \mathbf{T} = t \). Then, let \( \nu(t) \in (0, \frac{\pi}{2}] \) be an angle such that \( \langle c(l), \mathbf{t} \rangle = b(t, l) \) if \( \| \mathbf{t} \| \geq \cos(\nu(t)) \).

Let \( \nu_{\parallel} \) be the angle between \( \parallel \mathbf{t} \parallel \) and \( \mathbf{t} \) and define \( \nu_{\min} \triangleq \min_{t \in T} \nu(t) \). Therefore,

\[
\mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \iff \nu(\mathbf{T}) < \nu_{\parallel}.
\]

Then,

\[
E \left[ \left( e^{\theta \langle c(l), \mathbf{t} \rangle} - 1 \right) \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right) \right]
\]

\[
= E \left[ \left( e^{\theta \| \mathbf{t} \|} - 1 \right) \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right) \mathbb{I}_{\{\mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \}} \right]
\]

\[
= E \left[ \left( e^{\theta \| \mathbf{t} \|} - 1 \right) \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right) \mathbb{I}_{\{\nu(\mathbf{T}) > \nu_{\parallel} \}} \right]
\]

\[
\leq E \left[ \left( e^{\theta \| \mathbf{t} \| \cot(\nu_{\parallel})} - 1 \right) \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right) \mathbb{I}_{\{\nu(\mathbf{T}) > \nu_{\parallel} \}} \right]
\]

\[
\leq E \left[ \left( e^{\theta \| \mathbf{t} \| \cot(\nu_{\min})} - 1 \right) \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right) \mathbb{I}_{\{\nu(\mathbf{T}) > \nu_{\parallel} \}} \right]
\]

\[
\leq E \left[ \left( e^{\theta \| \mathbf{t} \| \cot(\nu_{\min})} - 1 \right)^2 \right] E \left[ \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right)^2 \right]
\]

where \( (a) \) and \( (b) \) hold because \( \cot(\nu) \) is decreasing for \( \nu \in (0, \frac{\pi}{2}] \); \( (c) \) holds because if \( \nu_{\parallel} \leq \nu(\mathbf{T}) \), then \( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 = 0 \); and \( (d) \) holds by Cauchy-Schwarz inequality.

Using an argument similar to the one at the end of Lemma 4.3.2, it can be proved that

\[
0 \leq E \left[ \left( e^{\theta \| \mathbf{t} \| \cot(\nu_{\min})} - 1 \right)^2 \right] \leq f(\epsilon) \epsilon^2,
\]

where \( f(\epsilon) \) converges to a constant as \( \epsilon \downarrow 0 \).

On the other hand,

\[
E \left[ \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right)^2 \right] = E \left[ \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right)^2 \mathbb{I}_{\{\mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \}} \right]
\]

\[
= E \left[ \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right)^2 \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} \right)^2 \mathbb{I}_{\{\mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \}} \right]
\]

\[
\leq \left( \frac{e^{\theta \langle \mathbf{B}_{\max} - \langle c(l), \mathbf{t} \rangle \rangle} - 1}{\theta e \langle \mathbf{B}_{\max} - \langle c(l), \mathbf{t} \rangle \rangle} \right)^2 \cdot \left( \frac{\mathbf{B} - \langle c(l), \mathbf{t} \rangle}{\theta e \langle \mathbf{B}_{\max} - \langle c(l), \mathbf{t} \rangle \rangle} \right)^2 P \left[ \mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \right]
\]

where \( \mathbf{B}_{\max} = \max_{t \in T} b(t, l) \). \[11\] prove that \( P \left[ \mathbf{B} \neq \langle c(l), \mathbf{t} \rangle \right] = K \epsilon \) for a finite constant \( K \), and their proof only depends on the geometry of SSC, so it also holds here. Therefore,

\[
0 \leq E \left[ \left( e^{\theta \langle \mathbf{B} - \langle c(l), \mathbf{t} \rangle \rangle} - 1 \right)^2 \right] \leq \hat{f}(\epsilon) \sqrt{e_{\parallel}},
\]

where \( \hat{f}(\epsilon) \) converges to a constant as \( \epsilon \downarrow 0 \).

This completes the proof.
C.3 Existence of MGF of $\epsilon \|q\|$ in the generalized switch

We prove the following lemma.

Lemma C.3.1. Consider a generalized switch parametrized by $\epsilon$ as described in Theorem 5.2.3. Then, for each $\epsilon > 0$ there exists $\Theta > 0$ such that $E \left[ e^{\theta \epsilon \|q\|} \right] < \infty$ for all $\theta \in [-\Theta, \Theta]$.

Proof. (of Lemma C.3.1) First observe that if $\theta = 0$ the lemma holds trivially. Therefore, in this proof we assume $\theta \neq 0$. We use Foster-Lyapunov theorem (Proposition 6.13 in [21]) with Lyapunov function $V(q) = e^{\theta \epsilon \|q\|}$. In this proof we use the notation

$$E_q [\cdot] \triangleq E [\cdot | q] \quad \text{and} \quad E_t [\cdot] \triangleq E [\cdot | T(k) = t]$$

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

$$E_q \left[ e^{\theta \epsilon \|q\|} - e^{\theta \epsilon \|q\|} \right] = E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} + 1 - e^{\theta \epsilon \|q\|} \right] + o(\epsilon^2)$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} + 1 \right] + o(\epsilon^2)$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} + 1 \right] + o(\epsilon^2)$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} + 1 \right] + o(\epsilon^2)$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} + 1 \right] + o(\epsilon^2)$$

where (a) holds expanding the product and rearranging terms in Lemma 5.3.2; (b) holds after adding and subtracting $E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right]$, and reorganizing terms; (c) holds because the arrival process is independent of the queue lengths and services processes.

But

$$E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right] = E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right]$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right]$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right]$$

$$= E_q \left[ e^{\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right]$$

where the last equality holds by Lemma 5.3.2 and because $\overline{B}$ is bounded.

Rearranging terms we obtain

$$E_q \left[ e^{\theta \epsilon \|q\|} - e^{\theta \epsilon \|q\|} \right] = 1 - E_q \left[ e^{\theta \epsilon \|q\|} \right] + E \left[ e^{\theta \epsilon \|q\|} \right] E_q \left[ e^{-\theta \epsilon \|q\|} - e^{-\theta \epsilon \|q\|} \right] + o(\epsilon^2)$$

(28)

Observe that the right hand side of Equation (28) is bounded because $\overline{q}_i \leq \overline{q}_i \leq S_{\text{max}}$ and $\overline{q}_i \leq A_{\text{max}}$ with probability 1 for all $i \in \{1, \ldots, n\}$. Also, $\overline{B}$ is bounded takes a finite number of values.

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

$$E \left[ e^{\theta \epsilon \|q\|} - 1 \right] < -\delta \quad \forall \theta \in [-\Theta, 0] \cup (0, \Theta]$$

The proof is similar to the proof of Equation (26) because $E \left[ \overline{B} \right] = b^{(l)}$, so we omit it. \qed
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 did in the proofs of our Theorems.

D.1  Proof of Claim 4.3.3

Proof. (of Claim 4.3.3) We have

$$0 \leq \frac{\theta \epsilon^2}{2} E \left[ \left( \sum_{i=1}^{n} \pi_i \right)^2 \right] \leq \epsilon^2 \left( \frac{\theta S_{\max} \theta^2}{2} \right) E \left[ \sum_{i=1}^{n} \pi_i \right] \leq \epsilon^3 \left( \frac{\theta S_{\max} \theta^2}{2} \right)$$

where (a) holds because, by definition of unused service, we have $\pi_i \leq \pi_i \leq S_{\max}$ and all terms are nonnegative; and (b) holds by Lemma 4.3.1.

Therefore,

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

\[Q.E.D.\]

D.2  Proof of Claim 5.3.4

Now we prove Claim 5.3.4.

Proof. (of Claim 5.3.4) We have

$$0 \leq \frac{\theta \epsilon^2}{2} E \left[ \left( \langle c^{(f)}, \overline{u} \rangle \right)^2 \right] \leq \epsilon^2 \left( \frac{\langle c^{(f)}, S_{\max} 1 \rangle \theta^2}{2} \right) E \left[ \langle c^{(f)}, \overline{u} \rangle \right] \leq \epsilon^3 \left( \frac{\langle c^{(f)}, S_{\max} 1 \rangle \theta^2}{2} \right)$$

where (a) holds because $\pi_i \leq \pi_i \leq S_{\max}$ and $c^{(f)} \geq 0$; and (b) holds by Lemma 5.3.1.

Therefore,

$$\frac{\theta \epsilon^2}{2} E \left[ \left( \langle c^{(f)}, \overline{u} \rangle \right)^2 \right] = O(\epsilon^3).$$

\[Q.E.D.\]

D.3  Proof of Claim 5.3.5

Now we prove Claim 5.3.5.

Proof. (of Claim 5.3.5) For the first expression, from Lemma 3.3.2 we have

$$E \left[ f_{\epsilon, \langle c^{(f)}, \overline{u} \rangle} (\theta) \right] = 1 + \frac{\theta \epsilon^2}{2} E \left[ \langle c^{(f)}, \overline{u} \rangle - \overline{B} \right] + \frac{\theta^2 \epsilon^2}{2} E \left[ \left( \langle c^{(f)}, \overline{u} \rangle - \overline{B} \right)^2 \right] + O(\epsilon^3)$$

$$= 1 + \frac{\theta^2 \epsilon^2}{2} E \left[ \left( \langle c^{(f)}, \overline{u} \rangle - \overline{B} \right)^2 \right] + O(\epsilon^3),$$

where (a) holds because, by definition of unused service, we have $\pi_i \leq \pi_i \leq S_{\max}$ and all terms are nonnegative; and (b) holds by Lemma 4.3.1.

Therefore,

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

\[Q.E.D.\]
where the last equality holds by Equation (18). Also,
\[
0 \leq \frac{(\theta \epsilon)^2}{2} E \left[ \left( \langle c^{(t)}, \overline{\sigma} \rangle - \overline{B} \right)^2 \right] \overset{(a)}{=} e^2 \left( \frac{\left( (c^{(t)}, A_{\text{max}}) + B_{\text{max}} \right) \theta^2}{2} \right) E \left[ \left( \langle c^{(t)}, \overline{\sigma} \rangle - \overline{B} \right)^2 \right] \overset{(b)}{=} e^3 \left( \frac{\left( (c^{(t)}, A_{\text{max}}) + B_{\text{max}} \right) \theta^2}{2} \right)
\]
where (a) holds because \( \overline{\sigma} \leq A_{\text{max}} \) with probability 1, \( c^{(t)} \geq 0 \), \( \overline{B} \) is bounded by a constant that we denote \( B_{\text{max}} \) and because all quantities are nonnegative; and (b) holds by Equation (18). Then,
\[
\frac{(\theta \epsilon)^2}{2} E \left[ \left( \langle c^{(t)}, \overline{\sigma} \rangle - \overline{B} \right)^2 \right] \text{ is } O(\epsilon^3).
\]
Therefore,
\[
E \left[ f_{\epsilon, \langle c^{(t)}, \overline{\sigma} \rangle} \right] = 1 + \theta \epsilon^2 + O(\epsilon^3).
\]
This proves the first equation of the claim.

For the second expression, using Lemma 3.3.2 we obtain
\[
E \left[ f_{\epsilon, \langle B - \langle c^{(t)}, \overline{\sigma} \rangle \rangle} \right] = 1 = \theta \epsilon E \left[ \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right] + \frac{(\theta \epsilon)^2}{2} E \left[ \left( \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right)^2 \right] + O(\epsilon^3).
\]
But
\[
0 \leq \frac{(\theta \epsilon)^2}{2} E \left[ \left( \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right)^2 \right] \overset{(a)}{=} e^2 \left( \frac{\left( B_{\text{max}} + \langle c^{(t)}, S_{\text{max}} \rangle \right) \theta^2}{2} \right) E \left[ \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right] \overset{(b)}{=} e^3 \left( \frac{\left( B_{\text{max}} + \langle c^{(t)}, S_{\text{max}} \rangle \right) \theta^2}{2} \right)
\]
where (a) holds because \( \overline{\sigma} \leq S_{\text{max}} \) with probability 1, \( c^{(t)} \geq 0 \), \( \overline{B} \leq B_{\text{max}} \) and all quantities are nonnegative (we clarify this statement below); and (b) holds by Lemma 5.3.1 and because \( E \left[ \langle c^{(t)}, \overline{\sigma} \rangle \right] \geq 0 \) since \( \overline{\sigma} \geq 0 \) and \( c^{(t)} \geq 0 \).

Observe that, even though \( \overline{\sigma} \) does not necessarily belong to the capacity region \( \mathcal{C} \), there exists \( t \in \mathcal{T} \) such that \( \overline{\sigma} \in \mathcal{S}^{(t)} \). Therefore, \( E \left[ \overline{\sigma} \right] \in \mathcal{C} \) and we obtain
\[
E \left[ \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right] = b^{(t)} - \langle c^{(t)}, E \left[ \overline{\sigma} \right] \rangle \geq 0,
\]
by the definition of \( \mathcal{C} \) given in Equation (14).

Then,
\[
\frac{(\theta \epsilon)^2}{2} E \left[ \left( \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right)^2 \right] \text{ is } O(\epsilon^3).
\]
Therefore,
\[
E \left[ f_{\epsilon, \langle B - \langle c^{(t)}, \overline{\sigma} \rangle \rangle} \right] = 1 = \theta \epsilon E \left[ \overline{B} - \langle c^{(t)}, \overline{\sigma} \rangle \right] + O(\epsilon^3).
\]
References

[1] Dimitris Bertsimas, David Gamarnik, and John N. Tsitsiklis. Performance of multiclass markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Probab.*, 11(4):1384–1428, 11 2001.

[2] A. Braverman. Steady-state analysis of the join the shortest queue model in the Halfin-Whitt regime. *arXiv preprint arXiv:1801.05121*, 2018.

[3] A. Braverman and J.G. Dai. Stein’s method for steady-state diffusion approximations of M/Ph/n+ M systems. *The Annals of Applied Probability*, 27(1):550–581, 2017.

[4] A. Braverman, I. Gurvich, and J. Huang. On the Taylor expansion of value functions. *arXiv preprint arXiv:1804.05011*, 2018.

[5] Anton Braverman, JG Dai, and Jiekun Feng. Stein’s method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stochastic Systems*, 6(2):301–366, 2017.

[6] Anton Braverman, JG Dai, and Masakiyo Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stochastic Systems*, 7(1):143–196, 2017.

[7] H. Chen and H.Q. Ye. Asymptotic optimality of balanced routing. *Operations Research*, 60(1):163–179, 2012.

[8] J.G. Dai. Steady-state approximations: Achievement lecture. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 1–1. ACM, 2018.

[9] Jim G Dai, Wuqin Lin, et al. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *The Annals of Applied Probability*, 18(6):2239–2299, 2008.

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

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

[12] P. Eschenfeldt and D. Gamarnik. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Mathematics of Operations Research*, 2018.

[13] G Foschini and JACK Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26(3):320–327, 1978.

[14] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, pages 56–90, 2006.

[15] Robert E Gaunt and Neil Walton. Stein’s method for the single server queue in heavy traffic. *arXiv preprint arXiv:1805.08003*, 2018.

[16] G.R. Gupta and N.B. Shroff. Delay analysis for wireless networks with single hop traffic and general interference constraints. *IEEE/ACM Transactions on Networking (TON)*, 18(2):393–405, 2010.

[17] Itai Gurvich et al. Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *The Annals of Applied Probability*, 24(6):2527–2559, 2014.

[18] Itai Gurvich, Junfei Huang, and Avishai Mandelbaum. Excursion-based universal approximations for the Erlang-A queue in steady-state. *Mathematics of Operations Research*, 39(2):325–373, 2013.
[19] A. Gut. *Probability: A Graduate Course*, volume 75. Springer Science & Business Media, 2012.

[20] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, pages 502–525, 1982.

[21] B. Hajek. *Random Processes for Engineers*. Cambridge university press, 2015.

[22] J. M. Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Ann. App. Probab.*, pages 822–848, 1998.

[23] J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, pages 339–368, 1999.

[24] J Michael Harrison and Marcel J López. Heavy traffic resource pooling in parallel-server systems. *Queueing systems*, 33(4):339–368, 1999.

[25] J.M. Harrison. Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 147–186. Springer, 1988.

[26] P.G. Harrison and N.M. Patel. *Performance Modeling of Communication Networks and Computer Architectures*. Addison-Wesley Longman Publishing Co., Inc., 1992.

[27] Junfei Huang and Itai Gurvich. Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Operations Research*, 66(4):1168–1188, 2018.

[28] D. Hurtado-Lange and S.T. Maguluri. A novel view of the drift method for heavy traffic limits of queueing systems. In *ACM SIGMETRICS Performance Evaluation Review*. ACM, 2018.

[29] W. N. Kang, F. P. Kelly, N. H. Lee, and R. J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Applied Probability*, pages 1719–1780, 2009.

[30] J. F. C. Kingman. Some inequalities for the queue GI/G/1. *Biometrika*, pages 315–324, 1962.

[31] JFC Kingman. The single server queue in heavy traffic. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 902–904. Cambridge University Press, 1961.

[32] JFC Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 383–392, 1962.

[33] J. Köllerström. Heavy traffic theory for queues with several servers. i. *Journal of Applied Probability*, 11(3):544–552, 1974.

[34] J. Lehoczky. Real-time queueing theory. *Real-Time Systems Symposium*, page 186, 1996.

[35] John P Lehoczky. Using real-time queueing theory to control lateness in real-time systems. *ACM SIGMETRICS Performance Evaluation Review*, 25(1):158–168, 1997.

[36] B. Li, X. Kong, and L. Wang. Optimal load-balancing for high-density wireless networks with flow-level dynamics. *arXiv preprint arXiv:1805.00158*, 2018.

[37] D.V. Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

[38] X. Liu and L. Ying. A simple steady-state analysis of load balancing algorithms in the sub-Halfin-Whitt regime. *arXiv preprint arXiv:1804.02622*, 2018.
[39] Y. Lu, Q. Xie, G. Kliot, A. Geller, J.R. Larus, and A. Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11):1056–1071, 2011.

[40] E. Lukacs. *Characteristic Functions*. Griffin, 1970.

[41] Siva Theja Maguluri, Sai Kiran Burle, and R Srikant. Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems*, 88(3-4):279–309, 2018.

[42] Siva Theja Maguluri and R. Srikant. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stoch. Syst.*, 6(1):211–250, 2016.

[43] Siva Theja Maguluri, R. Srikant, and Lei Ying. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation*, 81:20–39, 2014.

[44] Siva Theja Maguluri, Rayadurgam Srikant, and Lei Ying. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation*, 81:20–39, 2014.

[45] Aviash Mandelbaum and Alexander L Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cµ-rule. *Operations Research*, 52(6):836–855, 2004.

[46] Kneale T Marshall. Some inequalities in queuing. *Operations research*, 16(3):651–668, 1968.

[47] Laurent Massoulié and James W Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication systems*, 15(1-2):185–201, 2000.

[48] N. McKeown, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input queued switch. In *Proceedings of IEEE INFOCOM*, pages 296–302, 1996.

[49] S.P. Meyn. Stability and asymptotic optimality of generalized MaxWeight policies. *SIAM J. Control and Optimization*, 2008. to appear.

[50] M. Mitzenmacher. Load balancing and density dependent jump Markov processes. In *focs*, page 213. IEEE, 1996.

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

[52] A.M. Mood. *Introduction to the Theory of Statistics*. McGraw-hill, 1950.

[53] R. Srikant and L. Ying. *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press, 2014.

[54] A. L. Stolyar. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, pages 1–53, 2004.

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

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

[57] M. van der Boor, S.C. Borst, J.S.H. van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *arXiv preprint arXiv:1806.05444*, 2018.

[58] Maria Vlasiou, Jiheng Zhang, and Bert Zwart. Insensitivity of proportional fairness in critically loaded bandwidth sharing networks. *arXiv preprint arXiv:1411.4841*, 2014.
N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmission*, 32(1):15–27, 1996.

Chang-Heng Wang, Siva Theja Maguluri, and Tara Javidi. Heavy traffic queue length behavior in switches with reconfiguration delay. In *INFOCOM 2017–IEEE Conference on Computer Communications*, IEEE, pages 1–9. IEEE, 2017.

W. Wang, S.T. Maguluri, R. Srikant, and L. Ying. Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies. arXiv preprint arXiv:1808.02120, 2018.

Weina Wang, Siva Theja Maguluri, R. Srikant, and Lei Ying. Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *SIGMETRICS Perform. Eval. Rev.*, 45(3):232–245, March 2018.

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

R. J. Williams. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems Theory and Applications*, pages 27–88, 1998.

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

Heng-Qing Ye and David D Yao. A stochastic network under proportional fair resource control—diffusion limit with multiple bottlenecks. *Operations Research*, 60(3):716–738, 2012.

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

Lei Ying. On the approximation error of mean-field models. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS ’16, pages 285–297, New York, NY, USA, 2016. ACM.

Lei Ying. Stein’s method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):12:1–12:27, June 2017.

X. Zhou, J. Tan, and N. Shroff. Flexible load balancing with multi-dimensional state-space collapse: Throughput and heavy-traffic delay optimality. arXiv preprint arXiv:1806.02939, 2018.