A Novel View of the Drift Method for Heavy-Traffic Limits of Queueing Systems

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

Abstract

The Drift method has been recently developed to study queueing systems in heavy-traffic \cite{2, 8, 9}. This method was successfully used to obtain the heavy-traffic scaled sum queue lengths of several systems, even when the so-called Complete Resource Pooling (CRP) condition is not satisfied, i.e., when there is more than one bottleneck resource. In this paper, we present an alternate view of the Drift method to explain why the method works to give these results. This view also points out the limitations of the method, and we show that there are major challenges in generalizing this method to obtain the moments of the individual queue lengths or the joint distribution of queue lengths in the case when the CRP condition is not satisfied. We do this by showing that the Drift method only gives an under-determined system of linear equations on the moments of queue lengths. However, these equations can be used to obtain bounds on the moments of the queue lengths as well as bounds on the tail probabilities.

1 Introduction

Heavy-traffic limits of various queueing systems have been studied in the literature using fluid limits, diffusion limits and Brownian Motion processes \cite{3}. In this approach, the queueing process is scaled appropriately, and the limiting fluid or diffusion process is studied. Typically, the limit of a diffusion scaled process is shown to converge to a Reflected Brownian Motion (RBM) process that lives in a lower dimensional subspace. This is called State Space Collapse (SSC). Then, one solves for the steady-state distribution of the lower dimensional RBM. It is then shown that the heavy-traffic scaled queue length process converges to the steady-state distribution of the RBM, by establishing a so-called limit interchange argument.

This program to obtain heavy-traffic limits using RBMs has been successfully applied to several queueing systems, where the state space collapses to a line, i.e., to a one-dimensional subspace. In these systems, there is exactly one constraint that is tight in heavy-traffic and, therefore, they are said to satisfy the Complete Resource Pooling (CRP) condition. Intuitively, this means that in the heavy-traffic limit, there is a single bottleneck resource. Under the CRP condition, in the diffusion limit one obtains an RBM on a line, whose steady-state distribution is known to be exponential. However, a major challenge is in using this program for queueing systems where the CRP condition is not satisfied (i.e., when there are multiple resources that are simultaneously in heavy-traffic) is that one needs to solve for the steady-state distribution of a RBM in a multidimensional subset of $\mathbb{R}^n$ \cite{5}, and this is not known in general. One of the simplest systems where the CRP condition is not satisfied is an input-queued switch \cite{11, 8, 9}.

More recently, the Drift method was developed to study heavy-traffic limits of queueing systems \cite{2, 8, 9}. This approach is a generalization of the Kingman’s bound in a $G/G/1$ queue \cite{6}. In this method, one works

∗This work was partially funded by the CONICYT PFCHA/DOCTORADO BECAS CHILE/2018 - 72190413
directly with the original queueing system, without a fluid or diffusion scaling. Therefore, the SSC results in the diffusion limit are not applicable here. A novel definition of SSC is developed, and it is established using a Lyapunov drift argument. We will not focus on the details of the SSC here, but we state the result in Section 2 for completeness. Then, a test function of the queue lengths vector is picked, and its drift is set to zero in steady-state. By the definition of steady-state, the drift of any function is zero, as long as its expectation is known to be finite. The choice of this test function is important in the Drift method. It was shown that the square of the norm of the projection of the queue length vector into the space of the SSC is a useful test function [2, 8, 9, 14]. Using this test function, the expectation of the heavy-traffic scaled sum queue lengths was characterized for many systems, including some systems where the CRP condition is not satisfied. When the CRP condition is satisfied, a similar approach with different test functions can be used to obtain all the moments of the queue length. It can be shown that one can then uniquely determine the steady-state distribution of the heavy-traffic scaled queue lengths in the heavy-traffic limit and that it is indeed the exponential distribution.

However, in the case when the CRP condition is not satisfied, the joint distribution of the heavy-traffic scaled queue lengths is not known in general, and this is an open problem. Solving this problem is equivalent to solving for the steady-state distribution of the corresponding multi-dimensional RBM. Obtaining the joint distribution of queue lengths is equivalent to obtaining all the moments of all linear combinations of the queue lengths (under some reasonable conditions on the joint distribution). So far, the key step in using the Drift method, is to design the correct test function to obtain all these moments. However, it is not clear a priori if there are test functions that give all these moments. In this paper, we present an alternate way of thinking about the Drift method. Instead of trying to guess the right test function, this point of view shows that one can think about solving a set of linear equations. This system of linear equations turns out to be under-determined, and the major challenge in using the Drift method to obtain the complete joint distribution of queue lengths when the CRP condition is not satisfied is to obtain more equations using the constraints in the system, in order to solve for all the unknowns. A similar approach of using linear equations in queueing systems was studied in a very different context by [7].

Another utility of this alternate view is the following. Even though the first moment of the sum of the heavy-traffic scaled queue lengths in an input-queued switch was obtained in [8] and [9], it was not completely clear why this method works. Given the notorious difficulty in solving the steady-state of the RBM, it has been a little surprising that a simple drift based argument can be used to obtain the expectation of sum of the components of the RBM. Consequently, it was not clear how far this method can be pushed to obtain the higher moments. The alternate point of view presented in the next sections not only explains why the Drift method has worked, but shows that there are major challenges to be overcome in order to obtain the higher moments.

Moreover, the system of linear equations that we obtain, even though is under-determined, it can be used to get bounds on the mean queue lengths and moments, as well as tail probability bounds by solving a linear program. In this paper, for simplicity of exposition, we focus on a simple system of three queues which can be thought of as a special case of a $2 \times 2$ input-queued switch operating under MaxWeight. However, the same approach can easily be generalized to larger systems. In Section 2 we describe the model, in Section 3 we present and prove our main results on the system of equations, in Section 4, we present the bounds by solving a linear program and in Section 5 we present a discussion of our results before concluding in Section 6.

### 1.1 Notation

In this subsection we introduce the notation we will use along the paper. We use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, $\mathbb{R}^n$ to denote the set of $n$-dimensional vectors with real elements and $\mathbb{R}^n_+$ to denote the set of $n$-dimensional vectors with nonnegative elements. We use bold letters to denote vectors and its elements are denoted by a nonbold letter with a subscript. For example, $\mathbf{x} \in \mathbb{R}^n$ has elements $x_i \in \mathbb{R}$ for all $i \in \{1, \ldots, n\}$. Given two vectors $\mathbf{x}$ and $\mathbf{y}$, we write $\langle \mathbf{x}, \mathbf{y} \rangle$ to denote the inner product between $\mathbf{x}$ and $\mathbf{y}$ and $\| \mathbf{x} \|$ to denote the euclidean norm, i.e. $\| \mathbf{x} \| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Given two random variables $X$ and $Y$, we denote $E[X]$ the expected value of $X$, $\text{Var}[X]$ the variance of $X$ and $\text{Cov}(X,Y)$
the covariance between $X$ and $Y$. Given an event $E$, we denote $P\{E\}$ the probability of $E$. We use $\binom{n}{k}$ to denote the binomial coefficient, i.e. $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

2 Model

Consider a three-queue system operating in discrete time. Each queue has an independent arrival process which is i.i.d. across time. Let $\{a_i(k) : k \geq 1\}$ be the arrival process associated to queue $i \in \{1, 2, 3\}$ and let $a(k) = (a_1(k), a_2(k), a_3(k))$, $k \geq 1$. For each $i \in \{1, 2, 3\}$, let $\lambda_i = E[a_i(1)]$, $\sigma_i^2 = Var[a_i(1)]$ and let $A_{\text{max}}$ be a finite constant such that $a_i(1) \leq A_{\text{max}}$ with probability 1 for all $i \in \{1, 2, 3\}$. We assume all packets are size 1, i.e. each of them takes exactly one time slot to be processed and that, in each time slot, service occurs after arrivals. Let $s_i(k)$ be the potential service in queue $i \in \{1, 2, 3\}$ in time slot $k$ and $s(k) = (s_1(k), s_2(k), s_3(k))$. In each time slot whether queues 1 and 2 can be served, or queue 3 alone, i.e. $s(k)$ can only take the values $s(k) = (1, 1, 0)$ or $s(k) = (0, 0, 1)$ for all $k \geq 1$. Therefore, a scheduling problem must be solved in each time slot. We call $s(k)$ vector of potential service because it represents what the servers offer, but it does not need to be the actual service. In particular, the potential service might be one even if the queue is empty and, in such case, the actual service is zero. The difference between the potential and actual service is the unused service. We denote $u_i(k)$ the unused service in queue $i \in \{1, 2, 3\}$ in time slot $k$ and $u(k) = (u_1(k), u_2(k), u_3(k))$. Let $q_i(k)$ the number of packets in queue $i \in \{1, 2, 3\}$ at the beginning of time slot $k$ and $q(k) = (q_1(k), q_2(k), q_3(k))$.

The scheduling decision in each time slot is made according to MaxWeight algorithm, i.e.

$$s(k) \in \arg \max \{\{x, q(k)\} : x \in \{(1, 1, 0), (0, 0, 1)\}\},$$

where ties are broken at random. Note that MaxWeight algorithm only uses the queue length at the beginning of each time slot to make a decision. Therefore, the number of arrivals in time slot $k$ is independent of the queue lengths and potential service.

The dynamics of queue $i \in \{1, 2, 3\}$ are described by the following equation

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k),$$

$$k \geq 1. \tag{2}$$

Since $s_i(k) \in \{0, 1\}$ and MaxWeight only uses $q(k)$, the following equations hold

$$q_i(k+1)u_i(k) = 0, \quad q_i(k)u_i(k) = 0, \quad a_i(k)u_i(k) = 0. \tag{3}$$

In other words, the unused service is nonzero only if the corresponding queue is currently empty and there are no arrivals. In such case, the queue will remain empty to the next time slot. However, for $i \neq j$, $q_i(k+1)u_j(k)$, $q_i(k)u_j(k)$ and $a_i(k)u_j(k)$ do not need to be zero.

We assume that $u_3(k) = 0$ for all $k \geq 1$ without great loss of generality because of the following reason. If $u_3(k) > 0$, then by definition of unused service we must have $q_3(k) = 0$, and by MaxWeight scheduling (equation (1)) we must have $q_3(k) \geq q_1(k) + q_2(k)$. Therefore, $u_3(k) > 0$ is possible only if $q_1(k) = q_2(k) = q_3(k) = 0$. In this case, we can pretend that we always pick $\mathbf{s} = (1, 1, 0)$. We lose some generality with this assumption because arrivals occur after MaxWeight algorithm is used to select a schedule. So, the choice of service vector when all queues are empty is not entirely irrelevant. However, we ignore this issue in this paper for simplicity of exposition, and the influence of this assumption in our analysis is negligible because we do heavy-traffic analysis.

2.1 Heavy-traffic analysis and state space collapse

The three-queue system described above corresponds to a $2 \times 2$ input-queued switch operating under MaxWeight, where there are no arrivals to one of the queues. We pick this model because it is one of the simplest systems that does not satisfy the CRP condition. It is known [10, 13] that the capacity region of this queueing system is

$$\mathcal{C} = \{\mathbf{\lambda} \in \mathbb{R}^3_+ : \lambda_1 + \lambda_3 \leq 1, \lambda_2 + \lambda_3 \leq 1\}. \tag{4}$$
In other words, for all $\lambda$ in the interior of $C$, the Discrete Time Markov Chain $\{q(k) : k \geq 1\}$ is positive recurrent.

To take the heavy-traffic limit, we fix $\eta \in (0, 1)$ and we consider a set of three-queue systems as described above, parametrized by $\epsilon \in (0, 1 - \eta)$. The parametrization is such that $\lambda^{(\epsilon)} = E[a^{(\epsilon)}(1)] = (\eta, \eta, 1 - \epsilon - \eta)$ and the $i$th queue has variance $\left(\sigma_i^{(\epsilon)}\right)^2$. In heavy-traffic, we take the limit as $\epsilon \downarrow 0$, so we approach the boundary of the capacity region $C$, where both the inequalities in (4) are tight. This corresponds to heavy-traffic analysis of a $2 \times 2$ incompletely saturated input-queued switch operating under MaxWeight, which was studied by [9]. In particular, SSC was established and the heavy-traffic limit of a linear combination of the queue lengths scaled by the heavy-traffic parameter was computed. In the rest of this section we describe these results for the particular case of the three-queue system.

We start with SSC. [9] proved that SSC occurs into $K = \{x \in \mathbb{R}^3_+: x_3 = x_1 + x_2\}$, which is a two-dimensional cone in $\mathbb{R}^3$. The cone $K$ spans the subspace

$$S = \{x \in \mathbb{R}^3: x_3 = x_1 + x_2\},$$

and [9] also proved that SSC can be thought of as collapse into the subspace $S$. In this paper we work with SSC into the subspace $S$.

Given a vector $y \in \mathbb{R}^3_+$, we denote $y\parallel$ its projection onto $S$ and $y\perp = y - y\parallel$. Using the orthogonality principle, it can be proved that

$$y\parallel = \left(\frac{2y_1 - y_2 + y_3}{3}, \frac{-y_1 + 2y_2 + y_3}{3}, \frac{y_1 + y_2 + 2y_3}{3}\right).$$

Observe that, denoting $y_{\parallel i}$ the $i$th element of $y\parallel$, we have $y_{\parallel 3} = y_{\parallel 1} + y_{\parallel 2}$. Now we present the SSC result as proved by [9].

**Proposition 2.1.** Consider a set of three-queue systems parametrized by $\epsilon \in (0, 1 - \eta)$ as described above. Suppose $\left(\sigma_i^{(\epsilon)}\right)^2 \leq \bar{\sigma}^2$ for all $i \in \{1, 2, 3\}$, for some $\bar{\sigma}^2$ that does not depend on $\epsilon$. Let $\overline{q}^{(\epsilon)}$ be a steady-state vector to which the queue lengths process of the parametrized system converges in distribution. Then, for each system with $0 < \epsilon < \frac{1}{2}\min\{\eta, 1 - \eta\}$ the steady-state vector of queue lengths satisfies

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

where $M_m$ is independent of $\epsilon$ for each $m = 1, 2, \ldots$.

It is known that $\left\|\overline{q}^{(\epsilon)}_\parallel\right\|$ is $O\left(\frac{1}{\epsilon}\right)$. Therefore, Proposition 2.1 implies that $\overline{q}^{(\epsilon)}_\perp$ is negligible in heavy-traffic. In other words, the queue lengths process in steady-state $\overline{q}$ behaves as if it belongs to $S$ in heavy-traffic. Since the two constraints of $C$ are tight in $S$, the three-queue system does not satisfy the CRP condition.

With this notion of SSC, [9] computed the heavy-traffic limit of a linear combination of the vector of queue lengths. In the next proposition we present the result for the specific case of the three-queue system.

**Proposition 2.2.** Consider a set of the three-queue systems indexed by the heavy-traffic parameter $\epsilon$, as described in Proposition 2.1. Further, assume that $\lim_{\epsilon \downarrow 0} \left(\sigma_i^{(\epsilon)}\right)^2 = \sigma_i^2$ for $i \in \{1, 2, 3\}$. Then,

$$\lim_{\epsilon \downarrow 0} \epsilon E\left[\overline{q}_1 + \overline{q}_2 + \overline{q}_3\right] = \frac{2}{3} \left(\sigma_1^2 + \sigma_2^2 + \sigma_3^2\right),$$

where $\overline{q}_1, \overline{q}_2, \overline{q}_3$ are the elements of $\overline{q}$.

In the proof of Proposition 2.2, the drift of the test function $V(q) = \left\|q\right\|^2$ is set to zero and SSC is used to obtain upper and lower bounds that match in heavy-traffic [9]. With this technique, the expected sum of queue lengths scaled by the heavy-traffic parameter is computed. It was shown by [9] that in addition to the sum of the queue lengths, the mean of some other linear combinations of queue lengths can be computed. However, not all possible linear combinations or higher moments are known. In Section 3 we show how drift based arguments alone are not sufficient to answer these questions.
3 Systems of Equations to Compute the Moments of Scaled Queue Lengths

In this section we present the main result of our paper. In the Drift method, one of the key challenges is to get a handle on the unused service. In general, when one sets to zero the drift of a polynomial test function in steady-state, terms of the form \( q_i(k+1)u_j(k) \) arise. The idea is to use a test function that captures the geometry of SSC so that we can show that all the \( q_i(k+1)u_j(k) \) terms are small. Therefore, the choice of the test function is important, and the region into which SSC happens must be used in this choice. The quadratic test function, \( V(q) = ||q||^2 \) has been successfully used in [2, 8, 9, 14] to obtain the mean of sum of the queue lengths, similar to Proposition 2.2. Typically one uses polynomial test functions of degree \( (m+1) \) to get bounds on the expected value of \( m^{th} \) powers of queue lengths. Therefore, in order to obtain bounds on the queue lengths, one must use quadratic test functions. In order to get all linear combinations of the queue lengths, one can search through all the quadratic test functions, and this is equivalent to searching through all the quadratic monomials. The following theorem presents the result of using all quadratic monomial test functions.

**Theorem 3.1.** Consider a set of three-queue systems as described in Section 2, indexed by the heavy-traffic parameter \( \epsilon \). Let \( \mathbf{q}^{(e)} \) be a steady-state vector to which the queue lengths process of the system indexed by \( \epsilon \) converges in distribution and let \( \mathbf{a}^{(e)} \) be a stationary random vector such that \( \mathbf{a}^{(e)} \) has the same distribution as \( \beta_i(1) \) for each \( i \in \{1, 2, 3\} \), and such that \( \mathbf{a}_1^{(e)}, \mathbf{a}_2^{(e)} \) and \( \mathbf{a}_3^{(e)} \) are independent. Let \( \mathbf{s}(\mathbf{q}^{(e)}) \) and \( \mathbf{u}(\mathbf{q}^{(e)}, \mathbf{a}^{(e)}) \) be the corresponding potential and unused service, respectively. Further, assume \( \epsilon_0 \) such that \( \epsilon \to 0 \) as \( \epsilon \to 0 \). Define \( \mathbf{s}^{(e)} = \mathbf{s}(\mathbf{q}^{(e)}) \) and \( \mathbf{a}^{(e)} = \mathbf{u}(\mathbf{q}^{(e)}, \mathbf{a}^{(e)}) \) and \( \mathbf{q}^{(e)} = \mathbf{q}^{(e)} + \mathbf{a}^{(e)} - \mathbf{u}^{(e)} \). Then, the following system of equations is satisfied

\[
\lim_{\epsilon \to 0} \epsilon E \left[ q_1^{(e)} \right] = \frac{4\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{6} - \lim_{\epsilon \to 0} E \left[ (q_1^{(e)})^+ \right] - \lim_{\epsilon \to 0} E \left[ (q_2^{(e)})^+ \right]
\]

\[
\lim_{\epsilon \to 0} \epsilon E \left[ q_2^{(e)} \right] = \frac{\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2}{6} - \lim_{\epsilon \to 0} E \left[ (q_2^{(e)})^+ \right] - \lim_{\epsilon \to 0} E \left[ (q_3^{(e)})^+ \right]
\]

\[
\lim_{\epsilon \to 0} \epsilon E \left[ q_1^{(e)} + q_2^{(e)} \right] = \frac{-2\sigma_1^2 - 2\sigma_2^2 + \sigma_3^2}{3} + 2 \lim_{\epsilon \to 0} E \left[ (q_1^{(e)})^+ \right] + \lim_{\epsilon \to 0} E \left[ (q_2^{(e)})^+ \right] + \lim_{\epsilon \to 0} E \left[ (q_3^{(e)})^+ \right]
\]

Observe that Theorem 3.1 gives 3 equations in 4 variables, viz. \( \lim_{\epsilon \to 0} \epsilon E \left[ q_1^{(e)} \right], \lim_{\epsilon \to 0} \epsilon E \left[ q_2^{(e)} \right], \lim_{\epsilon \to 0} \epsilon E \left[ q_3^{(e)} \right] \) and \( \lim_{\epsilon \to 0} \epsilon E \left[ (q_1^{(e)})^+ \right] + \lim_{\epsilon \to 0} \epsilon E \left[ (q_2^{(e)})^+ \right] \). Therefore, it is under-determined and it does not have a unique solution. We need one more equation to solve all the four unknowns. Now we present a proof of Theorem 3.1.

**Proof of Theorem 3.1.** For ease of exposition, in this proof we omit the dependence on \( \epsilon \) of the variables.

SSC implies that for all \( i \in \{1, 2, 3\} \)

\[
\lim_{\epsilon \to 0} \epsilon E \left[ \mathbf{q}_i \right] = \lim_{\epsilon \to 0} \epsilon E \left[ \mathbf{q}_{||i} \right],
\]

where \( \mathbf{q}_{||1}, \mathbf{q}_{||2} \) and \( \mathbf{q}_{||3} \) are the elements of \( \mathbf{q}_{||} \). Also, by equation (6), \( \mathbf{q}_{||3} = \mathbf{q}_{||1} + \mathbf{q}_{||2} \). Therefore, the most general quadratic test function is

\[
V(q) = \beta_1 q_{||1}^2 + \beta_2 q_{||2}^2 + \beta_3 q_{||1} q_{||2},
\]

where \( \beta_1, \beta_2, \beta_3 \in \mathbb{R} \). Setting the drift of \( V(q) \) to zero is equivalent to setting to zero the drift of the following three test functions

\[
V_1(q) = q_{||1}^2, \quad V_2(q) = q_{||2}^2 \quad \text{and} \quad V_3(q) = q_{||1} q_{||2}
\]
We first work with $V_1(q)$. Setting its drift to zero in steady-state and using equation (2) yields

$$
0 = E \left[ (\pi_{||1})^2 - \pi_{||1}^2 \right]
= E \left[ (\pi_{||1} - \pi_{||1} + \bar{\pi}_{||1})^2 - \bar{\pi}_{||1}^2 \right]
= E \left[ (\pi_{||1} - \pi_{||1})^2 \right] + 2E \left[ \bar{\pi}_{||1} (\pi_{||1} - \pi_{||1}) \right] - E \left[ \bar{\pi}_{||1}^2 \right] + 2E \left[ q_{||1} \bar{\pi}_{||1} \right],
(12)
$$

where the last term is obtained using equation (2) and reorganizing terms.

We compute term by term. First, recall that $\bar{\pi} = (1, 1, 0)$ or $\bar{\pi} = (0, 0, 1)$. Then, from (6) we obtain that $\bar{\pi}_{||1} = \frac{1}{4}$. Therefore,

$$
E \left[ (\pi_{||1} - \pi_{||1})^2 \right] = E \left[ \left( \pi_{||1} - \frac{1}{3} \right)^2 \right]
= E \left[ \pi_{||1}^2 \right] + \frac{1}{9} - \frac{2}{3} E \left[ \pi_{||1} \right]
\overset{(a)}{=} \text{Var} \left[ \pi_{||1} \right] \left( E \left[ \pi_{||1} \right] \right)^2 + \frac{1}{9} - \frac{2}{3} E \left[ \pi_{||1} \right]
\overset{(b)}{=} 4\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \epsilon^2,
$$

where $(a)$ holds by definition of variance, and $(b)$ holds because $\bar{\pi}_{||1} = \frac{\pi_1 - \pi_2 + \pi_3}{3}$, and by definition of $\bar{\pi}$.

Using again that $\bar{\pi}_{||1} = \frac{1}{4}$ for the second term of (12), we obtain

$$
2E \left[ \bar{\pi}_{||1} (\pi_{||1} - \pi_{||1}) \right] = 2E \left[ \bar{\pi}_{||1} \left( \pi_{||1} - \frac{1}{3} \right) \right]
\overset{(a)}{=} 2E \left[ \bar{\pi}_{||1} \right] E \left[ \pi_{||1} - \frac{1}{3} \right]
\overset{(b)}{=} - \frac{2}{3} \epsilon E \left[ \pi_{||1} \right]
$$

where $(a)$ holds because the queue lengths are independent of the arrival processes.

On the other hand, observe

$$
0 \leq E \left[ \pi_{||1}^2 \right] \leq E \left[ \| \pi \|^2 \right] \leq 4\epsilon,
$$

where the last inequality was proved by [9]. Therefore, $E \left[ \pi_{||1}^2 \right]$ is $O(\epsilon)$.

To compute the next term we use SSC. By definition of $\bar{\pi}_{||1}$ and $\bar{\pi}_{\perp}$, we have

$$
2E \left[ q_{||1} \pi_{||1} \right] = 2E \left[ q_{||1} \pi_{||1} \right] - 2E \left[ q_{||1} \bar{\pi}_{||1} \right],
$$

where

$$
E \left[ \pi_{||1}^2 \right] \leq E \left[ \pi_{||1}^2 | \pi_{||1} \right]
\leq E \left[ \left( \pi_{||1}^2, \pi_{||2}^2, \pi_{||3}^2 \right) | \pi_{||1}, \pi_{||2}, \pi_{||3} \right)
\overset{(a)}{=} \sqrt{E \left[ \| \pi_{||1} \|^2 \right] E \left[ \| \pi_{||2} \|^2 \right]}
\overset{(b)}{=} \sqrt{4\epsilon M_2},
$$

6
where (a) holds by Cauchy-Schwarz inequality, and (b) by SSC and the upper bound we used above for $E \left[ \| \tilde{v} \|_{1}^{2} \right]$. Similarly, $E \left[ \tilde{v}_{1}^+ \tilde{v}_{1} \right] \geq -\sqrt{\epsilon} M_{2}$. Therefore,

$$2E \left[ \tilde{v}_{1}^+ \tilde{v}_{1} \right] = 2E \left[ \tilde{v}_{1}^+ \tilde{v}_{1} \right] + O(\sqrt{\epsilon}).$$

Using equation (6) we obtain that

$$\bar{v}_{1} = \frac{1}{3} (2\bar{v}_{1} - \bar{v}_{2} + \bar{v}_{3}) = \frac{1}{3} (2\bar{v}_{1} - \bar{v}_{2}),$$

because we assumed $\bar{v}_{3} = 0$ (see last paragraph of Section 2). Additionally, by equation (3) we have $\tilde{v}_{1}^+ \bar{v}_{1} = 0$. Then,

$$E \left[ \tilde{v}_{1}^+ \bar{v}_{1} \right] = \frac{1}{3} E \left[ \tilde{v}_{1}^+ (2\bar{v}_{1} - \bar{v}_{2} + \bar{v}_{3}) \right] = -\frac{1}{3} E \left[ \tilde{v}_{1}^+ \bar{v}_{2} \right]$$

Therefore,

$$2E \left[ \tilde{v}_{1}^+ \bar{v}_{1} \right] = -\frac{2}{3} E \left[ \tilde{v}_{1}^+ \bar{v}_{2} \right] + O(\sqrt{\epsilon})$$

Taking the heavy-traffic limit and reorganizing terms in (12) we obtain

$$\lim_{\epsilon \downarrow 0} \epsilon E \left[ \tilde{v}_{1} \bar{v}_{1} \right] = \frac{4\sigma_{2}^{2} + \sigma_{3}^{2} + \epsilon^{2}}{6} - \lim_{\epsilon \downarrow 0} E \left[ \tilde{v}_{1}^+ \bar{v}_{2} \right],$$

which is equivalent to equation (8), by equation (11). Similarly, setting the drift of $V_{3}(q)$ to zero in steady-state we obtain equation (9). We omit the details here because the proof is similar to working with the drift of $V_{1}(q)$.

Finally, we set to zero the drift of $V_{3}(q)$ in steady-state. We obtain

$$0 = E \left[ \tilde{v}_{1}^+ \bar{v}_{1} - \tilde{v}_{1}^+ \bar{v}_{2} \right]$$

$$= E \left[ (\tilde{v}_{1}^+ - \bar{v}_{1}) (\bar{v}_{1}^+ + \bar{v}_{1}) (\bar{v}_{2} - \bar{v}_{1}) + \frac{2}{3} E \left[ \tilde{v}_{1}^+ \bar{v}_{2} + \bar{v}_{1}^+ \bar{v}_{1} \right] \right]$$

where the last equality is obtained based on previous computations and

$$E \left[ (\bar{v}_{1} - \bar{v}_{1}) (\bar{v}_{2} - \bar{v}_{2}) \right] = E \left[ \left( \bar{v}_{1} - \frac{1}{3} \right) (\bar{v}_{2} - \frac{1}{3}) \right]$$

$$= E \left[ \bar{v}_{1} \bar{v}_{2} \right] - \frac{1}{3} E \left[ \bar{v}_{1} \right] - \frac{1}{3} E \left[ \bar{v}_{2} \right] + \frac{1}{9}$$

$$(a) = \text{Cov} \left( \bar{v}_{1}, \bar{v}_{2} \right) + E \left[ \bar{v}_{1} \right] E \left[ \bar{v}_{2} \right] - \frac{1}{3} E \left[ \bar{v}_{1} \right] - \frac{1}{3} E \left[ \bar{v}_{2} \right] + \frac{1}{9}$$

$$(b) = -2\sigma_{1}^{2} - 2\sigma_{3}^{2} + \epsilon^{2}$$

where (a) holds by definition of covariance, and (b) holds by (6) and by definition of $\bar{v}_{1}$.

Taking the heavy-traffic limit and reorganizing terms on (13) we obtain

$$\lim_{\epsilon \downarrow 0} \epsilon E \left[ \tilde{v}_{1} + \tilde{v}_{2} \right] = \frac{-2\sigma_{1}^{2} - 2\sigma_{3}^{2} + \epsilon^{2}}{3} + 2 \lim_{\epsilon \downarrow 0} E \left[ \tilde{v}_{1}^+ \bar{v}_{2} + \bar{v}_{1}^+ \bar{v}_{2} \right],$$

which is equation (10).

This proves the theorem. \qed
In the following corollary of Theorem 3.1 we present a special case of the three-queue system where the system of equations can be solved. In other words, we add an assumption to obtain additional equations that allow us to solve the system of equations.

**Corollary 3.2.** Consider the queueing system described in Theorem 3.1. If the arrival processes to queue 1 and queue 2 have the same distribution, then

\[
\begin{align*}
\lim_{\epsilon \downarrow 0} \epsilon E[q_1] &= \lim_{\epsilon \downarrow 0} \epsilon E[q_2] = \frac{2\sigma_1^2 + \sigma_3^2}{6}, \\
\lim_{\epsilon \downarrow 0} E[q_1^+ \pi_2] &= \lim_{\epsilon \downarrow 0} E[q_2^+ \pi_1] = \frac{\sigma_1^2}{2}
\end{align*}
\]

**Proof.** First, observe that since \(\pi_1\) and \(\pi_2\) are i.i.d., we have \(\sigma_1^2 = \sigma_2^2\). Also, from the scheduling policy \((1)\) observe that queue 1 is served if and only if queue 2 is served. Therefore, by symmetry, \(q_1\) and \(q_2\) have the same distribution. Hence,

\[
\begin{align*}
\lim_{\epsilon \downarrow 0} \epsilon E[q_1] &= \lim_{\epsilon \downarrow 0} \epsilon E[q_2] \quad \text{and} \quad \lim_{\epsilon \downarrow 0} E[q_1^+ \pi_2] = \lim_{\epsilon \downarrow 0} E[q_2^+ \pi_1]
\end{align*}
\]

If we use these conditions in the system of equations from Theorem 3.1 we obtain the following system of equations

\[
\begin{align*}
\lim_{\epsilon \downarrow 0} \epsilon E[q_1] &= \frac{5\sigma_1^2 + \sigma_3^2}{6} - \lim_{\epsilon \downarrow 0} E[q_1^+ \pi_2] \quad \text{(14)} \\
\lim_{\epsilon \downarrow 0} \epsilon E[q_1] &= -\frac{4\sigma_1^2 + \sigma_3^2}{6} + 2 \lim_{\epsilon \downarrow 0} E[q_1^+ \pi_2], \quad \text{(15)}
\end{align*}
\]

which has two equations and two variables. Therefore, we can solve it. Taking \(2(14) + (15)\) we obtain

\[
\lim_{\epsilon \downarrow 0} \epsilon E[q_1^+] = \frac{2\sigma_1^2 + \sigma_3^2}{6} \quad \text{(16)}
\]

and using this result in equation (14) we obtain

\[
\lim_{\epsilon \downarrow 0} E[q_1^+ \pi_2] = \frac{\sigma_1^2}{2}.
\]

Finally, recall that \(q_3 = q_1 + q_2\) and that SSC implies

\[
\lim_{\epsilon \downarrow 0} \epsilon E[q_i] = \lim_{\epsilon \downarrow 0} \epsilon E[q_i] \quad \forall i \in \{1, 2, 3\}.
\]

Therefore,

\[
\begin{align*}
\lim_{\epsilon \downarrow 0} \epsilon E[q_3] &= \lim_{\epsilon \downarrow 0} \epsilon E[q_3] \\
&= \lim_{\epsilon \downarrow 0} \epsilon E[q_1 + q_2] \\
&= \lim_{\epsilon \downarrow 0} \epsilon E[q_1 + q_2] \\
&= \frac{2\sigma_1^2 + \sigma_3^2}{3}
\end{align*}
\]

where the last equality holds by equation (16). \(\square\)
When the CRP condition is satisfied, it has been shown that the Drift method can be used to obtain all the moments of the scaled queue lengths [2]. This is done using test functions of the form \( \|\varphi\|^m \) for \( m = 1, 2, \ldots \) inductively, where \( \varphi \) represents the projection of vector of queue lengths (of the corresponding queueing system) on the subspace where SSC occurs. Since this subspace is one-dimensional under the CRP condition, \( \|\varphi\|^m \) is a monomial of degree \( m \). In general, if the test function is a monomial of degree \( m \), we obtain the \((m - 1)\)th moment. Under certain conditions, the moments of a random variable completely determine the joint distribution. This is known as the moments problem \([12]\). Therefore, if these (mild) conditions are satisfied, the Drift method can be used to obtain the joint distribution of queue lengths in heavy-traffic. On the other hand, if SSC occurs onto a multidimensional subspace, we need the moments of all the linear combinations of the elements of the projection of the queue lengths on the subspace where SSC occurs in order to determine the joint distribution of queue lengths (if the conditions of the moments problem are satisfied). Therefore, we need to inductively take polynomial test functions of degree \( m = 1, 2, \ldots \) in the next theorem, we present a system of equations on the second degree monomials of \( \|\varphi\| \).

**Theorem 3.3.** Consider the three-queue system described in Theorem 3.1. Then, the following system of equations must be satisfied.

\[
\begin{align*}
\lim_{\epsilon \to 0} \epsilon^2 E \left[ q_1^2 \right] &= \left( \frac{4\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_1 \right] - \lim_{\epsilon \to 0} \epsilon E \left[ (q_1)^2 \right] \\
\lim_{\epsilon \to 0} \epsilon^2 E \left[ q_2^2 \right] &= \left( \frac{\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_2 \right] - \lim_{\epsilon \to 0} \epsilon E \left[ (q_2)^2 \right] \\
\lim_{\epsilon \to 0} \epsilon^2 E \left[ q_1^2 + 2q_2^2 \right] &= \left( \frac{4\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_2 \right] + \left( \frac{-4\sigma_1^2 - 4\sigma_2^2 + 2\sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_1 \right] + 2 \lim_{\epsilon \to 0} \epsilon E \left[ (q_1)^2 \right] \\
\lim_{\epsilon \to 0} \epsilon^2 E \left[ q_2^2 + 2q_2^2 \right] &= \left( \frac{\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_1 \right] + \left( \frac{-4\sigma_1^2 - 4\sigma_2^2 + 2\sigma_3^2}{3} \right) \lim_{\epsilon \to 0} \epsilon E \left[ q_2 \right] + 2 \lim_{\epsilon \to 0} \epsilon E \left[ (q_2)^2 \right]
\end{align*}
\]

where we omitted the dependence on \( \epsilon \) of the variables for ease of exposition.

The proof of Theorem 3.3 follows after setting to zero the drift of the following test functions

\[ V_1(q) = \|q\|_1^3, \quad V_2(q) = \|q\|_2^3, \quad V_3(q) = q_1^2 \|q\|_2, \quad \text{and} \quad V_4(q) = q_2^2 \|q\|_2, \]

which are all the degree 3 monomials on \( q\|_1 \) and \( q\|_2 \). After setting to zero the drift of these test functions, the proof follows using the arguments similar to those in the proof of Theorem 3.1, and so we omit it.

Observe that this system of equations is also under-determined and it uses the variables from the system of equations presented in Theorem 3.1. Therefore, in order to solve it, we need to first solve the system presented in Theorem 3.1. If we add the symmetry condition of Corollary 3.2 we have a value for the variables from the system of equations presented in Theorem 3.1. However, if we add this condition then equations (17) and (18) are equivalent, and equations (19) and (20) are equivalent too. Then, we obtain a system of two equations and three variables, meaning \( \lim_{\epsilon \to 0} \epsilon^2 E \left[ q_1^2 \right], \lim_{\epsilon \to 0} \epsilon^2 E \left[ q_1 q_2 \right] \) and \( \lim_{\epsilon \to 0} \epsilon E \left[ (q_1)^2 \right] \). Therefore, even under symmetric traffic to queues 1 and 2, the Drift method with polynomial test functions does not provide enough information to find higher moments of the queue lengths.

In general, to obtain all the linear combinations of the \( m \)th moment of the scaled queue lengths we take polynomial test functions of degree \( m + 1 \). There are \( m + 2 \) monomials among all \( (m + 1) \) degree polynomials of \( q\|_1 \) and \( q\|_2 \) and they are \( q_k \|^m \) for \( k = 0, 1, \ldots, m + 1 \). This leads to \( m + 2 \) equations. On the other hand, we have \( m + 3 \) ‘new’ variables. We say a variable is ‘new’ for the system of equations that arises after setting to zero the monomials of degree \( m + 1 \) if it does not appear in any system of equations of degree \( \ell < m + 1 \). For example, the ‘new’ variables in the system of equations of Theorem 3.3 are \( \lim_{\epsilon \to 0} \epsilon^2 E \left[ q_1^2 \right], \)

\[9\]
lim_{\epsilon \downarrow 0} \epsilon^2 E \left[ 2 \bar{q}_2^2 \right], \lim_{\epsilon \downarrow 0} \epsilon^2 E \left[ \bar{q}_1 \bar{q}_2 \right], \lim_{\epsilon \downarrow 0} \epsilon E \left[ \left( \bar{q}_1^+ \right)^2 \bar{q}_2 \right] \text{ and } \lim_{\epsilon \downarrow 0} \epsilon E \left[ \left( \bar{q}_2^+ \right)^2 \bar{q}_1 \right]. \text{ In general, the } (m+3) \text{ 'new' variables are the } (m+1) \text{ monomials in queue length of degree } m, \text{ and the two variables, } \lim_{\epsilon \downarrow 0} \epsilon E \left[ \left( \bar{q}_1^+ \right)^m \bar{q}_2 \right] \text{ and } \lim_{\epsilon \downarrow 0} \epsilon E \left[ \left( \bar{q}_2^+ \right)^m \bar{q}_1 \right]. \text{ It can be shown that the terms involving all other powers of } \bar{q}_1 \text{ and } \bar{q}_2 \text{ vanish.}

If we add the symmetric traffic condition, i.e. that } \bar{q}_1 \text{ and } \bar{q}_2 \text{ are equidistributed, to the system of equations that arises after setting to zero the drift of all monomials of degree } m + 1 \text{ we obtain a system of } \frac{m + 2}{2} \text{ equations and } \frac{m + 2}{2} + 1 \text{ 'new' variables if } m \text{ is even, and a system of } \frac{m + 3}{2} \text{ equations and } \frac{m + 3}{2} \text{ 'new' variables if } m \text{ is odd. However, this does not mean that we can find a solution for the systems of equations where } m \text{ is odd, because they contain 'old' (not 'new') variables from the systems of equations of degree } \ell \leq m, \text{ where } \ell \text{ can be even. Thus, we need to find additional independent equations that the system satisfies in order to solve for all the variables, and this is the key challenge in pushing forward the Drift method. One possible way of doing this is to use the MGF method [4], which is a natural generalization of the Drift method.}

4 Bounds on the Mean Queue Length

In Section 3 we presented linear systems of equations that the moments of vector of queue lengths must satisfy in heavy-traffic. In this section we use these systems of equations to obtain bounds on linear combinations of the expected queue lengths in heavy-traffic. A similar approach was studied by [7] and [1], where an under-determined set of linear systems of equations was obtained and linear programming was used to obtain bounds. However, the focus in those papers was on queueing networks under fixed arrival and service rates, as opposed to the heavy-traffic analysis in the current paper.

In the next theorem we provide an upper and a lower bound for the heavy-traffic limit of the expected value of any linear combination of the queue lengths in a three-queue system.

**Theorem 4.1.** Let

\[
P = \left\{ (x_1, x_2, y_1, y_2) \in \mathbb{R} : \begin{align*}
    x_1 + y_1 &= \frac{4\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2} \\
    x_2 + y_2 &= \frac{6}{\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2} \\
    x_1 + x_2 - 2y_1 - 2y_2 &= \frac{-2\sigma_1^2 - 2\sigma_2^2 + \sigma_3^2}{3} \\
    x_1, x_2, y_1, y_2 &\geq 0
\end{align*} \right\}.
\]

For } \alpha, \beta \in \mathbb{R}, \text{ define } f(\alpha, \beta) = \min\{\alpha x_1 + \beta x_2 : \exists (y_1, y_2) \text{ such that } (x_1, x_2, y_1, y_2) \in P\} \text{ and } T(\alpha, \beta) = \max\{\alpha x_1 + \beta x_2 : \exists (y_1, y_2) \text{ such that } (x_1, x_2, y_1, y_2) \in P\}. \text{ Then,}

\[
f(\alpha, \beta) \leq \lim_{\epsilon \downarrow 0} \epsilon E [\alpha \bar{q}_1 + \beta \bar{q}_2] \leq T(\alpha, \beta),
\]

where } \bar{q}_1, \bar{q}_2 \text{ and } \epsilon \text{ are defined as in Theorem 3.1. Furthermore, for any } B \in \mathbb{R}_+, \text{ }

\[
P \left\{ \lim_{\epsilon \downarrow 0} \epsilon (\alpha \bar{q}_1 + \beta \bar{q}_2) \geq B \right\} \leq \frac{T(\alpha, \beta)}{B}.
\]

**Proof.** If let

\[
    x_1 = \lim_{\epsilon \downarrow 0} \epsilon E [\bar{q}_1],
\]

\[
    x_2 = \lim_{\epsilon \downarrow 0} \epsilon E [\bar{q}_2],
\]

\[
    y_1 = \lim_{\epsilon \downarrow 0} \epsilon E [\bar{q}_1 \bar{q}_2],
\]

\[
    y_2 = \lim_{\epsilon \downarrow 0} \epsilon E [\bar{q}_2^2 \bar{q}_1],
\]
then the proof follows from Theorem 3.1, because the set \( P \) represents the system of equations presented there.

Also, from Markov’s inequality we know
\[
P \left\{ \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} (\alpha \overline{q}_1 + \beta \overline{q}_2) \geq B \right\} \leq \frac{\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\alpha \overline{q}_1 + \beta \overline{q}_2]}{B}.
\]

Particular cases of Theorem 4.1 are when \( \alpha = 0 \) or \( \beta = 0 \), where we obtain bounds on \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_2] \) and \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1] \), respectively.

Theorem 4.1 gives explicit bounds for all linear combinations of the expected scaled queue lengths. Similar linear programs can be written to obtain bounds on higher moments, and consequently tighter tail probabilities.

5 Discussion

In this section we discuss the results presented in Sections 3. We proved that using the Drift method with polynomial test functions does not provide enough information to find the heavy-traffic limit of the expected scaled queue lengths. In particular, in Theorem 3.1 we proved that setting to zero the drift of all monomials of degree 2 leads to a system of 3 equations in 4 variables. Therefore, the solution is not unique. However, in Proposition 2.2 we provide a result from [9], where they compute \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1 + \overline{q}_2 + \overline{q}_3] \) using a quadratic test function in the Drift method. Now we show that this result follows from Theorem 3.1.

From SSC we know
\[
\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1 + \overline{q}_2 + \overline{q}_3] = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1 + \overline{q}_2 + \overline{q}_3] \]
\[
= \frac{2}{3} (\sigma_1^2 + \sigma_2^2 + \sigma_3^2)
\]
where (a) follows from the fact that \( \overline{q}_3 = \overline{q}_1 + \overline{q}_2 \) since \( q_\parallel \) is defined to be in the region of SSC defined in (5). The last equality follows from the following linear combination of the equations in Theorem 3.1: \( 2(8)+2(9)+(10) \), because the variables \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1 \overline{q}_2] \) and \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_3 \overline{q}_1] \) cancel out. In fact considering the linear combination \( 2(8)+2(9)+(10) \) is equivalent to using the test function
\[
2\overline{q}_1^2 + 2\overline{q}_2^2 + 2\overline{q}_1 \overline{q}_2 = \overline{q}_1^2 + \overline{q}_2^2 + (\overline{q}_1 \overline{q}_2)^2
\]
\[
= (\overline{q}_1^2 + \overline{q}_2^2 + \overline{q}_3^2)
\]
where (a) follows from the definition of the region of SSC, (5). This matches with the test function used by [9]. In fact this test function was chosen by [9] for the primary reason of canceling out the terms \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_1 \overline{q}_2] \) and \( \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} [\overline{q}_3 \overline{q}_1] \).

To obtain other linear combinations of the expected heavy-traffic scaled queue lengths we need to actually work with all the variables of the system of equations. Therefore, we need an additional equation. To better understand this argument, consider a tandem queue system with memoryless interarrival and service times in any (not necessarily heavy) traffic. We know that the steady-state joint distribution is product of two geometries, and can be obtained using reversibility arguments. Using the drift approach described above, we again get 3 equations in 4 unknowns. However, in addition to the drift arguments, if we use reversibility to
separately prove that the queues are independent in steady-state and impose it as an additional condition, we can solve for all the unknowns.

One way to add an equation is to assume symmetric traffic, as we did in Corollary 3.2. However, this only solves the system for a particular case and it does not allow us to obtain higher moments. As shown in the discussion after Theorem 3.3, if we add the same assumption to the system of equations for moments of degree 2, we obtain an under-determined system of equations and, therefore, we cannot find a unique solution.

5.1 Generalization to other queueing systems

In Section 3 we focused on a three-queue system in heavy-traffic. We chose this system because it is one of the simplest queueing systems where the CRP condition is not satisfied. However, the same approach can be applied to any queueing system where the CRP condition is not met. In this subsection we discuss what happens in a more general case.

[2] showed how to compute the moments of $\|q\|$ using the Drift method in queueing systems that satisfy the CRP condition. In this case, setting to zero the drift of $V(q) = \|q\|^{m+1}$ in steady-state and using SSC allows to compute the $m$th moment because of the following reason. When one sets to zero the drift of $V(q)$, terms of the form $q_i^+u_i$ arise and, since $q_i^+$ and $u_i$ belong to the same one-dimensional subspace, these terms can be approximated by $q_i^+u_i$, which is zero by definition of unused service.

On the other hand, if the CRP condition is not satisfied, then $q_i$ lives in a $d$-dimensional subspace, where $d > 1$. In this case, for each $i$, $q_i^+u_i$ cannot be approximated by $q_i^+u_i$ because the vectors $q_i^+$ and $u_i$ live in a $d$-dimensional subspace. Therefore, in heavy-traffic we only have the approximation (with some abuse of notation) $q_i^+u_i \approx q_i^+(u_{k_1} + u_{k_2} + \cdots + u_{k_d})$, where $k_1, \ldots, k_d$ represent the $d$ dimensions that characterize SSC. In other words, crossed-terms arise exactly as the ‘qu’ terms in Theorems 3.1 and 3.3 for the three-queue system. In the following analysis we present the number of equations and variables that appear in a general queueing system with $d$-dimensional SSC.

In order to obtain the $m$th moment of the queue lengths, we should construct a system of equations that yields from setting to zero the drift of all monomials of degree $m+1$. Since SSC occurs into a $d$-dimensional subspace, we need to consider all possible monomials of degree $m+1$ in $d$ variables. Setting to zero the drift of each monomial will lead to an equation, so we will have $\binom{m+d}{d-1}$ equations. Now we count the number of ‘new’ variables with respect to the system of equations that arises after setting to zero the drift of monomials of degree $k$, for all $k \leq m$. Here we use the same notion of ‘new’ and ‘old’ variables that we introduced in the discussion after Theorem 3.3. Observe that there are two types of ‘new’ variables that do not vanish in the heavy-traffic limit. On one hand, we have the heavy-traffic limit of the expected value of products of the elements of $q_i$ and, on the other hand, we have the heavy-traffic limit of the expected value of the product between the elements of $q_i$ and of the vector of unused service. We will call them the ‘q’ variables and the ‘qu’ variables, respectively. Specifically, the ‘q’ variables are all monomials of degree $m$ in $d$ variables, so there are $\binom{m+d}{d-1}$ ‘q’ variables. The ‘qu’ that do not vanish in heavy-traffic are of degree $m$ in ‘q’ and degree 1 in ‘u’. Also, the element corresponding to the unused service vector has to be different to the elements of the vector of queue lengths because the product between the queue length and the unused service of the same queue is zero by definition of unused service. Therefore, for each element of $u_i$ we need to consider all possible combinations of ‘q’s, i.e. all monomials of degree $m$ in $d - 1$ variables. Therefore, there are $d\binom{m+d-2}{d-2}$ ‘qu’ variables. Thus, in total we have $\binom{m+d-1}{d-1} + d\binom{m+d-2}{d-2}$ variables and this number is larger than the number of equations.

Summarizing, if we use the method introduced in Section 3 to compute the $m$th moment of the queue lengths of a queueing system that experiences $d$-dimensional SSC, we obtain a system of equations of $(\binom{m+d}{d-1})$ equations and $(\binom{m+d-1}{d-1}) + d\binom{m+d-2}{d-2}$ variables. Therefore, it is under-determined. In other words, we need extra equations to find a unique solution to this system of equations. This analysis shows that the issues illustrated in Theorems 3.1 and 3.3 arise in any queueing system with multidimensional SSC.
6 Conclusion and Future Work

In this paper we present an alternative way to use the Drift method, that allows us to avoid the guesswork required to come up with the right test function. We propose that, instead of coming up with a polynomial test function, we can set to zero the drift of all the monomials and write these conditions as a linear system of equations. Then, after solving the system we can obtain any linear combination of the moments of the queue lengths in heavy-traffic. However, we proved that polynomial test functions do not provide all the information that is required to find these moments. The proof consists of showing that setting to zero the drift of the monomials leads to an under-determined system of equations.

Future work is to add constraints to the system of equations, so that we can solve it. One way to do this is to write an equation that represents an additional constraint of the system (such as reversibility in the example of a tandem queue that we discussed in Section 5). Another perspective is to consider a different class of test functions. For example, [4] explored the use of exponential test functions to obtain the Moment Generating Function of the queue lengths in heavy-traffic.

References

[1] Dimitris Bertsimas, Ioannis Ch Paschalidis, and John N Tsitsiklis. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability*, pages 43–75, 1994.

[2] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012. ISSN 0257-0130.

[3] J. Michael Harrison. *Brownian Models of Performance and Control*. Cambridge University Press, 2013. doi: 10.1017/CBO9781139087698.

[4] Daniela Hurtado-Lange and Siva Theja Maguluri. Transform methods for heavy-traffic analysis. *arXiv preprint arXiv:1811.05595*, 2018.

[5] WN Kang and RJ Williams. Diffusion approximation for an input-queued switch operating under a maximum weight matching policy. *Stochastic Systems*, 2(2):277–321, 2012.

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

[7] S. Kumar and P. R. Kumar. Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 39(8):1600–1611, Aug 1994. ISSN 0018-9286. doi: 10.1109/9.310033.

[8] 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. URL http://dx.doi.org/10.1214/15-SSY193.

[9] 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.

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

[11] Devavrat Shah, JohnN. Tsitsiklis, and Yuan Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3-4):375–384, 2011. ISSN 0257-0130.

[12] James Alexander Shohat and Jacob David Tamarkin. *The problem of moments*. Number 1. American Mathematical Soc., 1943.
[13] 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.

[14] 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. ISSN 0163-5999. doi: 10.1145/3199524.3199565. URL http://doi.acm.org/10.1145/3199524.3199565.