ASYMPTOTIC ANALYSIS OF SINGLE-HOP STOCHASTIC PROCESSING NETWORKS USING THE DRIFT METHOD

A Dissertation
Presented to
The Academic Faculty

By

Daniela Hurtado-Lange

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Engineering
Department of Industrial and Systems Engineering

Georgia Institute of Technology

December 2021

ASYMPTOTIC ANALYSIS OF SINGLE-HOP STOCHASTIC PROCESSING NETWORKS USING THE DRIFT METHOD

Thesis committee:

Dr. Siva Theja Maguluri
Department of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Debankur Mukherjee
Department of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. J.G. "Jim" Dai
School of Operations Research and Information Engineering
Cornell University

Dr. Devavrat Shah
School of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Dr. Robert Foley
Department of Industrial and Systems Engineering
Georgia Institute of Technology

Date approved: October 15th, 2021

You may not always have a comfortable life and you will not always be able to solve all of the world's problems at once, but don't ever underestimate the importance you can have because history has shown us that courage can be contagious and hope can take on a life of its own.

Michelle Obama

For Paulina, Manuel, Bárbara, Tomás, Sebastián and Spin

ACKNOWLEDGMENTS

First, I must thank my family for their love and support. I wouldn't have made it this far without them. Specifically, I thank Spin, Sebastián, Hineva and Dayna for their unconditional love and support when things got intense. I must also thank my friends, old and new, who have helped keep me sane during the last five years, especially during this pandemic.

I'm also thankful to those in the Applied Probability community with whom I haven't yet worked directly, but who have nonetheless encouraged and supported me along the way. Special thanks also to the ISyE community at Georgia Tech, without whom none of this would have been possible.

This work was funded through an NSF grant, and the Tennenbaum fellowship. I also thank ANID Becas Chile for supporting my studies.

Special thanks to my committee members, Dr. J.G. "Jim" Dai, Dr. Robert Foley, Dr. Debankur Mukherjee and Dr. Devavrat Shah , for their advice and knowledge these past several years. Their suggestions and comments have made this work possible.

Lastly, but certainly not least, thank you to Dr. Siva Theja Maguluri for everything. He has provided excellent guidance, support, and mentoring. I can't describe how thankful I am that I was lucky enough to be part of his group.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Today's era of cloud computing and big data is powered by massive data centers. The focus of my dissertation is on resource allocation problems that arise in the operation of these large-scale data centers. Analyzing these systems exactly is usually intractable, and a usual approach is to study them in various asymptotic regimes with heavy traffic being a popular one. We use the drift method, which is a two-step procedure to obtain bounds that are asymptotically tight. In the first step, one shows state-space collapse, which, intuitively, means that one detects the bottleneck(s) of the system. In the second step, one sets to zero the drift of a carefully chosen test function. Then, using state-space collapse, one can obtain the desired bounds.

This dissertation focuses on exploiting the properties of the drift method and providing conditions under which one can completely determine the asymptotic distribution of the queue lengths. In chapter 1 we present the motivation, research background, and main contributions.

In chapter 2 we revisit some well-known definitions and results that will be repeatedly used in the following chapters.

In chapter 3, chapter 4 and chapter 5 we focus on load-balancing systems, also known as supermarket checkout systems. In the load-balancing system, there are a certain number of servers, and jobs arrive in a single stream. Once they come, they join the queue associated with one of the servers, and they wait in line until the corresponding server processes them.

In chapter 3 we introduce the moment generating function (MGF) method. The MGF, also known as two-sided Laplace form, is an invertible transformation of the random variable's distribution and, hence, it provides the same information as the cumulative distribution function or the density (when it exists). The MGF method is a two-step procedure to compute the MGF of the delay in stochastic processing networks (SPNs) that satisfy the complete resource pooling (CRP) condition. Intuitively, CRP means that the SPN has a

single bottleneck in heavy traffic.

A popular routing algorithm is power-of-d choices, under which one selects d servers at random and routes the new arrivals to the shortest queue among those d. The power-of-d choices algorithm has been widely studied in load-balancing systems with homogeneous servers. However, it is not well understood when the servers are different. In chapter 4 we study this routing policy under heterogeneous servers. Specifically, we provide necessary and sufficient conditions on the service rates so that the load-balancing system achieves throughput and heavy-traffic optimality. We use the MGF method to show heavy-traffic optimality.

In chapter 5 we study the load-balancing system in the many-server heavy-traffic regime, which means that we analyze the limit as the number of servers and the load increase together. Specifically, we are interested in studying how fast the number of servers can grow with respect to the load if we want to observe the same probabilistic behavior of the delay as a system with a fixed number of servers in heavy traffic. We show two approaches to obtain the results: the MGF method and Stein's method.

In chapter 6 we apply the MGF method to a generalized switch, which is one of the most general single-hop SPNs with control on the service process. Many systems, such as ad hoc wireless networks, input-queued switches, and parallel-server systems, can be modeled as special cases of the generalized switch.

Most of the literature in SPNs (including the previous chapters of this thesis) focuses on systems that satisfy the CRP condition in heavy traffic, i.e., systems that behave as single-server queues in the limit. In chapter 7 we study systems that do not satisfy this condition and, hence, may have multiple bottlenecks. We specify conditions under which the drift method is sufficient to obtain the distribution function of the delay, and when it can only be used to obtain information about its mean value. Our results are valid for both, the CRP and non-CRP cases and they are immediately applicable to a variety of systems. Additionally, we provide a mathematical proof that shows a limitation of the drift method.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Introduction

A stochastic processing network (SPN) is a system that receives job requests and has servers that complete them. Typically, the interarrival and processing times are random variables (thus the name *stochastic*). An important example is the use of the internet. Every time we use a smartphone app, purchase something online, or do a web search, we request service from a data center. Given the ubiquitous use of the internet, data centers must process a massive amount of data every second to fulfill every user's requirement, and minimizing response time becomes essential. Studies show that if an Amazon's website is five milliseconds slower than its competition, it can lose up to $4 billion per millisecond in revenues [1]. Hence, a crucial problem in the operation of data center networks is to minimize the delay. In this thesis, the main focus is on understanding delay from a queue length perspective. Specifically, we analyze the behavior of the queue lengths in several SPNs.

Exact analysis of queueing systems that arise in the study of SPNs is usually intractable, so it is common to analyze them in various asymptotic regimes to get insights on their behavior. A very popular regime in the literature is the heavy-traffic regime, where the system is loaded very close to its maximum capacity. This regime is sometimes called the classical or conventional heavy-traffic regime. One of the advantages of the heavy-traffic limit is that many queueing systems behave as if they live in a much lower dimensional subspace of the state space in the limit. This phenomenon is known as state space collapse (SSC). If the heavy-traffic limit is taken such that SSC occurs into a line, then the system is said to satisfy the Complete Resource Pooling (CRP) condition and, intuitively, it behaves

as a single server queue in the limit [2, 3, 4].

Most of the literature on heavy-traffic analysis is on systems that satisfy the CRP condition. A popular framework is the diffusion limits approach [2, 3, 5, 6, 7, 8, 9]. In this approach, the scaled queue lengths are shown to converge to a reflected brownian motion (RBM) process, and then the steady-state behavior of this RBM is studied using SSC. The last step is to show interchange of limits, which is usually challenging. Under the CRP condition, one obtains an RBM on a line, which is well understood. However, a major challenge is in using this program for SPNs where the CRP condition is not satisfied (i.e., when there are multiple resources that are simultaneously in heavy traffic). In such case, one needs to solve for the steady-state distribution of a RBM in a multidimensional subset of $\mathbb{R}^n$, and this is not known in general, as shown by [10].

More recently, three 'direct methods' have been proposed to perform heavy-traffic analysis [11]. In these approaches, there is no need to show interchange of limits, as one directly works with the queue length process instead of working with an RBM. The methods are: (i) The BAR approach, (ii) the drift method, and (iii) Stein's method. We present an overview of these methods and of the diffusion limits approach in section 1.3.

As opposed to the classical heavy-traffic regime, where the number of servers is fixed, in the many-server heavy-traffic regime both, the load and the number of servers increase together. A popular SPN in the many-server heavy-traffic literature is the load balancing system, also known as supermarket checkout system. In the load balancing system there are servers with infinite buffers, and the jobs arrive in a single stream. Upon arrival, they must be routed to one of the servers. There are many routing algorithms, but in this thesis we work with join the shortest queue (JSQ) and power-of-d choices (also known as JSQ(d)). Under JSQ, the new arrivals are routed to the queue with the least number of jobs in line, and under power-of-d choices, they are routed to the shortest queue among d servers that are sampled uniformly at random. A formal definition of the load balancing system and these routing algorithms is presented in chapter 3.

Depending on how fast the load increases with respect to the number of servers, the queue lengths in a load balancing system exhibit different behaviors. In section 5.2 we discuss some well-known results in this area.

In this thesis, we focus on heavy-traffic analysis of single-hop SPNs with control in the arrival or the service process (not both). In particular, we study load balancing systems as described above (control in the arrival process), and generalized switches (control in the service process). The generalized switch is a model that was first introduced in [8], and subsumes several SPNs that are of practical interest, such as ad hoc wireless networks, wireless networks in presence of fading, input-queued switches, and parallel-server systems. A formal definition of the generalized switch and these specific systems is presented in chapter 6 and chapter 7.

For systems that satisfy the CRP condition, we propose a transform method based on drift arguments that yields the heavy-traffic distribution of the queue lengths in two steps. Additionally, we study the essential question of throughput optimality in load balancing systems with heterogeneous servers. We show these results in chapter 3, chapter 4 and chapter 6.

Additionally, we study the load balancing system in the many-server heavy-traffic regime. Our goal is to determine how fast the load has to increase with respect to the number of servers to obtain the heavy-traffic behavior of the average queue length. In chapter 5 we show that the answer depends on the routing policy. Additionally, we provide two proofs of the result: one using the transform method mentioned above, and one using Stein's method.

According to [10, 12], one of the simplest queueing systems where the CRP condition is not satisfied is an input-queued switch, and [13] identifies it as a focus of study in the SPN literature since it serves as a guiding example to study more general systems that do not satisfy CRP. Recently, the drift method was used to characterize the heavy-traffic scaled sum of queue lengths in input-queued switches [14, 15], solving a question that remained open for over a decade. In chapter 7 we generalize this result to the generalized

switch, and we show the power and flexibility of the result in several cases of SPNs that do not satisfy the CRP condition. Further, some of the corollaries of the main theorem solve open questions. We additionally show a limitation of the drift method with polynomial test functions. Specifically, we show that we can only obtain certain linear combinations of the queue lengths.

In the next section we describe the main contributions of this thesis, and in section 1.3 we provide an overview of the methods for heavy-traffic analysis that we briefly introduced above. We finalize this chapter establishing the notation that we use in the remainder of this document, including the notation for a general queueing system.

1.2 Main contributions

In this section we present an overview of the main contributions of this thesis.

1.2.1 Transform method

In chapter 3 we develop the transform 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 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 immediately yields the stationary distribution.

To introduce the method and highlight the main steps, we present a sketch of the MGF method in section 2.5 in the case of a single server queue operating in discrete time. Then, in chapter 3 we formalize the framework and we apply it to a load balancing system operating under join the shortest queue [16, 17] and power-of-2 choices [18, 19, 20].

In chapter 6 we study generalized switches [8] under the CRP condition, operating under MaxWeight scheduling algorithm [21]. We also show that an ad hoc wireless networks

and an input-queued switch operating under MaxWeight scheduling algorithm satisfy our assumptions. All these systems are assumed to satisfy the CRP condition, and they are operated under algorithms that ensure that SSC occurs into a one-dimensional subspace. We show that the stationary distribution of this one-dimensional component is exponential.

1.2.2 Power-of-d choices with heterogeneous servers

In the literature, most of the work on the power-of-d choices algorithm assumes homogeneous servers. In chapter 4 we provide necessary and sufficient conditions on the vector of service rates that ensure that power-of-d choices is throughput optimal for the load balancing system. Specifically, we show that power-of-d choices is throughput optimal if and only if the vector of service rates normalized by the total service rate is majorized by a vector $\boldsymbol{\nu}$ where the i^{th} element represents the probability of routing the new arrivals to the i^{th} longest queue. This notion of majorization determines a polytope $\mathcal{M}^{(d)}$ where the service-rate vector should lie.

Additionally, we show that the load balancing system operating under power-of-d choices satisfies the CRP condition if the service-rate vector lies in the interior of the polytope $\mathcal{M}^{(d)}$. Hence, under this condition, power-of-d choices is heavy-traffic optimal and we show that the vector of queue lengths converges in distribution to a vector of exponential random variables. We use the transform method introduced in chapter 3 to prove the last result.

The third contribution of this chapter is a set of sufficient conditions to ensure throughput optimality of any routing policy that samples a subset of servers at random, and routes to the shortest queue from the set.

1.2.3 Load balancing system in the many-server heavy-traffic regime

In chapter 5 we focus on the many-server heavy-traffic regime of the load balancing system. In this regime, the load and the number of servers increase together, as opposed to the

classical heavy-traffic regime, where the number of servers is fixed. The goal of this chapter is to characterize how fast can the load grow with respect to the number of servers to observe the classical heavy-traffic behavior, i.e., to obtain convergence in distribution of the average queue length to an exponential random variable.

We obtain conditions on the parameters of the system for JSQ and power-of-d choices routing. We provide two proofs of our result: one using transform method and one using Stein's method, where we additionally obtain the rate of convergence in Wasserstein's distance.

These proofs are empowered by multiplicative SSC, where we obtain error bounds that increase with the number of servers, but become negligible with respect to the average queue length. A key component of our proof that is novel relative to prior literature, is the use of Stein's method in the presence of a multiplicative SSC.

To prove our result using the transform method, we generalize the framework from the classical heavy-traffic regime to the many-server heavy-traffic regime. Additionally, we work with both, a discrete-time and a continuous-time system. Hence, we also generalize the transform method to be used in SPNs modeled in continuous time.

To finalize the chapter, we show the rate of convergence in mean of the average queue length. In the previous results, we show convergence in distribution. However, convergence in distribution does not necessarily imply convergence in expected value. We pursue this additional step and we use the drift method in our proof.

1.2.4 Generalized switch with no complete resource pooling

In chapter 7 we study one of the most general single-hop SPNs with control in the service process: the generalized switch. We study the heavy-traffic limit of the expected queue lengths without assuming that the CRP condition is satisfied. Indeed, our results are valid in the case of one-dimensional or multi-dimensional SSC.

The main contribution of chapter 7 is the characterization of the heavy-traffic scaled

mean of certain linear combinations of the queue lengths in a generalized switch. Since the generalized switch model subsumes several SPNs, we additionally show how to apply the main theorem to specific systems such as the input-queued switch, ad hoc wireless networks, parallel-server systems, and the so-called $\mathcal{N}$-system. Additionally, we show that if SSC occurs into a full-dimensional subspace, the correlation between the arrival processes does not affect the mean of the linear combination of the queue lengths. These results are contributions on themselves, even though their proof is a simple application of the main theorem.

In the last part of chapter 7 we present an alternate view of the drift method. Traditionally, the choice of the test function plays a key role in the drift method, and it is known that using polynomial test functions yields heavy-traffic results for the moments of the queue lengths. We propose that, instead of searching for the right polynomial test function, we can use all the possible monomials of a certain degree as test functions, and build a linear system of equations by setting their drift to zero. We show that, when the CRP condition is not satisfied, this system of equations is under-determined and, hence, one cannot solve for the individual mean queue lengths. However, if one takes the right linear combination of the equations, some of the variables cancel out and we obtain an explicit expression.

This system of equations explains the success of the drift method. While it is known that it is notoriously hard to solve the stationary distribution of a multidimensional RBM, it has been a little surprising that simple drift-based arguments give the mean of sum of the queue lengths in several systems, such as the ones studied in [14, 15, 22]. The system of equations shows that due to the difficulty of the underlying problem, it is not possible to get all the mean queue lengths individually. However, because of the structure of the system of equations, it is possible to obtain certain linear combinations. In the case of input-queued switch and the bandwidth sharing system, the sum of the queue lengths was one of the linear combinations that is easy to obtain.

Even though one cannot explicitly solve for the individual queue lengths, we propose

to use the under-determined system of equations as the feasible region of a linear program to obtain bounds. We present the details of this approach and numerical results.

1.3 Diffusion limits and direct methods for heavy-traffic analysis

One of the most famous approaches for heavy-traffic analysis is diffusion limits. In this framework, one scales time and space, and proves process level convergence to a Reflected Brownian Motion (RBM) process in heavy traffic. Then, the stationary distribution of the RBM is computed. The last step is to prove interchange of limits, which is challenging in many cases because one needs to prove tightness. Kingman was the first one to use this approach, and he computed the stationary distribution of the scaled waiting time in a $G/G/1$ queue [23]. The same approach has been used in a variety of queueing systems that satisfy the CRP condition [2, 5, 6, 7, 9, 8]. When the CRP condition is not satisfied, one needs to solve a multi-dimensional RBM and this is a major challenge that has not been solved [10].

More recently, three methods that do not require the interchange of limits step have been developed. They are known as 'direct' methods [11], and they are the BAR method, Stein's method and the drift method. BAR stands for Basic Adjoint Relationship, which is an integral equation that must be satisfied by a Markov process. The BAR method is developed in [24] in the context of a Generalized Jackson Network. It is shown that a sequence of one-sided Laplace transforms asymptotically satisfies the BAR equation. The authors in [24] work with a continuous time Markov process, so one of the challenges in their proof is to handle the jumps of the queue lengths process. They overcome this difficulty by carefully choosing an exponential test function that helps to eliminate the jump term associated with the BAR.

Stein's method for analyzing SPNs was first introduced in [25], and it has now emerged as a simple yet powerful method that can be used not only to show asymptotic convergence, but also to bound the rate of convergence in Wasserstein's distance. It has become a popular

approach for both, the mean-field, the classical heavy-traffic and the many-server heavy-traffic regimes [26, 27, 28, 29, 30, 31, 32]. In [27, 28] the authors establish a three-step framework that is inspired by the classic paper [33] and on the recent work [25]. We present a proof using this approach in chapter 5, where we obtain the rate of convergence in Wasserstein's distance of the total queue length to the many-server heavy-traffic regime.

The drift method was developed in [34] in the context of queueing systems that satisfy the CRP condition, operating in discrete time. The main idea is to set to zero the drift of a carefully chosen test function in steady-state and, using SSC, one obtains bounds on the moments of the queue lengths that are tight in heavy-traffic. In the process of choosing the test function, it is essential to consider the region where SSC occurs. Since the drift method does not involve RBMs, a new notion of SSC is required. The authors in [34] bound the error of approximating the original queue length vector with its projection on the region where SSC occurs, and they show that this error is negligible in the heavy-traffic limit. Therefore, the vector of queue lengths can be well approximated by a vector in the subspace where SSC occurs. In this document we focus on the drift method, and we present a detailed overview of the method in section 2.4.

The drift method was first used in queueing systems that satisfy the CRP condition [34], such as the load balancing system and ad hoc wireless networks (where the CRP condition is imposed). Here the authors compute all the moments of the queue lengths in heavy traffic and, therefore, they are able to draw conclusions about the distribution of the vector of queue lengths. Later, the drift method was generalized to compute the mean of the total queue length in an input-queued switch [14, 15], and the distribution in bandwidth sharing networks [22].

1.4 Notation

In this section we establish the notation that we use throughout this document.

Vectors and numbers

We use $\mathbb{R}$ and $\mathbb{Z}$ to denote the set of real and integer numbers, respectively. We add a subscript $+$ to denote subset of nonnegative numbers and a superscript to denote vector spaces. For example, given a nonnegative integer number n, we use $\mathbb{R}_+^n$ to denote the set of n-dimensional vectors with nonnegative elements.

For any $n \in \mathbb{Z}_+$, we use $[n]$ to denote the set of integers between 1 and n, both included. For example, $[4] = \{1, 2, 3, 4\}$.

We use bold letters to denote vectors, and nonbold letters with an integer subscript to denote their elements. For example, the vector $\boldsymbol{x} \in \mathbb{R}^n$ has elements x_i for $i \in [n]$. We write $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ for convenience, but we treat vectors as column vectors unless otherwise stated.

Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we use $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ to denote the dot product, and for any $p \in \mathbb{Z}_+$, we use $\|\boldsymbol{x}\|_p$ to denote the p-norm, i.e., $\|\boldsymbol{x}\|_p^p = \sum_{i=1}^n |x_i|^p$, where $|x_i|$ denotes the absolute value of x_i. When $p = 2$, we may omit the subscript if it is clear that we mean Euclidean norm from the context.

Given a matrix A, we write A^T to denote its transpose. Given two matrices of the same size, A and B we use $A \circ B$ to denote the Hadamard's product between A and B. For example, if

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix},$$

then

$$A \circ B = \begin{bmatrix} a_{1,1}b_{1,1} & a_{1,2}b_{1,2} & a_{1,3}b_{1,3} \\ a_{2,1}b_{2,1} & a_{2,2}b_{2,2} & a_{2,3}b_{2,3} \\ a_{3,1}b_{3,1} & a_{3,2}b_{3,2} & a_{3,3}b_{3,3} \end{bmatrix}.$$

Let $\mathbb{I}_n$ be the identity matrix of $n \times n$. We use $\boldsymbol{e}^{(i,n)}$ to denote the i^{th} canonical vector in $\mathbb{R}^n$, i.e., a vector with a 1 in the i^{th} element and zeros in all other entries. With this notation, we may write the identity matrix as $\mathbb{I}_n = \left[\boldsymbol{e}^{(1,n)}\ \boldsymbol{e}^{(2,n)}\ \cdots\ \boldsymbol{e}^{(n,n)}\right]$. We may omit the n if the dimension is clear from the context.

Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, the notation $x_{(i)}$ refers to the i^{th} smallest element of $\boldsymbol{x}$. Then, $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the elements of $\boldsymbol{x}$ ordered from smallest to largest.

We use $\binom{n}{k}$ to denote the binomial coefficient, i.e. $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where $n!$ represents the factorial. Given $x \in \mathbb{R}$, the symbol $\lceil x \rceil$ represents the smallest integer which is greater than or equal to x, also known as ceiling function. For two integers k and n, where $n \leq k$, we write $k \bmod n$ to denote the mod function, i.e., $k \bmod n$ is the remainder after dividing n by k. For example, $5 \bmod 3 = 2$.

Probability and Random Processes

Given two random variables X and Y, we denote $\mathbb{E}[X]$ 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 $\mathbb{P}[E]$ the probability of E.

Given a random process $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ (that will be later defined as the queue lengths process), we use $\mathbb{E}_{\boldsymbol{q}}[\cdot] \stackrel{\triangle}{=} \mathbb{E}[\cdot|\boldsymbol{q}(k) = \boldsymbol{q}]$. We use $\Rightarrow$ to denote convergence in distribution.

Functions and asymptotic notation

We use e^x and $\exp(x)$ to denote the exponential function, depending on the size of the argument.

For a function f with domain $Dom(f)$ we denote $\|f\| \stackrel{\triangle}{=} \sup_{x \in Dom(f)} |f(x)|$, and we use f', f'' and f''' for its first, second and third derivative, respectively (provided their existence).

We say $f(n)$ is of order $o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$, and we say that $f(n)$ is of order $O(g(n))$ if there exists a constant C such that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = C$. Similarly, for the limit as

$\epsilon \downarrow 0$, we say $f(\epsilon)$ is of order $o(g(\epsilon))$ if $\lim_{\epsilon \downarrow 0} \frac{f(\epsilon)}{g(\epsilon)} = 0$, and we say that $f(\epsilon)$ is of order $O(g(\epsilon))$ if there exists a constant C such that $\lim_{\epsilon \downarrow 0} \frac{f(\epsilon)}{g(\epsilon)} = C$.

1.5 A general single-hop SPN

In this thesis we mainly work with two SPNs: the load balancing system and the generalized switch. Even though these systems are very different, they share some characteristics and properties that are essential to our analysis. Below we present a general queueing model, that has both of these systems as special cases.

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 each $k \in \mathbb{Z}_+$, let $\boldsymbol{q}(k)$ be the vector of queue lengths at the beginning of time slot k. Throughout this document, we use the term 'queue length' to talk about the number of jobs in the queue, including the job in service (if any). Then, for $i \in [n]$, $q_i(k)$ is the number of jobs (or queue length) in the i^{th} queue at the beginning of time slot k.

Given that the vector of queue lengths in time slot k is $\boldsymbol{q}(k)$, let $a_i\left(\boldsymbol{q}(k)\right)$ be the number of arrivals to the i^{th} queue in time slot k and $s_i\left(\boldsymbol{q}(k)\right)$ be the potential number of jobs that can be served from queue i in time slot k. We say $s_i\left(\boldsymbol{q}(k)\right)$ is potential service because, if there are not enough jobs in line, then less than $s_i\left(\boldsymbol{q}(k)\right)$ jobs are processed. For ease of exposition, and with a slight abuse of notation, from now on we write $\boldsymbol{a}(k)$ and $\boldsymbol{s}(k)$ instead of $\boldsymbol{a}\left(\boldsymbol{q}(k)\right)$ and $\boldsymbol{s}\left(\boldsymbol{q}(k)\right)$, respectively. We assume that $a_i(k)$ and $s_i(k)$ are upper bounded by constants. Specifically, let $A_{\max}$ and $S_{\max}$ be finite constants such that $a_i(k) \leq A_{\max}$ and $s_i(k) \leq S_{\max}$ with probability 1 for all $i \in [n]$ and all $k \in \mathbb{Z}_+$. The difference between potential and actual service is called unused service, which we denote $u_i\left(\boldsymbol{q}(k)\right)$. We also write $\boldsymbol{u}(k)$ instead of $\boldsymbol{u}\left(\boldsymbol{q}(k)\right)$ from now on, for ease of exposition. In some queueing systems, the control problem is to decide the vector $\boldsymbol{a}(k)$ in each time slot (e.g. the load balancing system) and, in others the vector $\boldsymbol{s}(k)$ (e.g. the generalized switch). We give more details about these selection processes in the systems that we describe in section 3.4

and section 6.3, respectively.

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

$$q_i(k+1) = \max\left\{q_i(k) + a_i(k) - s_i(k), 0\right\} \qquad \forall i \in [n],\, \forall k \in \mathbb{Z}_+. \tag{1.1}$$

For each $i \in [n]$ the variables $a_i(k)$ and $s_i(k)$ depend only on $\boldsymbol{q}(k)$, (or they are independent of $\boldsymbol{q}(k)$), then Equation 1.1 implies that the process $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is a DTMC.

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

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k) \qquad \forall i \in [n],\, \forall k \in \mathbb{Z}_+. \tag{1.2}$$

Observe that Equation 1.2 implies

$$q_i(k+1)u_i(k) = 0 \qquad \forall i \in [n],\, \forall k \in \mathbb{Z}_+ \tag{1.3}$$

with probability 1, 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. If $i \neq j$, then we do not necessarily have $q_i(k+1)u_j(k) = 0$ because the fact that queue j is empty at the end of time slot k does not imply that queue i will be empty at the beginning of time slot $k+1$, and vice versa. It turns out that getting a handle on the unused service plays an important role in heavy-traffic analysis and Equation 1.3 will be an essential tool in the analysis. Equation 1.3 can be thought of as a defining property of the queueing process and is analogous to the Skorohod map [35].

In this thesis, we add a line on top of the variables and vectors to denote steady state. Specifically, let $\overline{\boldsymbol{q}}$, $\overline{\boldsymbol{a}} \stackrel{\triangle}{=} \boldsymbol{a}(\overline{\boldsymbol{q}})$, $\overline{\boldsymbol{s}} \stackrel{\triangle}{=} \boldsymbol{s}(\overline{\boldsymbol{q}})$ and $\overline{\boldsymbol{u}} \stackrel{\triangle}{=} \boldsymbol{u}(\overline{\boldsymbol{q}})$ be steady-state vectors that represent the queue lengths at the beginning of a time slot, and arrivals, potential service and unused service in one time slot in steady-state, respectively. Let $\overline{\boldsymbol{q}}^{+} \stackrel{\triangle}{=} \overline{\boldsymbol{q}} + \overline{\boldsymbol{a}} - \overline{\boldsymbol{s}} + \overline{\boldsymbol{u}}$ 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 once the specific characteristics of each system are established.

CHAPTER 2

PRELIMINARIES

As explained in chapter 1, heavy-traffic analysis has been focused on systems that satisfy the CRP condition. Under this condition, one can completely determine the distribution of the queue lengths in heavy traffic. However, if it is not satisfied, one can only partially study the mean queue lengths (as shown in chapter 7). We start with a formal definition of the CRP condition. We use the definition provided [8].

Definition 2.1 (CRP condition)**.** *Consider a set of queueing systems as described in section 1.5, where the capacity region is $\mathcal{C}$. Suppose that in heavy traffic, the vector of arrival rates approaches a point $\boldsymbol{\nu}$ in the boundary of $\mathcal{C}$. We say that the queueing system satisfies the Complete Resource Pooling (CRP) condition if the outer normal vector to $\mathcal{C}$ at $\boldsymbol{\nu}$ is unique up to a scalar coefficient.*

In other words, the CRP condition implies that the system can be operated such that all the servers pool together in the heavy-traffic limit [2, 3, 4]. Hence, under this condition, the queueing system intuitively behaves as a one-dimensional queueing system (i.e., as a single server queue) if it is operated under a 'good' control algorithm. Therefore, the MGF method is essentially similar to the proof of Theorem 2.11 after one establishes SSC on a one-dimensional subspace of the state space.

In this thesis, many of the results are proved using the drift of a function of the vector of queue lengths. In this section, we provide essential definitions and results for our analysis. Most of the work in this thesis is based on discrete-time models of SPNs. However, in Chapter 5 we additionally study a continuous-time model for the load balancing system (that we describe in detail therein). We present the definitions and preliminary results in both cases.

2.1 Definition of drift

Intuitively, the drift of a function Z at a state x is a random variable that measures the amount of change in the value of $Z(x)$. In the case of discrete-time processes, it measures the change in one time slot, and in the case of continuous-time processes, it measures the change in the next transition. We present both definitions below. We start with the discrete-time version.

Definition 2.2. *For an irreducible and aperiodic DTMC $\{X(k) : k \in \mathbb{Z}_+\}$ over a countable state space $\mathcal{X}$, suppose $Z : \mathcal{X} \to \mathbb{R}_+$ is a Lyapunov function. Define the drift of Z at x as*

$$\Delta Z(x) \stackrel{\triangle}{=} \left[Z\left(X(k+1)\right) - Z\left(X(k)\right) \right] \mathbb{1}_{\{X(k)=x\}}. \tag{2.1}$$

The corresponding definition for continuous-time processes is presented below.

Definition 2.3. *For a Continuous-Time Markov Chain (CTMC) $\{X(t) : t \in \mathbb{R}_+\}$ with countable state space $\mathcal{X}$ and transition-rate matrix G^X, suppose $Z : \mathcal{X} \to \mathbb{R}_+$ is a Lyapunov function. Define the drift of Z at x as*

$$\Delta Z(x) \stackrel{\triangle}{=} \sum_{x' \in \mathcal{X} \setminus \{x\}} G^X_{x,x'} \left(Z(x') - Z(x) \right). \tag{2.2}$$

In both cases (discrete and continuous time), we say that we set to zero the drift of Z in steady-state when we use the property $\mathbb{E}\left[\Delta Z(\overline{X})\right] = 0$ under stationary distribution, provided that $\mathbb{E}\left[Z(\overline{X})\right] < \infty$. Here, $\overline{X}$ is a random variable that is limit in distribution of the Markov chain $\{X(k) : k \in \mathbb{Z}_+\}$ in the discrete-time case, or the Markov chain $\{X(t) : t \in \mathbb{R}_+\}$ in continuous time. Hence, in order to set to zero the drift of a function, the first step is to show positive recurrence of the corresponding Markov chain. We provide a certificate of positive recurrence for discrete and continuous-time processes in section 2.2.

2.2 Stability criteria

We work with queueing systems in steady state and, hence, it is essential to show that the embedded Markov chains are positive recurrent. We present a certificate of positive recurrence for discrete-time and continuous-time processes.

In the following result, we present a certificate that a DTMC is positive recurrent. We state the result as presented in [36, Theorem 3.3.7].

Theorem 2.4 (Foster-Lyapunov Theorem for discrete-time processes)**.** *Let* $\{X(k) : k \in \mathbb{Z}_+\}$ *be an irreducible DTMC with countable state space* $\mathcal{X}$*. Suppose that there exists a function* $Z : \mathcal{X} \to \mathbb{R}_+$ *and a finite set* $\mathcal{B} \subseteq \mathcal{X}$ *satisfying the following conditions:*

(i) $\mathbb{E}\left[\Delta Z(x) | X(k) = x\right] \leq -\eta$ *if* $x \in \mathcal{X} \setminus \mathcal{B}$ *for some* $\eta > 0$*,*

(ii) $\mathbb{E}\left[\Delta Z(x) | X(k) = x\right] \leq D$ *if* $x \in \mathcal{B}$ *for some finite constant* D*.*

Then, the DTMC $\{X(k) : k \in \mathbb{Z}_+\}$ *is positive recurrent.*

In the following result, we present a certificate that a DTMC is not positive recurrent, i.e., that is either null recurrent or transient. We state the result as presented in [36, Theorem 3.3.10].

Theorem 2.5. *An irreducible DTMC* $\{X(k) : k \in \mathbb{Z}_+\}$ *with countable state space* $\mathcal{X}$ *is either transient or null recurrent if there exists a function* $Z : \mathcal{X} \to \mathbb{R}_+$ *and a finite set* $\mathcal{B} \subseteq \mathcal{X}$ *satisfying the following conditions:*

(i) $\mathbb{E}\left[\Delta Z(x) | X(k) = x\right] \geq 0$ *for all* $x \in \mathcal{X} \setminus \mathcal{B}$*,*

(ii) There exists some $x \in \mathcal{X} \setminus \mathcal{B}$ *such that* $Z(x) > Z(y)$ *for any* $y \in \mathcal{B}$*,*

(iii) $\mathbb{E}\left[|\Delta Z(x)| \, | X(k) = x\right] \leq D$ *for all* $x \in \mathcal{X}$ *for some* $D < \infty$*.*

To end this section, we present a certificate that a continuous-time process is positive recurrent. We present the result as stated in [37, Corollary 6.18].

Theorem 2.6. *Let $\{X(t) : t \in \mathbb{R}_+\}$ be a CTMC with state space $\mathcal{X}$. Suppose the functions $Z : \mathcal{X} \to \mathbb{R}_+$, $f : \mathcal{X} \to \mathbb{R}_+$ and $g : \mathcal{X} \to \mathbb{R}_+$ satisfy:*

(i) $\Delta Z(x) \leq -f(x) + g(x)$ *for all* $x \in \mathcal{X}$

(ii) There exists $\epsilon > 0$ *such that the set* $\{x \in \mathcal{X} : f(x) < g(x) + \epsilon\}$ *is finite*

(iii) The set $\{x \in \mathcal{X} : Z(x) \leq K\}$ *is finite for any constant* K.

Then, $\{X(t) : t \in \mathbb{R}_+\}$ *is positive recurrent, so it converges in distribution to a steady-state random variable* $\overline{X}$. *Further,* $\mathbb{E}\left[f(\overline{X})\right] \leq \mathbb{E}\left[g(\overline{X})\right]$.

2.3 Moment bounds based on drift arguments

An essential step in heavy-traffic analysis is State Space Collapse (SSC). In this work, we show SSC by bounding the error between the actual vector of queue lengths and its projection on the subspace where SSC occurs. To show such a result, we use moment bounds based on drift arguments.

The following result was first established in [14, Lemma 3] for discrete-time processes, and combines the more general results proved in [38] and [39, Theorem 1].

Lemma 2.7. *For an irreducible and aperiodic DTMC $\{X(k) : k \in \mathbb{Z}_+\}$ over a countable state space $\mathcal{X}$, suppose $Z : \mathcal{X} \to \mathbb{R}_+$ is a Lyapunov function and consider its drift at x, $\Delta Z(x)$. Suppose the following conditions are satisfied.*

(C1) There exists $\eta > 0$ and $\kappa < \infty$ such that $\mathbb{E}\left[\Delta Z(x) \,|\, X(k) = x\right] \leq -\eta$ for all $x \in \mathcal{X}$ with $Z(x) \geq \kappa$.

(C2) There exists $D < \infty$ such that $\mathbb{P}\left[|\Delta Z(x)| \leq D\right] = 1$ for all $x \in \mathcal{X}$.

Further, assume that the Markov chain $\{X(k) : k \in \mathbb{Z}_+\}$ converges in distribution to a random variable $\overline{X}$ as $k \uparrow \infty$. Then, for any $j \in \mathbb{Z}_+$ with $j \geq 1$,

$$\mathbb{E}\left[Z(\overline{X})^j\right] \leq (2\kappa)^j + (4D)^j \left(\frac{D+\eta}{\eta}\right)^j j!$$

Lemma 2.7 gives conditions on the drift of a function Z, under which the moments of such function in steady state can be explicitly bounded. In the following result, we additionally present a bound on the MGF of $Z(\overline{X})$.

Lemma 2.8. *Let $\{X(k) : k \in \mathbb{Z}_+\}$ be a DTMC as described in Lemma 2.7, and suppose it satisfies the two conditions therein. Let $\theta \in \mathbb{R}$ be such that $|\theta| \leq \frac{1}{2D} \log\left(1 + \frac{\eta}{D}\right)$.Then,*

$$\mathbb{E}\left[e^{\theta Z(\overline{X})}\right] \leq e^{\theta\kappa}\left(\frac{\eta}{\eta + D(1 - e^{2\theta D})}\right).$$

The proof of Lemma 2.8 is similar to the proof of Lemma 2.10, so we omit it for brevity.

A continuous-time version of Lemma 2.7 is proved in [22, Lemma 4.1]. We state it below.

Lemma 2.9. *Let $\{X(t) : t \in \mathbb{R}_+\}$ be a CTMC over a countable state space $\mathcal{X}$, with transition rate matrix G^X. Suppose that it is irreducible, nonexplosive and positive recurrent, and it converges in distribution to a random variable $\overline{X}$. Consider a Lyapunov function $Z : \mathcal{X} \to \mathbb{R}_+$ and suppose its drift satisfies the following conditions:*

(C1) There exist constants $\eta > 0$ and $\kappa > 0$ such that $\Delta Z(x) \leq -\eta$ for any $x \in \mathcal{X}$ with $Z(x) > \kappa$,

(C2) $\nu_{\max} \stackrel{\triangle}{=} \sup\left\{|Z(x') - Z(x)| : x, x' \in \mathcal{X} \text{ and } G^X_{x,x'} > 0\right\}$ is finite,

(C3) $\overline{G} \stackrel{\triangle}{=} \sup\left\{-G^X_{x,x} : x \in \mathcal{X}\right\}$ is finite.

Then, for any $j \in \mathbb{Z}_+$ with $j \geq 1$, we have

$$\mathbb{P}\left[Z(\overline{X}) > \kappa + 2\nu_{\max} j\right] \leq \left(\frac{G_{\max}\nu_{\max}}{G_{\max}\nu_{\max} + \eta}\right)^{j+1}, \tag{2.3}$$

where

$$G_{\max} \stackrel{\triangle}{=} \sup\left\{\sum_{x' \in \mathcal{X} : Z(x) < Z(x')} G^X_{x,x'} : x \in \mathcal{X}\right\}.$$

As a result, for any positive integer j, the j^{th} moment of $Z(\overline{X})$ can be bounded as follows:

$$\mathbb{E}\left[Z(\overline{X})^j\right] \leq (2\kappa)^j + (4\nu_{\max})^j \left(\frac{G_{\max}\nu_{\max}+\eta}{\eta}\right)^j j! \tag{2.4}$$

Lemma 2.10. *Let $\{X(t) : t \in \mathbb{R}_+\}$ be a CTMC as described in Lemma 2.9, and suppose it satisfies the three conditions therein. Let $\theta \in \mathbb{R}$ be such that $|\theta| < \frac{1}{2\nu_{\max}} \log\left(1 + \frac{\eta}{G_{\max}\nu_{\max}}\right)$. Then,*

$$\mathbb{E}\left[e^{\theta Z(\overline{X})}\right] \leq \frac{e^{\theta\kappa}\eta}{\eta + G_{\max}\nu_{\max}(1 - e^{2\nu_{\max}\theta})}.$$

Before ending this section, we prove Lemma 2.10.

Proof of Lemma 2.10. First observe that $Z(\overline{X}) \geq 0$ by assumption. Then,

$$e^{\theta Z(\overline{X})} \leq e^{|\theta| Z(\overline{X})}.$$

We compute an upper bound for $\mathbb{E}\left[e^{|\theta|Z(\overline{X})}\right]$. Let $F_Z(x)$ be the cumulative distribution function of $Z(\overline{X})$. Then,

$$\begin{aligned}
&\mathbb{E}\left[e^{|\theta|Z(\overline{X})}\right] \\
&= \int_0^\infty e^{|\theta|x}\, dF_Z(x) \\
&\overset{(a)}{=} \left[-e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > x\right]\right]_0^\infty + |\theta| \int_0^\infty e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > x\right]\, dx \\
&= \mathbb{P}\left[Z(\overline{X}) > 0\right] + |\theta| \int_0^\kappa e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > x\right]\, dx + |\theta| \int_\kappa^\infty e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > x\right]\, dx \\
&\overset{(b)}{\leq} e^{|\theta|\kappa} + \sum_{i=0}^\infty \int_{\kappa+2\nu_{\max}i}^{\kappa+2\nu_{\max}(i+1)} |\theta| e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > x\right]\, dx \\
&\overset{(c)}{\leq} e^{|\theta|\kappa} + \sum_{i=0}^\infty \int_{\kappa+2\nu_{\max}i}^{\kappa+2\nu_{\max}(i+1)} |\theta| e^{|\theta|x}\mathbb{P}\left[Z(\overline{X}) > \kappa + 2\nu_{\max}i\right]\, dx \\
&\overset{(d)}{\leq} e^{|\theta|\kappa} + e^{|\theta|\kappa}\left(e^{2|\theta|\nu_{\max}} - 1\right)\left(\frac{G_{\max}\nu_{\max}}{G_{\max}\nu_{\max}+\eta}\right) \sum_{i=0}^\infty \left(\frac{G_{\max}\nu_{\max}e^{2|\theta|\nu_{\max}}}{G_{\max}\nu_{\max}+\eta}\right)^i
\end{aligned}$$

$$\stackrel{(e)}{=} \frac{e^{|\theta|\kappa}\eta}{\eta + G_{\max}\nu_{\max}(1 - e^{2\nu_{\max}|\theta|})},$$

where (a) holds integrating by parts; (b) holds because probabilities are upper bounded by 1, solving $\int_0^\kappa e^{|\theta|x}\, dx$, and breaking the last integral into intervals; (c) holds because $f(x) = 1 - F_Z(x) = \mathbb{P}\left[Z(\overline{X}) > x\right]$ is a nonincreasing function; (d) holds by Equation 2.3 and solving the integral; and (e) holds after solving the geometric summation and reorganizing terms, because $|\theta| < \frac{1}{2\nu_{\max}} \log\left(1 + \frac{\eta}{G_{\max}\nu_{\max}}\right)$ by assumption and, hence, the geometric sum converges. □

2.4 Overview of the drift method

The drift method is a direct approach to study the heavy-traffic behavior of the vector of queue lengths in heavy traffic. We call the method 'direct' because one studies the heavy-traffic behavior of the steady-state queue lengths *directly* and, hence, the interchange of limits issue does not arise.

The drift method is based on the analysis developed in [23] for a single-server queue with general inter-arrival and service time distributions. It has been developed in a series of articles [34, 14, 15, 22], and can be summarized in two steps: (i) State Space Collapse (SSC), and (ii) Asymptotically tight bounds. We describe each of these steps in subsection 2.4.1 and subsection 2.4.2.

2.4.1 State space collapse

In this step, the goal is to show that the vector of queue lengths can be a approximated by a vector that lies in a subspace of the state space that frequently has a lower dimension. Denote $\mathcal{K}$ the subspace where SSC occurs, and let $\boldsymbol{q}_{\|}$ be the projection of the vector of queue lengths $\boldsymbol{q}$ on $\mathcal{K}$. Additionally, define $\boldsymbol{q}_{\perp} \stackrel{\triangle}{=} \boldsymbol{q} - \boldsymbol{q}_{\|}$ and observe that $\boldsymbol{q}_{\perp}$ represents the error of approximating $\boldsymbol{q}$ by $\boldsymbol{q}_{\|}$.

The goal is to show that the moments of $\|\boldsymbol{q}_{\perp}\|$ are negligible in heavy traffic. To prove

such a result, we use the lemmas stated in section 2.3. In these lemmas, one uses Lyapunov-drift arguments to show moment bounds. Since the goal is to bound the moments of $\|\boldsymbol{q}_\perp\|$, we use $Z(\boldsymbol{q}) = \|\boldsymbol{q}_\perp\|$ and show that the conditions are satisfied. The most challenging condition to show is the first condition of Lemma 2.7 and Lemma 2.9, where one needs to show negative drift outside a bounded set. We sketch the steps that are commonly followed to show this condition.

Define

$$W_\perp(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_\perp\|,\ V(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}\|^2,\ \text{and}\ V_\parallel(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_\parallel\|^2.$$

Then, since $W_\perp(\boldsymbol{q}) = \sqrt{\|\boldsymbol{q}_\perp\|^2}$ and $f(x) = \sqrt{x}$ is a concave function, we have the following inequality:

$$\Delta W_\perp(\boldsymbol{q}) \leq \frac{1}{2W_\perp(\boldsymbol{q})} \left(\Delta V(\boldsymbol{q}) - \Delta V_\parallel(\boldsymbol{q})\right). \tag{2.5}$$

Equation 2.5 was proved for discrete-time processes in [34], and we prove it for the continuous-time process that we study in chapter 5 in subsection 5.8.3.

Then, it suffices to find an upper bound for $V(\boldsymbol{q})$ and a lower bound for $V_\parallel(\boldsymbol{q})$.

2.4.2 Asymptotically tight bounds

In this step we set to zero the drift of a test function, and use SSC to compute bounds that are tight in heavy traffic. The choice of the function plays a key role, and it should be related to the region where SSC occurs in order to obtain meaningful bounds. A popular function is $V_\parallel(\boldsymbol{q}) = \|\boldsymbol{q}_\parallel\|^2$. In [34, 14, 15], the authors show that if we set to zero the drift of this test function, we obtain bounds on the total queue length. In general, if we use polynomial test functions of degree $m+1$, we can obtain bounds on the m^{th} moment of the queue lengths.

In the case of systems that satisfy the CRP condition, one can explicitly obtain all the

moments of the queue lengths in heavy traffic and, hence, prove convergence in distribution additional mild conditions (see [40, Section 4.10] for a discussion of these conditions). For example, in the case of a single server queue, one can inductively use $q^2, q^3, q^4, \ldots$ as test function, to obtain the expected value of $q, q^2, q^3, \ldots$ (where q denotes the queue length).

When the CRP condition is not satisfied, one cannot obtain the moments of the queue lengths in general. We show this result formally in section 7.8.

2.5 Transform method based on the drift method

In this section we introduce a variant of the drift method, which uses a test function motivated by the moment generating function (MGF), and is a contribution of this thesis. The key insight is that, instead of using the polynomial test functions of increasing degrees inductively as in the drift method, all the polynomials can be combined in Taylor series to obtain an exponential test function. For example, in the case of a single server queue, combining $q, q^2, q^3, \ldots$ in Taylor series (with appropriate coefficients), we obtain $e^{\theta q}$ for some constant θ, and $\mathbb{E}\left[e^{\theta q}\right]$ is the MGF of q. We exemplify the method in the context of a single server queue below, and we highlight the main steps. We present the details of the method and an illustration of its use in the context of a load balancing system in chapter 3. We additionally apply the method to a generalized switch in chapter 6.

We provide a proof of the well-known result that the scaled queue length of a single server queue has an exponential distribution in heavy-traffic to illustrate the method and to show its simplicity. We do not provide all the details of our proofs, since the single server queue is a special case of the load balancing system ($n = 1$) and this system is studied in detail in chapter 3.

Consider a single server queue operating in discrete time, i.e., a queueing system as described in section 1.5 with $n = 1$. Arrivals and potential service in each time slot are assumed to be independent sequences of i.i.d. random variables. Since they are also assumed to be finite with probability 1 (as specified in section 1.5), their MGFs $\mathbb{E}\left[e^{\theta a(1)}\right]$ and

$\mathbb{E}\left[e^{\theta s(1)}\right]$ exist for all $\theta \in \mathbb{R}$.

Let $\lambda \stackrel{\triangle}{=} \mathbb{E}\left[a(1)\right]$ and $\mu \stackrel{\triangle}{=} \mathbb{E}\left[s(1)\right]$. Observe that λ and μ are the rates of arrival and service, respectively, since they are the expected number of arrival/services in one time slot. Then, the capacity region of the single server queue is $\mathcal{C} = \{\lambda \in \mathbb{R}_+ : \ \lambda \leq \mu\}$. To study de heavy-traffic asymptotics, we parametrize the arrival process by $\epsilon \in (0, \mu)$ as follows. We consider a set of single server queues with a fixed service process of rate μ and arrival rate $\lambda^{(\epsilon)} \stackrel{\triangle}{=} \mu - \epsilon$. For each $k \in \mathbb{Z}_+$, let $q^{(\epsilon)}(k)$, $a^{(\epsilon)}(k)$ and $u^{(\epsilon)}(k)$ be the queue length, number of arrivals and unused service in time slot k in the system parametrized by ϵ.

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

In the rest of this section we prove Theorem 2.11. This is a well-known result and there are proofs using diffusion limits [23] and the drift method [34] 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 2.11. *Let $\epsilon \in (0, \mu)$ and consider a set of single server queues parametrized by ϵ as described above. Let $\overline{q}^{(\epsilon)}$ be a steady-state random variable such that $\{q^{(\epsilon)}(k) : \ k \geq 1\}$ converges in distribution to $\overline{q}^{(\epsilon)}$ as $k \uparrow \infty$. Further, assume $\lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a$. Then, $\epsilon \overline{q}^{(\epsilon)} \Rightarrow \Upsilon$ as $\epsilon \downarrow 0$, where Υ is an exponential random variable with mean $\frac{\sigma_a^2 + \sigma_s^2}{2}$.*

It is well-known that for all $\epsilon \in (0, \mu)$, the Markov chain $\{q^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent. A proof using Foster-Lyapunov theorem (Theorem 2.4) can be found in [36, Theorem 3.4.2]. Then, $\overline{q}^{(\epsilon)}$ is well defined.

Before presenting the proof, we prove two lemmas. The first lemma is a different version of Equation 1.3 and is key in the MGF method. For other queueing systems we use a weaker version of this lemma, that is sufficient for the MGF method (see Step 1 in section 3.3 for more details).

Lemma 2.12. *Consider a single server queue parametrized by ϵ as described above. Then, for all $\alpha, \beta \in \mathbb{R}$ and each $k \in \mathbb{Z}_+$ we have*

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

Proof of Lemma 2.12. It follows from Equation 1.3 and because $e^x - 1 = 0$ if and only if $x = 0$. □

The next Lemma gives the first moment of the unused service in steady state, and it will be used at the end of the proof of Theorem 2.11.

Lemma 2.13. *Consider a single server queue parametrized by $\epsilon \in (0, \mu)$ as described above. Then,*

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

Proof of Lemma 2.13. We set to zero the drift of the linear test function $V_1(q) = q$, and we obtain

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

where the last equality holds because $\left(\overline{q}^{(\epsilon)}\right)^+ \stackrel{\triangle}{=} \overline{q}^{(\epsilon)} + \overline{a}^{(\epsilon)} - \overline{s} + \overline{u}^{(\epsilon)}$ by definition. Rearranging terms we obtain

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

□

Now we prove Theorem 2.11.

Proof of Theorem 2.11. If we expand the product in Lemma 2.12 and rearrange terms we obtain

$$e^{\theta\epsilon q^{(\epsilon)}(k+1)} - e^{\theta\epsilon\left(q^{(\epsilon)}(k)+a^{(\epsilon)}(k)-s(k)\right)} = 1 - e^{-\theta\epsilon u^{(\epsilon)}(k)}. \tag{2.6}$$

Observe that Equation 2.6 holds for all $k \in \mathbb{Z}_+$. In particular, it holds in steady-state. Also, it can be shown that $\mathbb{E}\left[e^{\theta\epsilon\overline{q}^{(\epsilon)}}\right] < \infty$ in an interval around 0. We omit the proof because we provide a proof for the load balancing system in Lemma 3.13, and the single server queue is a particular case of the load balancing system ($n = 1$). Therefore, $\mathbb{E}\left[e^{\theta\epsilon\left(\overline{q}^{(\epsilon)}\right)^+}\right] = \mathbb{E}\left[e^{\theta\epsilon\overline{q}^{(\epsilon)}}\right]$. Taking expected value with respect to the stationary distribution in Equation 2.6 we obtain

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

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

$$\mathbb{E}\left[e^{\theta\epsilon\overline{q}^{(\epsilon)}}\right] = \frac{1 - \mathbb{E}\left[e^{-\theta\epsilon\overline{u}^{(\epsilon)}}\right]}{1 - \mathbb{E}\left[e^{\theta\epsilon\left(\overline{a}^{(\epsilon)}-\overline{s}\right)}\right]} \tag{2.7}$$

Observe that Equation 2.7 gives an expression for the MGF of $\epsilon\overline{q}^{(\epsilon)}$ that is valid for all traffic intensities. However, it does not give an explicit expression because the right-hand side depends on the unused service, which is a function of $\overline{q}^{(\epsilon)}$. In the rest of this section, we take the heavy-traffic limit of Equation 2.7 and we obtain an explicit expression.

Observe that the right-hand side of Equation 2.7 yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$. Then, we take Taylor series of each term with respect to θ, around $\theta = 0$. The technical details of why this expansion can be done are established in Lemma 3.1, which is presented in section 3.3 For the numerator we obtain

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

$$=\theta\epsilon^2 + O(\epsilon^3), \tag{2.8}$$

where the last equality holds by Lemma 2.13 and because $\mathbb{E}\left[\left(\overline{u}^{(\epsilon)}\right)^2\right]$ is $O(\epsilon)$. Details of this argument will be provided in chapter 3 for the load balancing system (see Claim 3.11). The main idea of the proof is that $0 \leq \overline{u}^{(\epsilon)} \leq \overline{s}$ by definition and, hence, $\overline{u}^{(\epsilon)}$ is a bounded random variable. For the denominator we obtain

$$\begin{aligned} 1 - \mathbb{E}\left[e^{\theta\epsilon(\overline{a}^{(\epsilon)} - \overline{s})}\right] =& -\theta\epsilon\mathbb{E}\left[\overline{a}^{(\epsilon)} - \overline{s}\right] - \frac{(\theta\epsilon)^2}{2}\mathbb{E}\left[(\overline{a}^{(\epsilon)} - \overline{s})^2\right] + O(\epsilon^3) \\ =& \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sigma_s^2 + \epsilon^2\right) + O(\epsilon^3), \end{aligned} \tag{2.9}$$

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

If we replace Equation 2.8 and Equation 2.9 in Equation 2.7, and cancel out $\theta\epsilon^2$ from numerator and denominator we obtain

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

Therefore, taking the heavy-traffic limit we obtain

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\theta\epsilon\overline{q}^{(\epsilon)}}\right] = \frac{1}{1 - \theta\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)}. \tag{2.10}$$

Since the right-hand side is the MGF of an exponential random variable with mean $\frac{\sigma_a^2 + \sigma_s^2}{2}$, Equation 2.10 implies that $\epsilon\overline{q}^{(\epsilon)}$ converges in distribution to such an exponential random variable [40, Theorem 9.5 in Section 5]. □

Remark 2.14. *In this thesis we introduce the transform method sketched above, and we showcase its simplicity and flexibility in the context of a load balancing system and a generalized switch. This method is the start of a line of work that is still ongoing. In particular, [41] uses this method to obtain the distribution of an input-queued switch where the arrival*

process is governed by a Markov chain, as opposed to the i.i.d. assumption here.

Remark 2.15. *One of the reasons why the proofs using the transform method are simple, is the fraction obtained in Equation 2.7, which corresponds to Equation 3.7 in the load balancing system and Equation 6.5 in the generalized switch. These equations give explicit expressions for the MGF of the scaled queue lengths and, hence, all we need to do to take the heavy-traffic limit is bounding the terms related to the unused service.*

However, this situation becomes more complex in other settings. In an input-queued switch that does not satisfy CRP [42] and two-sided queues [43], one obtains an implicit equation for the MGF of the queue lengths after step 1. Therefore, one needs to show uniqueness of the solution. This last step is usually challenging.

CHAPTER 3

HEAVY-TRAFFIC ANALYSIS OF LOAD-BALANCING SYSTEMS

Based on:

D. Hurtado-Lange and S. T. Maguluri, "Transform methods for heavy-traffic analysis," *Stochastic Systems*, vol. 10, no. 4, pp. 275–309, 2020

3.1 Introduction

The primary contribution of this chapter 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 a load balancing system. We also show 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 chapter 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. Under the CRP condition and routing algorithms that ensure SSC to a one-dimensional subspace, we show that even under correlated services, the heavy-traffic scaled stationary distribution continues to be exponential (see Theorem 3.5). It is possible to allow for this generalization using other methods, but we illustrate the simplicity of such generalizations using the MGF method.

3.2 Related work

Moment generating functions have been used in the literature to study queueing systems such as the classical analysis of $M/G/1$ queue [45]. 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 in [46, 47] to study heavy-traffic queueing systems, and moment generating functions were used in [48, 49]. In contrast, the primary focus of this work is to develop transform methods for heavy-traffic analysis that incorporates SSC.

The single server queue was first studied in heavy-traffic in [47] using characteristic functions and tools from complex analysis. Characteristic functions are also used in [46] to study the single server queue. The diffusion limit method to study queueing systems was developed by studying the single server queue [23]. The well known Kingman bound for the expected waiting time in a single server queue was developed in the 60's [50], and later the second moment was computed using similar arguments [51]. These formed the basis for the drift method, that was developed in [34]. The single server queue was also presented as an illustrative example of the BAR method [24]. Most of these papers study the delay in $G/G/1$ queue in continuous time, which evolves according to Lindley's equation [52]. Similar to [34], in this chapter we study the queue length in discrete time. The queue lengths process evolves from one time slot to the next according to Equation 1.2, which is equivalent to Lindley's equation for the waiting time of $(k+1)^{\text{th}}$ 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 [17, 53, 54]. Heavy-traffic

optimality of the JSQ policy in a system with two servers was established in [16] 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 [55, 56, 26, 57, 31, 32, 58, 59, 60, 61]. One lower complexity algorithm that has received attention is the power-of-two choices algorithm [18, 19, 20, 55]. An exhaustive survey of literature on load balancing is presented in [62]. The most relevant work for our purposes is the study of the JSQ policy under the drift method [34] and that of the power-of-two choices algorithm [63].

3.3 General MGF framework

In section 2.5 we proved a well-known result using the MGF method in the case of the simplest queueing system, i.e., the single server queue. In this section we describe the method in detail, generalizing the key steps from section 2.5 to more general queueing systems that satisfy the CRP condition. Then, we apply it to a load balancing system in section 3.5. Later, in section 6.3, we apply it to a generalized switch that satisfies CRP.

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

Prerequisite 1. Positive recurrence.

Before using the MGF method, one needs to prove that the Markov chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent for $\epsilon > 0$. Positive recurrence is a requirement to ensure the queue length process $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ converges in distribution to a steady-state random vector, that we denote $\overline{\boldsymbol{q}}^{(\epsilon)}$, as $k \uparrow \infty$.

Prerequisite 2. State Space Collapse.

To use the MGF method, one also needs to prove SSC into a one-dimensional subspace. Let $\boldsymbol{c} \geq \mathbf{0}$ be the direction into which SSC occurs. For simplicity, we assume $\|\boldsymbol{c}\| = 1$. Then

$\mathcal{K} = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \xi \boldsymbol{c}\,,\, \xi \in \mathbb{R}_+\}$ is the cone where the state space collapses in heavy traffic. For any n-dimensional vector $\boldsymbol{x}$, let $\boldsymbol{x}_\parallel \stackrel{\triangle}{=} \langle \boldsymbol{x}, \boldsymbol{c}\rangle \boldsymbol{c}$ be the projection of $\boldsymbol{x}$ on $\mathcal{K}$ and let $\boldsymbol{x}_\perp \stackrel{\triangle}{=} \boldsymbol{x} - \boldsymbol{x}_\parallel$. In this step it should be proved that $\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^2\right]$ is $o\left(\frac{1}{\epsilon^2}\right)$, which is equivalent to proving that $\epsilon^2 \mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^2\right]$ is $o(1)$.

The queueing systems that we study in this dissertation actually exhibit a stronger form of SSC, where $\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^j\right]$ is $O(1)$ for all $j \in \mathbb{Z}_+$ with $j \geq 1$. However, a weaker form of SSC is studied in [22, 64].

From this notion of SSC, we conclude that

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

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

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

Step 1. Prove an equation of the form of

$$\mathbb{E}\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}, \left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^+\rangle} - 1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{u}}^{(\epsilon)}\rangle} - 1\right)\right] \quad \textit{is } o(\epsilon^2) \tag{3.1}$$

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

The key in the MGF method is to handle unused service and its interaction with the queue lengths, arrivals and potential service. In the drift method, the unused service is handled with Equation 1.3. However, in this case we want to work with an exponential transform of the queue lengths, so we need to write Equation 1.3 in a different way. In the case of the single server queue, we used Lemma 2.12 which, in fact, it is much stronger than what we actually use in the MGF method. For more general queueing systems we use Equation 3.1.

To prove an equation of the form of Equation 3.1 it is essential to use SSC. After proving Equation 3.1, we need to obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$ that is valid for

all traffic. Below we sketch some algebraic steps that are useful to do it. Expanding the product on the left-hand side of Equation 3.1 we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(e^{\theta\epsilon\langle \boldsymbol{c},\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}\rangle}-1\right)\left(e^{-\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}-1\right)\right] \\
=&\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}-\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]-\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}\rangle}\right]+1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right] \qquad (3.2)\\
\overset{(a)}{=}&\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{q}}^{(\epsilon)}+\overline{\boldsymbol{a}}^{(\epsilon)}-\overline{\boldsymbol{s}}^{(\epsilon)}\rangle}\right]-\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}\rangle}\right]+1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right] \\
\overset{(b)}{=}&\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{q}}^{(\epsilon)}+\overline{\boldsymbol{a}}^{(\epsilon)}-\overline{\boldsymbol{s}}^{(\epsilon)}\rangle}\right]-\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{q}}^{(\epsilon)}\rangle}\right]+1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right], \qquad (3.3)
\end{aligned}$$

where (a) holds by the dynamics of the queues described in Equation 1.2 and by definition of $\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}$; and (b) holds if the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists in an interval around 0 (this must be proved). In such case, by definition of steady state we have $\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+}\rangle}\right] = \mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{q}}^{(\epsilon)}\rangle}\right]$, which is equivalent to setting the drift of the test function $V(\boldsymbol{q}) = e^{\theta\epsilon\langle \boldsymbol{c},\boldsymbol{q}\rangle}$ to zero.

Observe that when we first expand the product in Equation 3.2, we obtain two terms that are related to the unused service (the first and the last term). We use the dynamics of the queues, as described by Equation 1.2, to deal with the first one, and we write $\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^{+} - \overline{\boldsymbol{u}}^{(\epsilon)}$ in terms of $\overline{\boldsymbol{q}}^{(\epsilon)}$, $\overline{\boldsymbol{a}}^{(\epsilon)}$ and $\overline{\boldsymbol{s}}^{(\epsilon)}$. The last term is the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{u}}^{(\epsilon)}\rangle$, and we deal with it in the second step of the MGF method.

Using Equation 3.3 in Equation 3.1 and reorganizing terms we obtain

$$\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{q}}^{(\epsilon)}\rangle}\left(1-e^{\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{a}}^{(\epsilon)}-\overline{\boldsymbol{s}}^{(\epsilon)}\rangle}\right)\right] = 1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]+o(\epsilon^2) \qquad (3.4)$$

From Equation 3.4 we can obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$ which is valid for all traffic. However, the steps to obtain it depend on the properties of each queueing system. For example, in the case of the single server queue we know that the arrival and potential service processes are independent of the queue lengths. Then, we can separate the product on the left-hand side and we obtain Equation 2.7.

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

Observe that the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{a}}^{(\epsilon)} \rangle$ and $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{s}}^{(\epsilon)} \rangle$ exist for all $\theta \in \mathbb{R}$, because the random variables are bounded by assumption. Further, by definition of unused service, we have $\mathbf{0} \leq \overline{\boldsymbol{u}}^{(\epsilon)} \leq \overline{\boldsymbol{s}}^{(\epsilon)}$ component-wise. Then, the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{u}}^{(\epsilon)} \rangle$ exists for all $\theta \in \mathbb{R}$. Also, in Step 1 (before obtaining Equation 3.3) it was proved that the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ exists in an interval around zero. Therefore, as $\epsilon \downarrow 0$, Equation 3.4 yields $0 = 0$. As mentioned above, depending on the queueing system we will use different approaches to obtain an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ that is valid for all traffic from Equation 3.4. For example, in the case of the single server queue we obtained Equation 2.7, which yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$. Therefore, to compute the heavy-traffic limit we take Taylor series of each term around $\theta = 0$, except for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$. To do that, we use the following lemma.

Lemma 3.1. *Let $X^{(\epsilon)}$ be a set of random variables indexed by $\epsilon > 0$. Assume $X^{(\epsilon)}$ is bounded for all ϵ, i.e., there exists a constant $K_{\max}$ (that does not depend on ϵ) such that $X^{(\epsilon)} \leq K_{\max}$ with probability 1. Define $f_{\epsilon,X}(\theta) \stackrel{\triangle}{=} e^{\theta\epsilon X^{(\epsilon)}}$. Then,*

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

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

$$\mathbb{E}\left[f_{\epsilon,X}(\theta)\right] = 1 + \theta\epsilon\mathbb{E}\left[X^{(\epsilon)}\right] + \frac{(\theta\epsilon)^2}{2}\mathbb{E}\left[\left(X^{(\epsilon)}\right)^2\right] + O(\epsilon^3). \tag{3.5}$$

Since we are working with a bounded random variable, the proof of Lemma 3.1 is straightforward. However, in general, one needs an assumption on the existence of the MGF. We present the proof below.

Proof of Lemma 3.1. Fix $\Theta > 0$ and $x \in \mathbb{R}$. Then, from Taylor approximation of $f_{\epsilon,x}(\theta) =$

$e^{\theta\epsilon x}$ at $\theta = 0$ we have

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

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

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

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

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

which is finite for every ϵ. Then,

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

Therefore,

$$\left|e^{\theta\epsilon x} - 1 - \theta\epsilon x - \frac{(\theta\epsilon)^2}{2}x^2\right| \leq \zeta_1\epsilon^3,$$

where $\zeta_1 = \frac{\Theta^3\kappa^3}{3!}$ is a finite constant. Now, since $X^{(\epsilon)}$ is bounded, we know the existence of $\kappa_{\max}$ such that $X^{(\epsilon)} \leq \kappa_{\max}$ with probability 1. Therefore,

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

which proves the lemma. □

Expanding each term on the right-hand side of Equation 3.4 in Taylor series according to Lemma 3.1 will yield terms related to the moments of the unused service. As illustrated in the case of the single server queue, it suffices to handle the first moment. To do that, we

set to zero the drift of the linear test function $V_1(\boldsymbol{q}) = \langle \boldsymbol{c}, \boldsymbol{q} \rangle$, i.e., we set $\mathbb{E}\left[\langle \boldsymbol{c}, \left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^+ \rangle\right] = \mathbb{E}\left[\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle\right]$ (which is finite because in Step 1 it was proved that the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ exists in an interval around 0). For example, see Lemma 2.13 in the case of the single server queue, which is used in Equation 2.8.

From this step we obtain an expression for the limit as $\epsilon \downarrow 0$ of the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$. This proves convergence in distribution of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ to a random variable Y, which turns out to be exponential in the cases we study in this chapter. Then, $\epsilon\overline{\boldsymbol{q}}_{\|}^{(\epsilon)} = \epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle \boldsymbol{c} \Rightarrow Y\boldsymbol{c}$ as $\epsilon \downarrow 0$ because $\boldsymbol{c}$ is a fixed vector. We also know from SSC in Prerequisite 2 that $\epsilon\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)} \to 0$ in probability as $\epsilon \downarrow 0$. Then, by Slutsky's theorem [40, Theorem 11.4 in Section 5], we obtain that $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} = \epsilon\overline{\boldsymbol{q}}_{\|}^{(\epsilon)} + \epsilon\boldsymbol{q}_{\perp}^{(\epsilon)} \Rightarrow Y\boldsymbol{c}$ as $\epsilon \downarrow 0$.

Remark 3.2. *In order to set* $\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}, \left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^+ \rangle}\right] = \mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle}\right]$ *in Step 1, one must first prove the existence of the MGF of* $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ *in an interval around zero. An alternative approach (where this difficulty does not arise), is to use characteristic functions, because they always exist. However, working with characteristic functions involves the use of complex analysis. Another way to overcome this difficulty is to use one-sided Laplace transform, i.e., to consider* $\theta < 0$. *One-sided Laplace transform of* $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$ *always exists because* ϵ, $\boldsymbol{c}$ *and* $\overline{\boldsymbol{q}}^{(\epsilon)}$ *are nonnegative. If one chooses to work with other transforms such as the characteristic function or one-sided Laplace transform to get around the issue of the existence of the MGF, then one needs to assume that certain moments exist in a counterpart of Lemma 3.1. For instance, Theorem 2.3.3. in [65] can be used when one is working with characteristic functions.*

3.4 Load balancing system model

Consider a system with n separate queues, as described in section 1.5. For each $i \in [n]$, $\{s_i(k) : k \in \mathbb{Z}_+\}$ is a sequence of i.i.d. random variables with $\mu_i \stackrel{\triangle}{=} \mathbb{E}[s_i(1)]$, and let $\mu_\Sigma \stackrel{\triangle}{=} \sum_{i=1}^{n} \mu_i$. We consider this system in a general setting, so we do not assume independence of

the servers. Let Σ_s be the covariance matrix of $\boldsymbol{s}(1)$. Then, for each pair $i, j \in [n]$, we have $(\Sigma_s)_{i,j} = \text{Cov}\left[s_i(1), s_j(1)\right]$.

There is a single stream of arrivals, that we model as a sequence $\{a(k) : \ k \in \mathbb{Z}_+\}$ of i.i.d. random variables such that $a(k)$ is the number of arrivals to the system in time slot k. In this queueing 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 $k \in \mathbb{Z}_+$, but we do not assume any particular policy. After routing, $a_i(k)$ is the number of arrivals routed to the i^{th} queue in time slot k, for $i \in [n]$. We assume $a(k) \leq A_{\max}$ with probability 1 for all $k \in \mathbb{Z}_+$, and that the arrival process is independent of the queue length and service processes. The dynamics of the queues are according to Equation 1.2. It is well known that the capacity region of the load balancing system is $\mathcal{C} = \{\lambda \in \mathbb{R}_+ : \ \lambda \leq \mu_\Sigma\}$. A proof can be found in Appendix A of [34].

3.5 MGF method applied to load balancing systems

In this section we use the MGF method in the context of load balancing systems, also known as supermarket checkout systems. The model is described in section 3.4.

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 with arrival process $\{a^{(\epsilon)}(k) : \ k \in \mathbb{Z}_+\}$, that satisfies $\mathbb{E}\left[a^{(\epsilon)}(1)\right] = \mu_\Sigma - \epsilon$ and $\text{Var}\left[a^{(\epsilon)}(1)\right] = \left(\sigma_a^{(\epsilon)}\right)^2$. In other words, the arrival rate approaches the point $\nu = \mu_\Sigma$ in the boundary of $\mathcal{C}$ as $\epsilon \downarrow 0$. Since the capacity region $\mathcal{C}$ of the load balancing system is one-dimensional, the CRP condition (as defined in Definition 2.1) is trivially satisfied.

In the rest of this section, we state the main theorem of this section and provide some examples. We prove the result using the MGF method as developed in section 3.3 in subsection 3.5.1. Before presenting the formal statement of the result we introduce the following definitions.

Definition 3.3 (Throughput optimality)**.** *A routing algorithm $\mathcal{A}$ is throughput optimal for*

the load balancing system described in section 3.4, if the Markov chain $\left\{\boldsymbol{q}^{(\epsilon)}(k) : \ k \in \mathbb{Z}_+\right\}$ *operating under* $\mathcal{A}$ *is positive recurrent for all* $\epsilon \in (0, \mu_\Sigma)$.

Definition 3.4 (State Space Collapse)**.** *Consider a routing algorithm* $\mathcal{A}$ *and let*

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

i.e., $\boldsymbol{c} = \frac{1}{\sqrt{n}}\mathbf{1}$. *For any vector* $\boldsymbol{y} \in \mathbb{R}^n$, *let* $\boldsymbol{y}_\parallel$ *be the projection of* $\boldsymbol{y}$ *on* $\mathcal{K}$ *and let* $\boldsymbol{y}_\perp \triangleq \boldsymbol{y} - \boldsymbol{y}_\parallel$. *We say that the algorithm* $\mathcal{A}$ *satisfies SSC if the load balancing system described in section 3.4 operating under* $\mathcal{A}$ *satisfies the following property.*

$$\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^2\right] \ \textit{is } o\left(\frac{1}{\epsilon^2}\right),$$

where $\overline{\boldsymbol{q}}^{(\epsilon)}$ *is a steady-state random vector such that* $\left\{\boldsymbol{q}^{(\epsilon)}(k) : \ k \in \mathbb{Z}_+\right\}$ *converges in distribution to* $\overline{\boldsymbol{q}}^{(\epsilon)}$ *if it is positive recurrent.*

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

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

Theorem 3.5. *Let* $\epsilon \in (0, \mu_\Sigma)$ *and consider a set of load balancing systems as described in section 3.4, parametrized by* ϵ *as described above. Suppose that the routing algorithm is throughput optimal and that it satisfies SSC. For each* $\epsilon \in (0, \mu_\Sigma)$, *let* $\overline{\boldsymbol{q}}^{(\epsilon)}$ *be a steady-state random vector such that the queue length process* $\{\boldsymbol{q}^{(\epsilon)}(k) : \ k \in \mathbb{Z}_+\}$ *converges in distribution to* $\overline{\boldsymbol{q}}^{(\epsilon)}$. *Assume the MGF of* $\epsilon \sum_{i=1}^n \overline{q}_i$ *exists, i.e.,* $\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^n \overline{q}_i^{(\epsilon)}}\right] < \infty$ *for* $\theta \in [-\Theta, \Theta]$ *where* $\Theta > 0$ *is a finite number, and that* $\lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a$. *Then* $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \Longrightarrow \widetilde{\Upsilon}\mathbf{1}$ *as* $\epsilon \downarrow 0$, *where* $\widetilde{\Upsilon}$ *is an exponential random variable with mean* $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$.

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

Definition 3.6 (JSQ and Power-of-two choices)**.** *Consider a load balancing system as described in section 3.4. Then, for each $k \in \mathbb{Z}_+$, given the vector of queue lengths $\boldsymbol{q}^{(\epsilon)}(k)$, a routing policy selects $i^*(k)$ routes the arrivals so that $\boldsymbol{a}^{(\epsilon)}(k) = \boldsymbol{e}^{(i^*(k))}a(k)$.*

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

$$i^*(k) \in \arg\min_{i\in[n]} \left\{ q_i^{(\epsilon)}(k) \right\},$$

breaking ties at random.

(b) The routing policy power-of-two choices selects two queues uniformly at random, say $i_1, i_2 \in [n]$ 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, then

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

breaking ties at random.

In the following two corollaries we show that these routing policies satisfy the assumptions of Theorem 3.5 and, therefore, the scaled vector of queue lengths in a load balancing system operating under any of these policies has an exponential distribution in heavy-traffic.

Corollary 3.7. *Consider a set of load balancing systems as described in section 3.4, parametrized by $\epsilon \in (0, \mu_\Sigma)$ as described above. Suppose the routing algorithm is JSQ. Then, $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \implies \widetilde{\Upsilon}_1\mathbf{1}$ as $\epsilon \downarrow 0$, where $\widetilde{\Upsilon}_1$ is an exponential random variable with mean $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$.*

A particular case of the queueing system described in Corollary 3.7 is the load balancing system operating under JSQ with independent servers. In this case, $\mathbf{1}^T\Sigma_S\mathbf{1}$ reduces to the sum of variances of the servers. This is one of the systems studied in [34].

Proof of Corollary 3.7. We only need to show that JSQ is throughput optimal, that it satisfies SSC, and that there exists $\Theta > 0$ such that $\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^n \overline{q}_i^{(\epsilon)}}\right] < \infty$ for all $\theta \in [-\Theta, \Theta]$. In [34], the authors 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 [34].

The SSC result proved [34] is stronger than the property presented in Definition 3.4. In fact, they prove that $\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^j\right]$ is upper bounded by a constant for each $j \in \mathbb{Z}_+$ with $j \geq 1$. This clearly implies that Definition 3.4 is satisfied. We provide a sketch of their proof of SSC in subsection 3.6.1.

The existence of MGF of $\epsilon\sum_{i=1}^n \overline{q}_i^{(\epsilon)}$ in an interval around 0 is proved in subsection 3.6.2. □

Corollary 3.8. *Consider a set of load balancing systems as described in section 3.4, parametrized by ϵ as described above. Suppose the routing algorithm is power-of-two choices and that all the servers are identical. Then, $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \Longrightarrow \widetilde{\Upsilon}_2\mathbf{1}$ as $\epsilon \downarrow 0$, where $\widetilde{\Upsilon}_2$ is an exponential random variable with mean $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$.*

Proof of Corollary 3.8. Similar to the proof of Corollary 3.7, we check throughput optimality, SSC and existence of MGF. SSC is proved in [63, Section 4.3] in the case of independent servers, but their proof holds true if this assumption is dropped. Their proof is along the lines of the proof for JSQ in subsection 3.6.1, so we do not present it here. Throughput optimality can be proved using the Foster-Lyapunov theorem (Theorem 2.4) and the calculations developed in [63] in the proof of SSC. The proof of existence of MGF is similar to the case of JSQ. We omit these proofs, since our goal is to introduce the MGF method. □

Observe that the assumption of identical servers is essential for the power-of-two choices algorithm to be throughput optimal. The case when the servers are not identical was studied in [55] using the diffusion limits approach. The routing policy there randomly selects d servers in each time slot, where the probability of choosing server i is proportional to its service rate μ_i, for all $i \in [n]$. 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 stated in Definition 3.4 is established. This is straightforward extension, and we do not present the details here because the focus is on illustrating the MGF approach.

In chapter 4 we provide necessary and sufficient conditions on the vector of service rates to ensure that power-of-d choices is throughput optimal in load balancing systems with heterogeneous servers. We additionally show SSC under similar conditions, and we compute the distribution of the vector of queue lengths in heavy traffic using the MGF method.

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

3.5.1 Proof of Theorem 3.5

In the rest of this section we prove Theorem 3.5 using the MGF method. Before presenting the proof we specify notation.

Let $\overline{a}^{(\epsilon)}$ be a steady-state random variable with the same distribution as $a^{(\epsilon)}(1)$ and let $\overline{\boldsymbol{a}}^{(\epsilon)} \stackrel{\triangle}{=} \boldsymbol{a}^{(\epsilon)}(\overline{\boldsymbol{q}})$ be the vector of arrivals to each queue after routing in steady-state.

The vector $\overline{\boldsymbol{u}}^{(\epsilon)}$ represents the unused service. Observe that, in this case, the vector $\overline{\boldsymbol{s}}$ is independent of $\overline{\boldsymbol{q}}^{(\epsilon)}$ and has the same distribution as $\boldsymbol{s}(1)$, because the potential service sequences $\{s_i(k) : k \in \mathbb{Z}_+\}$ are i.i.d. and independent of the queue length process.

Proof of Theorem 3.5. For ease of exposition, we omit the dependence on ϵ of the variables in this proof. We use the MGF method. Before applying the steps, we need to verify that the prerequisites are satisfied, i.e., we need to check positive recurrence and SSC. In fact, one of the assumptions of the theorem is that the routing policy is throughput optimal. Therefore, for any $\epsilon > 0$ the Markov chain $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent. Also, SSC is satisfied by assumption. Now we go through the steps of the MGF method.

Step 1. Prove an equation of the form of Equation 3.1 and compute an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$.

We first prove the following lemma.

Lemma 3.9. *Consider a load balancing system parametrized by ϵ as described in Theorem 3.5. Then, there exists $\theta_{\max} > 0$ finite such that for any real number $\theta \in [-\theta_{\max}, \theta_{\max}]$ we have*

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

We present the proof of Lemma 3.9 in subsection 3.6.3.

Since $\langle \boldsymbol{c}, \boldsymbol{q}\rangle = \dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^{n} q_i$, proving an equation of the form of Equation 3.1 is equivalent to Lemma 3.9 using $\frac{\theta}{\sqrt{n}}$ instead of θ. For ease of exposition, we work with θ in the rest of this proof.

Note that $\mathbb{P}\left[\overline{a} - \sum_{i=1}^{n}\overline{s}_i \neq 0\right] > 0$ whenever $\epsilon > 0$. If we expand the product in the expression of Lemma 3.9 and we follow the steps sketched after Step 1 in Section

section 3.3 we obtain

$$\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i}\left(1-e^{\theta\epsilon\sum_{i=1}^{n}(\overline{a}_i-\overline{s}_i)}\right)\right]=1-\mathbb{E}\left[e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}\right]+o(\epsilon^2). \tag{3.6}$$

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

$$\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i}\right]=\frac{1-\mathbb{E}\left[e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}\right]+o(\epsilon^2)}{1-\mathbb{E}\left[e^{\theta\epsilon\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)}\right]}, \tag{3.7}$$

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

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

Equation 3.7 yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$, just like Equation 2.7 in the case of the single server queue. Equivalently, we can observe that Equation 3.6 yields $0 = 0$ in the limit as $\epsilon \downarrow 0$. Then, we take Taylor series of the numerator and the denominator of Equation 3.7 at $\theta = 0$ to obtain the limit. To take Taylor expansion we use Lemma 3.1.

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

Lemma 3.10. *Consider a load balancing system as described in section 3.4, parametrized by $\epsilon \in (0, \mu_\Sigma)$ as described at the beginning of section 3.5, operating under a throughput optimal routing policy. Then,*

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

Proof of Lemma 3.10. We set to zero the drift of $V_1(\boldsymbol{q}) = \langle \boldsymbol{c}, \boldsymbol{q}\rangle$ in steady state. In this case, from the definition of $\mathcal{K}$ in Definition 3.4 we have $\boldsymbol{c} = \frac{1}{\sqrt{n}}\mathbf{1}$. Then, we obtain

$$0 = \mathbb{E}\left[V_1\left(\overline{\boldsymbol{q}}^{+}\right) - V_1\left(\overline{\boldsymbol{q}}\right)\right]$$

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

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

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

where (a) holds because $\mathbb{E}\left[\overline{a}\right] = \mu_\Sigma - \epsilon$; and (b) holds by definition of μ_Σ. □

Now we expand the numerator and denominator of Equation 3.7 in Taylor series. We start with the numerator, and we obtain

$$
\begin{aligned}
1 - \mathbb{E}\left[e^{-\theta\epsilon \sum_{i=1}^{n} \overline{u}_i}\right] &= 1 - \mathbb{E}\left[f_{\epsilon, -\sum_{i=1}^{n} \overline{u}_i}(\theta)\right] \\
&= \theta\epsilon \mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] - \frac{(\theta\epsilon)^2}{2} \mathbb{E}\left[\left(\sum_{i=1}^{n} \overline{u}_i\right)^2\right] + O(\epsilon^3) \\
&= \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2} \mathbb{E}\left[\left(\sum_{i=1}^{n} \overline{u}_i\right)^2\right] + O(\epsilon^3), \qquad (3.8)
\end{aligned}
$$

where the last equality holds by Lemma 3.10. Now we need to bound the second moment of the sum of unused services.

Claim 3.11. *Consider a load balancing system as described in Theorem 3.5. Then,*

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

We prove the claim in subsection 3.6.4. Using the Claim 3.11 in Equation 3.8 we obtain

$$1 - \mathbb{E}\left[e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}\right] = \theta\epsilon^2 + O(\epsilon^3), \tag{3.9}$$

For the denominator, we obtain

$$\begin{aligned}
&1 - \mathbb{E}\left[e^{\theta\epsilon\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)}\right] \\
&= 1 - \mathbb{E}\left[f_{\epsilon,\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)}(\theta)\right] \\
&= -\theta\epsilon\mathbb{E}\left[\overline{a} - \sum_{i=1}^{n}\overline{s}_i\right] - \frac{(\theta\epsilon)^2}{2}\mathbb{E}\left[\left(\overline{a} - \sum_{i=1}^{n}\overline{s}_i\right)^2\right] + O(\epsilon^3) \\
&= \theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathrm{Cov}\left[s_i, s_{i'}\right] + \epsilon^2\right) + O(\epsilon^3),
\end{aligned} \tag{3.10}$$

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

Using Equation 3.9 and Equation 3.10 in Equation 3.7, and since $O(\epsilon^3)$ is $o(\epsilon^2)$, we obtain

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

Canceling $\theta\epsilon^2$ from the numerator and denominator, and noticing that $\sum_{i=1}^{n}\sum_{i'=1}^{n}\mathrm{Cov}\left[s_i, s_{i'}\right] = \mathbf{1}^T\Sigma_s\mathbf{1}$, we obtain

$$\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i}\right] = \frac{1 + o(1)}{1 - \frac{\theta}{2}\left(\left(\sigma_a^{(\epsilon)}\right)^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right) + O(\epsilon)}.$$

Therefore, taking the limit we obtain

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i}\right] = \frac{1}{1 - \frac{\theta}{2}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)},$$

which is the MGF of an exponential random variable with mean $\frac{1}{2}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$. Then, $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}\rangle \boldsymbol{c} = \epsilon\left(\frac{1}{n}\sum_{i=1}^n \overline{q}_i\right)\mathbf{1} \Rightarrow \widetilde{\Upsilon}\mathbf{1}$ as $\epsilon \downarrow 0$, where $\widetilde{\Upsilon}$ is an exponential random variable with mean $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$. Therefore, we conclude that $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} = \epsilon\overline{\boldsymbol{q}}_{\parallel}^{(\epsilon)} + \epsilon\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)} \Rightarrow \widetilde{\Upsilon}\mathbf{1}$ as $\epsilon \downarrow 0$. This proves Theorem 3.5. □

3.6 Details of the proofs of section 3.5

In this section we provide the details of the proofs from section 3.5.

3.6.1 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 [34]. 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 3.12. *Consider a load balancing system as described in Corollary 3.7. Then, for each $j \in \mathbb{Z}_+$ with $j \geq 1$ there exists a finite constant J_j such that*

$$\mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^j\right] \leq J_j.$$

This proof is based on Lemma 2.7.

Proof of Proposition 3.12. In [34], the authors use the Lyapunov function $Z(\boldsymbol{q}) = \|\boldsymbol{q}_{\perp}^{(\epsilon)}\|$ and they prove that

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

where δ is a fixed constant in $(0, \mu_{\min})$. The proof is based on the fact that $\|\boldsymbol{x}\| = \sqrt{\|\boldsymbol{x}\|^2}$, that square root is a concave function and that JSQ sends all arrivals to the shortest queue in each time slot. This verifies condition (C1) of Lemma 2.7.

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

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

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

Also, for $\epsilon > 0$ the Markov Chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent and, since projection is nonexpansive, we have $\|\boldsymbol{q}^{(\epsilon)}_\perp(k)\| \leq \|\boldsymbol{q}^{(\epsilon)}(k)\|$. Hence, $\{\boldsymbol{q}^{(\epsilon)}_\perp(k) : k \in \mathbb{Z}_+\}$ is positive recurrent, which proves the result. □

3.6.2 Existence of MGF of $\epsilon \sum_{i=1}^n \overline{q}_i^{(\epsilon)}$ in the load balancing system operating under JSQ

We first state the result formally.

Lemma 3.13. *Consider a load balancing system operating under JSQ, parametrized by $\epsilon \in (0, \mu_\Sigma)$ as described in Corollary 3.7. Then, there exists $\Theta > 0$ (which is independent of ϵ) such that $\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^n \overline{q}_i^{(\epsilon)}}\right] < \infty$ for all $\theta \in [-\Theta, \Theta]$.*

Proof of Lemma 3.13. We omit the dependence on ϵ of the variables for ease of exposition. First observe that if $\theta \leq 0$, then $\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^n \overline{q}_i}\right] < \infty$ trivially because $\overline{\boldsymbol{q}} \geq \mathbf{0}$ by definition of queue length.

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

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

Hence, it suffices to show that $\sum_{i=1}^n \mathbb{E}\left[e^{\theta\epsilon\overline{q}_i}\right] < \infty$ for $\theta \leq \Theta$. We show that the conditions of the Foster-Lyapunov theorem (Theorem 2.4) are satisfied with Lyapunov function $V(\boldsymbol{q}) = \sum_{i=1}^n e^{\theta\epsilon q_i}$. Then, one can easily obtain a finite bound on $\mathbb{E}\left[V(\overline{\boldsymbol{q}})\right]$ [37, Proposition 6.14].

Using Lemma 2.12 for each of the n queues and rearranging terms we obtain that, for

each $i \in [n]$ and $k \in \mathbb{Z}_+$

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

Then, using the notation $\mathbb{E}_{\boldsymbol{q}}[\cdot] \stackrel{\triangle}{=} \mathbb{E}[\cdot \mid \boldsymbol{q}(k) = \boldsymbol{q}]$, we obtain

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

Since $\mathbb{E}_{\boldsymbol{q}}\left[e^{-\theta\epsilon u_i(k)}\right] \geq 0$, we have

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

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

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

Given $\boldsymbol{q}(k) = \boldsymbol{q}$, let $i^* \in \arg\min_{i\in[n]} \{q_i(k)\}$ be the queue where arrivals in time slot k are routed. Then,

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

where we used the notation $M_X(\theta) = \mathbb{E}\left[e^{\theta\epsilon X}\right] - 1$. Next we take the truncated Taylor

series around 0. Then, for some $\xi_i \in (0, \theta)$ for each $i \in [n]$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} e^{\theta\epsilon q_i} \left(\mathbb{E}_{\boldsymbol{q}}\left[e^{\theta\epsilon(a_i(k)-s_i(k))}\right] - 1\right) \\
&\overset{(a)}{=} e^{\theta\epsilon q_{i^*}} \left(\theta\epsilon(\lambda - \mu_{i^*}) + \frac{\theta^2}{2} M''_{a-s_{i^*}}(\xi_{i^*})\right) + \sum_{\substack{i=1 \\ i\neq i^*}}^{n} e^{\theta\epsilon q_i} \left(-\theta\epsilon\mu_i + \frac{\theta^2}{2} M''_{s_i}(\xi_i)\right) \\
&\overset{(b)}{\leq} e^{\theta\epsilon q_{i^*}} \theta\epsilon(\lambda - \mu_{i^*}) - \sum_{\substack{i=1 \\ i\neq i^*}}^{n} e^{\theta\epsilon q_i} \theta\epsilon\mu_i + \frac{\theta^2\epsilon^2}{2}(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}} \sum_{i=1}^{n} e^{\theta\epsilon q_i}, \qquad (3.11)
\end{aligned}$$

where (a) holds by definition of i^* and the routing algorithm; and (b) holds after bounding the second derivatives as follows. Since $\epsilon < \mu_\Sigma$ by definition, we have

$$\begin{aligned}
M''_{a-s_{i^*}}(\theta) &= \mathbb{E}\left[\epsilon^2(a - s_{i^*})^2 e^{\theta\epsilon(a-s_{i^*})}\right] \leq \epsilon^2(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}} \\
M''_{-s_i}(\theta) &= \mathbb{E}\left[\epsilon^2 s_i^2 e^{-\theta\epsilon s_i}\right] \leq \epsilon^2 S_{\max}^2 \leq \epsilon^2(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}}.
\end{aligned}$$

For each $i \in [n]$, define $\lambda_i \overset{\triangle}{=} \mu_i - \frac{\epsilon}{n}$, and observe $\lambda = \sum_{i=1}^{n} \lambda_i$. Then, we obtain

$$\begin{aligned}
&\textit{Equation } 3.11 \\
&= \theta\epsilon \sum_{i=1}^{n} e^{\theta\epsilon q_i}(\lambda_i - \mu_i) + \frac{\theta^2\epsilon^2}{2}(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}} \sum_{i=1}^{n} e^{\theta\epsilon q_i} + \theta\epsilon \sum_{i=1}^{n} \lambda_i \left(e^{\theta\epsilon q_{i^*}} - e^{\theta\epsilon q_i}\right) \\
&\overset{(a)}{\leq} \theta\epsilon^2 \left(-\frac{1}{n} + \frac{\theta}{2}(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}}\right) \sum_{i=1}^{n} e^{\theta\epsilon q_i},
\end{aligned}$$

where (a) holds because $q_{i^*} \leq q_i$ for all $i \in [n]$ by definition of i^*, because $\lambda_i - \mu_i = -\frac{\epsilon}{n}$ and reorganizing terms. Therefore, it suffices to show the existence of $\Theta > 0$ such that

$$-\frac{1}{n} + \frac{\theta}{2}(A_{\max}^2 + S_{\max}^2) e^{\theta\mu_\Sigma A_{\max}} \leq -\frac{1}{2n} \quad \forall \theta \leq \Theta.$$

Solving the inequality yields

$$\Theta = \frac{1}{\mu_\Sigma A_{\max}} W_0\left(\frac{A_{\max}\mu_\Sigma}{n(A_{\max}^2+S_{\max}^2)}\right),$$

where $W_0(\,\cdot\,)$ is the principal branch of the Lambert W function, which has been studied in [66] among others. Observe that Θ is independent of ϵ. This completes the proof. □

3.6.3 Proof of Lemma 3.9

To prove Lemma 3.9 we use the following result.

Lemma 3.14. *Consider the load balancing system indexed by ϵ described in Theorem 3.5. Then, for any $\alpha \in \mathbb{R}$ and for all $k \in \mathbb{Z}_+$ we have*

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

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

Proof of Lemma 3.14. If $\alpha = 0$, the equation trivially holds. So now assume $\alpha \neq 0$. Since $q_i(k+1)u_i(k) = 0$ for all $i \in [n]$, we have

$$u_i(k)(e^{-\alpha q_i(k+1)} - 1) = 0 \quad \forall i \in [n].$$

Then, summing over $i \in [n]$ we obtain

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

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

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

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

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

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

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

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

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

□

In the proof of Lemma 3.9 we use Lemma 3.14 and the following facts:

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

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

(iii) For all $x \in \mathbb{R}_+$, $\dfrac{e^x - 1}{x} < e^x$

All these facts can be shown using calculus techniques, so we omit the proof. Now we prove Lemma 3.9.

Proof of Lemma 3.9. First observe that if $\theta = 0$ the statement trivially holds. If $\theta \neq 0$, by properties of expectation and absolute value we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i^{+}} - 1\right)\left(e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i} - 1\right)\right]\right| \\
\leq &\mathbb{E}\left[\left|\left(e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i^{+}} - 1\right)\left(e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i} - 1\right)\right|\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=}|\theta|\epsilon\mathbb{E}\left[\left|\left(\sum_{i=1}^{n}\overline{u}_i\right)\left(e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i^{+}}-1\right)\left(\frac{e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}-1}{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}\right)\right|\mathbb{1}_{\left\{\sum_{i=1}^{n}\overline{u}_i\neq 0\right\}}\right]\\
&\overset{(b)}{\leq}|\theta|\epsilon\left(\frac{e^{|\theta|\epsilon nS_{\max}}-1}{|\theta|\epsilon nS_{\max}}\right)\mathbb{E}\left[\left|\sum_{i=1}^{n}\overline{u}_i\left(e^{\theta\epsilon\sum_{j=1}^{n}\overline{q}_j^{+}}-1\right)\right|\right]\\
&\overset{(c)}{\leq}|\theta|\epsilon\left(\frac{e^{|\theta|\epsilon nS_{\max}}-1}{|\theta|\epsilon nS_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i\left|e^{-\theta\epsilon n\overline{q}_{\perp i}}-1\right|\right]\\
&\overset{(d)}{\leq}|\theta|\epsilon\left(\frac{e^{|\theta|\epsilon S_{\max}}-1}{|\theta|\epsilon S_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i^{j}\right]^{\frac{1}{j}}\mathbb{E}\left[\sum_{i=1}^{n}\left|e^{-\theta\epsilon n\overline{q}_{\perp i}}-1\right|^{\frac{j}{j-1}}\right]^{\frac{j-1}{j}}\\
&\overset{(e)}{\leq}|\theta|\epsilon^{1+\frac{1}{j}}S_{\max}^{\frac{j-1}{j}}\left(\frac{e^{|\theta|\epsilon S_{\max}}-1}{|\theta|\epsilon S_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\left|e^{-\theta\epsilon n\overline{q}_{\perp i}}-1\right|^{\frac{j}{j-1}}\right]^{\frac{j-1}{j}}\\
&=\theta^2\epsilon^{2+\frac{1}{j}}S_{\max}^{\frac{j-1}{j}}n\left(\frac{e^{|\theta|\epsilon S_{\max}}-1}{|\theta|\epsilon S_{\max}}\right)\left(\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{e^{-\theta\epsilon n\overline{q}_{\perp i}}-1}{-\theta\epsilon n\overline{q}_{\perp i}}\right|^{\frac{j}{j-1}}|\overline{q}_{\perp i}|^{\frac{j}{j-1}}\mathbb{1}_{\{\overline{q}_{\perp i}\neq 0\}}\right]\right)^{\frac{j-1}{j}},
\end{aligned}
\tag{3.12}
$$

where $j\in\mathbb{Z}_+$ satisfies $j>1$. Here (a) holds because if $\sum_{i=1}^{n}\overline{u}_i=0$ then $e^{-\theta\epsilon\sum_{i=1}^{n}\overline{u}_i}-1=0$, and by multiplying and dividing everything by $|\theta\epsilon\sum_{i=1}^{n}\overline{u}_i|$; (b) holds by the fact item (i) stated above, because $\overline{u}_i\leq S_{\max}$ for all $i\in[n]$ and because $0\leq\mathbb{1}_{\left\{\sum_{i=1}^{n}\overline{u}_i\neq 0\right\}}\leq 1$; (c) holds by triangle inequality and Lemma 3.14; (d) holds by Hölder's inequality; and (e) holds because $\overline{u}_i\leq S_{\max}$ for all $i\in[n]$, because $\sum_{i=1}^{n}\mathbb{E}\left[\overline{u}_i\right]=\epsilon$ and because $x^{\frac{1}{p}}$ is an increasing function for $x\geq 0$.

By L'Hospital's rule we have

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

Then, the last step is to prove that the last expression in Equation 3.12 is $O(1)$. To do that we show the following claim at the end of this section.

Claim 3.15. *Consider a load balancing system as described in Lemma 3.9. Then, there*

exists $\theta_{\max} > 0$ *finite such that for all* $|\theta| < \theta_{\max}$ *we have*

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

An expression for $\theta_{\max}$ *is provided in Equation 3.13.*

Therefore,

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

□

Now we prove the claim.

Proof of Claim 3.15. By Hölder's inequality, for each $i \in [n]$

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{e^{-\theta\epsilon\overline{q}_{\perp i}}-1}{-\theta\epsilon\overline{q}_{\perp i}}\right|^{\frac{j}{j-1}} |\overline{q}_{\perp i}|^{\frac{j}{j-1}} \mathbb{1}_{\{\overline{q}_{\perp i}\neq 0\}}\right] \\
&\leq \mathbb{E}\left[\left|\frac{e^{-\theta\epsilon\overline{q}_{\perp i}}-1}{-\theta\epsilon\overline{q}_{\perp i}}\right|^{\left(\frac{j}{j-1}\right)\left(\frac{j'}{j'-1}\right)} \mathbb{1}_{\{\overline{q}_{\perp i}\neq 0\}}\right]^{\frac{j'-1}{j'}} \mathbb{E}\left[|\overline{q}_{\perp i}|^{\left(\frac{j}{j-1}\right)j'}\right]^{\frac{1}{j'}},
\end{aligned}$$

where $j' \in \mathbb{Z}_+$ satisfies $j' > 1$. On one hand, we can choose j large so that $\frac{j}{j-1} \approx 1$, and $j' > 1$ such that $\left(\frac{j}{j-1}\right) j' = 2$. Then, $\mathbb{E}\left[|\overline{q}_{\perp i}|^{\left(\frac{j}{j-1}\right)j'}\right]^{\frac{1}{j'}}$ is $O(1)$ by SSC.

Also, by SSC we know that $\epsilon|\overline{q}_{\perp i}|$ converges to zero in the mean-square sense and, therefore, in distribution. Then, by the continuous mapping theorem [40, Theorem 10.4 in Section 5] we have that

$$\left(\frac{e^{-\theta\epsilon|\overline{q}_{\perp i}|}-1}{-\theta\epsilon|\overline{q}_{\perp i}|}\right)^{\left(\frac{j}{j-1}\right)\left(\frac{j'}{j'-1}\right)} \Longrightarrow 1.$$

It remains to prove that $\mathbb{E}\left[\frac{e^{-\theta\epsilon|\overline{q}_{\perp i}|}-1}{-\theta\epsilon|\overline{q}_{\perp i}|}\right]$ is finite to conclude that its expected value also

converges to 1. In fact, we have

$$-\theta\epsilon|\overline{q}_{\perp i}| \leq |\theta|\epsilon|\overline{q}_{\perp i}| \leq |\theta|\epsilon\|\overline{\boldsymbol{q}}_{\perp}\|$$

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

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

Therefore,

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

where (a) holds because projection is nonexpansive; (b) holds because norm-1 is greater than Euclidean norm; and (c) holds by assumption of Theorem 3.5, choosing

$$|\theta|\left(\frac{j}{j-1}\right)\left(\frac{j'}{j'-1}\right) \leq \Theta.$$

Then, the claim holds with

$$\theta_{\max} = \Theta\left(\frac{j'-1}{2}\right), \tag{3.13}$$

where we used that $\left(\frac{j}{j-1}\right)j' = 2$. This completes the proof. □

3.6.4 Proof of Claim 3.11

Proof of Claim 3.11. We have

$$0 \leq \frac{(\theta\epsilon)^2}{2}\mathbb{E}\left[\left(\sum_{i=1}^{n}\overline{u}_i\right)^2\right] \overset{(a)}{\leq} \epsilon^2\left(\frac{nS_{\max}\theta^2}{2}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i\right] \overset{(b)}{=} \epsilon^3\left(\frac{nS_{\max}\theta^2}{2}\right)$$

where (a) holds because, by definition of unused service, we have $\overline{u}_i \leq \overline{s}_i \leq S_{\max}$ and all terms are nonnegative; and (b) holds by Lemma 3.10.

Therefore,

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

□

3.7 Conclusion and future work

In this chapter we introduced transform methods to compute the steady-state distribution of the scaled queue lengths in heavy traffic. We focused on two-sided Laplace transform, which is also known as MGF. We motivated the method with a single server queue in section 2.5 and we applied it in the load balancing system under the CRP condition. In chapter 6 we also apply it to 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.

We discuss future work directions at the end of chapter 6 (see section 6.6).

CHAPTER 4

POWER-OF-D CHOICES UNDER HETEROGENEOUS SERVERS

Based on:

D. Hurtado-Lange and S. T. Maguluri, "Throughput and delay optimality of power-of-d choices in inhomogeneous load balancing systems," *Operations Research Letters*, 2021

4.1 Introduction

Two popular routing algorithms for the load balancing system are join the shortest queue (JSQ) and power-of-d choices. Under JSQ, the arrivals of each time slot are routed to the server with the shortest queue among all. A disadvantage of JSQ is that, if the number of servers is high, finding the shortest queue among all may take considerable time. However, JSQ only uses the state of the system (queue length vector) and does not require any information about the parameters of the system, such as the service rates.

Under power-of-d choices, one samples d queues uniformly at random, and routes the arrivals to the shortest queue among these d. In this case, we also do not use information about the service rates, but if the servers are not equal, the queue length vector may not be stable. In other words, if the servers are heterogeneous we may have infinite queue lengths in steady state, even when the arrival rate is strictly less than the total service rate. Hence, under heterogeneous servers, power-of-d choices may not be throughput optimal.

The primary contribution of this chapter is the computation of necessary and sufficient conditions for throughput optimality of power-of-d choices, that only depend on the mean service rate vector. Specifically, we characterize a polytope where the service rate vectors should lie. In particular, if the servers are identical our conditions are satisfied. Our result formalizes the idea that, in order to have throughput optimality, all the queues need to be sampled frequently enough. Then, given that power-of-d selects d queues uniformly at

random, our result implies that the service rates of different servers should be close to each other; but not necessarily equal.

The second contribution of this chapter is the computation of the distribution of the scaled vector of queue lengths in heavy traffic. We show that, if the heterogeneous service rates lie in the interior of the polytope proposed for throughput optimality, the load balancing system operating under power-of-d choices has the same limiting distribution as a load balancing system operating under JSQ. Therefore, our results imply that power-of-d choices is heavy-traffic optimal.

The third contribution of this chapter is a sufficient condition for throughput optimality under a larger class of routing policies. Specifically, we consider the following generalization of power-of-d choices. In power-of-d choices, only sets of size d are sampled, and all of them are observed with the same probability. In the last part of this chapter, we consider a routing policy that selects any subset of servers with certain probability, and routes the arrivals to the server with the shortest queue in the set. Then, we prove sufficient conditions on the sampling probabilities for throughput optimality.

4.2 Related work

Throughput and delay optimality of the power-of-d choices routing algorithm have been proved only when the servers are identical. If the service rates are different, there are known counterexamples for throughput optimality [68]. In other words, if the servers are different, power-of-d may reduce the stability region of the load balancing system. If the dispatcher knows the service rates, throughput and delay optimality of a modified version of power-of-d choices have been proved in [55, 69]. In this adaptation, the probability of sampling each server is proportional to its mean service rate. However, we are interested in studying the cases when service rates may be unknown to the dispatcher.

In [70] the authors address a similar question. They study stability of a general load balancing system, and they obtain sufficient conditions for throughput optimality. However,

they approach the problem from a different perspective, and they provide conditions that depend on the queue length processes. In this chapter, we provide conditions that only depend on the service rates and the sampling scheme. Hence, our conditions are easier to check.

Heavy-traffic analysis of the load balancing system operating under power-of-d choices has been done in the literature, but only under the assumption of identical and independent servers [63]. To the best of our knowledge, we are the first ones to obtain the heavy-traffic behavior of this queueing system with heterogeneous servers, and without modifying the probability of sampling each server.

4.3 Throughput optimality of power-of-d choices

The goal of this chapter is to provide necessary and sufficient conditions on the vector of service rates such that the load balancing system operating under the power-of-d choices routing algorithm (see Definition 4.1) is throughput and heavy-traffic optimal. Before presenting the result, we establish the details of the model.

We consider a load balancing system as described in section 3.4 but, instead of assuming that the arrivals and potential service per time slot are bounded, we only assume finite second moment. Let $\lambda \stackrel{\triangle}{=} \mathbb{E}\left[a(1)\right]$, $\boldsymbol{\mu} \stackrel{\triangle}{=} \mathbb{E}\left[\boldsymbol{s}(1)\right]$ and $\mu_\Sigma \stackrel{\triangle}{=} \sum_{i=1}^{n} \mu_i$. Without loss of generality, we assume the vector $\boldsymbol{\mu}$ is ordered from minimum to maximum, i.e., $\mu_i = \mu_{(i)}$ for all $i \in [n]$. Let $\sigma_a^2 \stackrel{\triangle}{=} \text{Var}\left[a(1)\right]$ be the variance of the arrival process and Σ_s the covariance matrix of $\boldsymbol{s}(1)$. For each $i \in [n]$, define $\sigma_{s_i}^2 \stackrel{\triangle}{=} (\Sigma_s)_{i,i}$. It is well known that the capacity region of the load balancing system is

$$\mathcal{C} \stackrel{\triangle}{=} \left\{x \in \mathbb{R}_+ : x \leq \mu_\Sigma\right\}, \tag{4.1}$$

i.e., for each $\lambda \in \text{Int}\left(\mathcal{C}\right)$, there exists a routing algorithm such that $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent, and if $\lambda \notin \mathcal{C}$, then $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is not positive recurrent for any

routing algorithm. A proof of this statement is presented in [34].

In this chapter we work with the power-of-d choices routing algorithm, also known as JSQ(d). We define it below.

Definition 4.1. *Fix $d \in [n]$. In each time slot, the power-of-d choices algorithm selects d queues uniformly at random, and then routes the arrivals to the shortest of these. Ties are broken at random. Formally, if queues $i_1, \ldots, i_d$ are selected uniformly at random, then the arrivals in time slot k are routed to the i^{*th} queue, where $i^* \in \arg\min_{i \in \{i_1,\ldots,i_d\}} \{q_i(k)\}$.*

Observe that the power-of-d choices algorithm does not require any information about arrival or service rates. It just requires observing the number of jobs at d of the queues in each time slot.

Before presenting the result we formally define throughput optimality.

Definition 4.2. *A routing algorithm $\mathcal{A}$ is throughput optimal if the queue length process $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ of the load balancing system operating under $\mathcal{A}$ is positive recurrent for all $\lambda \in$ Int $(\mathcal{C})$, where $\mathcal{C}$ is defined in Equation 4.1.*

Now we present the main theorem of this chapter. Recall that for a vector $\boldsymbol{x} \in \mathbb{R}^n$ we use $x_{(i)}$ for its i^{th} smallest element. Then, $x_{(1)} = \min_{i\in[n]} x_i$ and $x_{(n)} = \max_{i\in[n]} x_i$, for example.

Theorem 4.3. *For any $d \in [n-1]$, define*

$$\mathcal{M}^{(d)} \stackrel{\triangle}{=} \left\{ \boldsymbol{y} \in \mathbb{R}_+^n : \frac{\sum_{i=1}^{\ell} y_{(i)}}{y_\Sigma} \geq \frac{\binom{\ell}{d}}{\binom{n}{d}} \; \forall d \leq \ell \leq n-1 \right\}, \tag{4.2}$$

where $y_\Sigma \stackrel{\triangle}{=} \sum_{i=1}^n y_i$. Then, the power-of-$d$ choices algorithm is throughput optimal for the load balancing system described above if and only if $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$.

Before presenting the proof of Theorem 4.3 we present some remarks that will help interpreting the result.

Remark 4.4. *Observe that we can equivalently define $\mathcal{M}^{(d)}$ for all $d \in [n]$ as follows*

$$\mathcal{M}^{(d)} \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}_+^n : \frac{\sum_{i=1}^{\ell} y_{(i)}}{y_\Sigma} \geq \frac{\binom{\ell}{d}}{\binom{n}{d}} \; \forall \ell \in [n] \right\},$$

where we use the convention $\binom{\ell}{d} = 0$ if $\ell < d$. Here we only added redundant constraints to $\mathcal{M}^{(d)}$, so we use the definition in Equation 4.2 to avoid confusion.

Remark 4.5. *An interpretation of Theorem 4.3 is the following. In order for power-of-d choices algorithm to be throughput optimal, faster servers should be sampled sufficiently often. If this does not happen, it leads to the counter example in [68]. Equation 4.2 characterizes the amount of imbalance between service rates that power-of-d choices can tolerate. Note that, when the number of servers is fixed, as d increases, power-of-d choices can tolerate more imbalance because the right-hand side of Equation 4.2 becomes smaller. If $d = 1$, which corresponds to random routing, the set $\mathcal{M}^{(d)}$ is exactly the set of vectors where all the service rates are equal. In the other extreme case, when $d = n$, all the inequalities in Equation 4.2 are redundant, and $\mathcal{M}^{(d)}$ is the set of all nonnegative vectors. This fact is consistent with the throughput optimality of JSQ for any vector of service rates.*

Remark 4.6. *For $i \in [n]$, define*

$$\nu_i = \begin{cases} 0 & \text{, if } 1 \leq i \leq d-1 \\ \frac{\binom{i-1}{d-1}}{\binom{n}{d}} & \text{, if } d \leq i \leq n. \end{cases}$$

and let $\boldsymbol{\nu}$ be a vector with elements ν_i. An equivalent characterization of $\mathcal{M}^{(d)}$ is the set of all nonnegative vectors $\boldsymbol{\mu}$ such that $\frac{\boldsymbol{\mu}}{\mu_\Sigma}$ is majorized by $\boldsymbol{\nu}$. Majorization captures the notion of imbalance, and several equivalent characterizations can be found in [71]. This notion has been used in the study of balls and bins models in [72], and to prove optimality of routing and servicing algorithms in [73]. This notion also shows that for fixed d and n, the vector $\boldsymbol{\mu} = \boldsymbol{\nu}$ is on the boundary of $\mathcal{M}^{(d)}$.

Remark 4.7. *Theorem 4.3 establishes that if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, then the power-of-$d$ choices is not throughput optimal. In other words, if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$ there are some values of $\lambda \in$ Int $(\mathcal{C})$ for which $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is not positive recurrent. In fact, if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, the queue length process is positive recurrent only if $\lambda \in$ Int $\left(\overline{\mathcal{C}}\right)$, where*

$$\overline{\mathcal{C}} \stackrel{\triangle}{=} \left\{ x \in \mathbb{R}_+ : \; x \leq \frac{\binom{n}{d}}{\binom{\ell}{d}} \sum_{i=1}^{\ell} \mu_i \quad \forall d-1 \leq \ell \leq n-1 \right\}.$$

Observe that $\overline{\mathcal{C}} \subsetneq \mathcal{C}$ if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, and $\mathcal{C} = \overline{\mathcal{C}}$ if $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$. We omit the proof of this remark, since it easily follows from the proof of Theorem 4.3.

In the proof of Theorem 4.3 we use the Foster-Lyapunov theorem (Theorem 2.4) and a certificate that a DTMC is not positive recurrent (Theorem 2.5). Both of them are stated in section 2.2.

Proof of Theorem 4.3. Let $\epsilon \stackrel{\triangle}{=} \mu_\Sigma - \lambda$, and observe that $\lambda \in$ Int $(\mathcal{C})$ if and only if $\epsilon \in (0, \mu_\Sigma)$. We first prove that if $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$, then the power-of-$d$ choices algorithm is throughput optimal. To do that, we use the Foster-Lyapunov theorem (Theorem 2.4) with Lyapunov function $Z(\boldsymbol{q}) = \|\boldsymbol{q}\|^2$. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[\Delta Z(\boldsymbol{q}(k))\right] \\
=&\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k+1)\|^2 - \|\boldsymbol{q}(k)\|^2\right] \\
\stackrel{(a)}{=}&\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k+1) - \boldsymbol{u}(k)\|^2 + \|\boldsymbol{u}(k)\|^2 + 2\langle \boldsymbol{q}(k+1) - \boldsymbol{u}(k), \boldsymbol{u}(k)\rangle - \|\boldsymbol{q}(k)\|^2\right] \\
\stackrel{(b)}{=}&\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k) + \boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2 - \|\boldsymbol{u}(k)\|^2 - \|\boldsymbol{q}(k)\|^2\right] \\
\stackrel{(c)}{\leq}&\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k) + \boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2 - \|\boldsymbol{q}(k)\|^2\right] \\
\stackrel{(d)}{=}&\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2\right] + 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle\right], \qquad (4.3)
\end{aligned}$$

where (a) holds after adding and subtracting $\boldsymbol{u}(k)$ to the first term, and expanding the square; (b) holds after usingthe dynamics of the queues presented in Equation 1.2 and the

key property of the unused service presented in Equation 1.3, and reorganizing terms; (c) holds because $\|\boldsymbol{u}(k)\|^2 \geq 0$; and (d) holds after expanding the first square and reorganizing terms.

We analyze each of the terms in Equation 4.3 separately. For the first term, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\|^2\right] &\overset{(a)}{\leq} \mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k)\|^2\right] + \mathbb{E}\left[\|\boldsymbol{s}(k)\|^2\right] \\ &\overset{(b)}{=} \mathbb{E}\left[a(k)^2\right] + \sum_{i=1}^{n}\mathbb{E}\left[s_i(k)^2\right] \\ &\overset{(c)}{=} \lambda^2 + \sigma_a^2 + \sum_{i=1}^{n}\left(\mu_i^2+\sigma_{s_i}^2\right),\end{aligned}$$

where (a) holds after expanding the square, noticing that $\langle \boldsymbol{a}(k), \boldsymbol{s}(k)\rangle \geq 0$ and because the potential service vector is independent of the queue lengths; (b) holds because all the arrivals in one time slot are routed to the same queue; and (c) holds by definition of variance. Define $\zeta_1 \triangleq \lambda^2 + \sigma_a^2 + \sum_{i=1}^{n}\left(\mu_i^2+\sigma_{s_i}^2\right)$, and observe ζ_1 is a finite constant. Then,

$$\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\|^2\right] \leq \zeta_1. \tag{4.4}$$

Observe that the computation of the bound in Equation 4.4 does not use any properties of the routing algorithm. In other words, the Equation 4.4 is valid for the load balancing system under any routing algorithm.

To compute the second term of Equation 4.3, we first compute $\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle\right]$. Recall that under power-of-d choices, d queues are chosen uniformly at random, and then the arrivals are sent to the shortest among them. Then, we have

$$\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle\right] = \lambda \sum_{i=1}^{n-d+1} q_{(i)} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \tag{4.5}$$

because there are $\binom{n-i}{d-1}$ ways pf sampling d queues, and make sure that $q_{(i)}$ is the shortest; and there are $\binom{n}{d}$ ways of sampling d queues uniformly at random. If there are ties on

the queue lengths, power-of-d breaks them at random. Hence, the result in Equation 4.5 remains valid.

Let $\phi(i)$ be the index of the i^{th} shortest queue given $\boldsymbol{q}(k) = \boldsymbol{q}$. Then, since the potential service is independent of the queue lengths, the second term of Equation 4.3 is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle\right] &= \mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle\right] - \langle \boldsymbol{q}, \boldsymbol{\mu}\rangle \\ &= \sum_{i=1}^{n-d+1} q_{(i)} \left(\frac{\lambda\binom{n-i}{d-1}}{\binom{n}{d}} - \mu_{\phi(i)} \right) - \sum_{i=n-d+2}^{n} q_{(i)}\mu_{\phi(i)}. \end{aligned} \tag{4.6}$$

Define

$$\alpha_i \triangleq \begin{cases} \dfrac{\lambda\binom{n-i}{d-1}}{\binom{n}{d}} - \mu_{\phi(i)} & \text{, if } 1 \le i \le n-d+1 \\ -\mu_{\phi(i)} & \text{, if } n-d+1 < i \le n. \end{cases} \tag{4.7}$$

Claim 4.8. *The parameters α_i defined in Equation 4.7 satisfy*

1. $\alpha_n \le -\mu_1$.
2. $\displaystyle\sum_{i=1}^{n} \alpha_i = -\epsilon$.
3. *For any $\ell \in \mathbb{Z}_+$ satisfying $2 \le \ell \le n-1$, we have $\displaystyle\sum_{i=\ell}^{n} \alpha_i \le -\zeta_2$, where $\zeta_2 \triangleq \min\left\{\mu_1, \frac{\epsilon}{\binom{n}{d}}\right\}$.*

We prove Claim 4.8 at the end of this section. Now we compute an upper bound for Equation 4.6. We obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle\right] &= \sum_{i=1}^{n} \alpha_i q_{(i)} \\ &= q_{(1)} \sum_{i=1}^{n} \alpha_i + \sum_{\ell=2}^{n} \left(\sum_{i=\ell}^{n} \alpha_i \right) \left(q_{(\ell)} - q_{(\ell-1)} \right) \\ &\overset{(a)}{\le} -\epsilon q_{(1)} - \zeta_2 \sum_{\ell=2}^{n} \left(q_{(\ell)} - q_{(\ell-1)} \right)\end{aligned}$$

$$\stackrel{(b)}{=} q_{(1)}\left(\zeta_2 - \epsilon\right) - \zeta_2 q_{(n)}$$
$$\stackrel{(c)}{\leq} -\zeta_2 q_{(n)}, \tag{4.8}$$

where (a) holds by the first two properties of Claim 4.8; (b) holds after solving the telescopic sum and rearranging terms; and (c) holds because $\zeta_2 \leq \frac{\epsilon}{\binom{n}{d}}$ by definition, and $\binom{n}{d} \geq 1$. Using Equation 4.4 and Equation 4.8 in Equation 4.3 we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta Z(\boldsymbol{q}(k))\right] \leq \zeta_1 - 2\zeta_2 q_{(n)}.$$

Let $\eta > 0$. Then, defining

$$D \stackrel{\triangle}{=} \zeta_1 \quad \text{and} \quad \mathcal{B} \stackrel{\triangle}{=} \left\{\boldsymbol{q} \in \mathbb{R}_+^n : \max_{i\in[n]} q_i \leq \frac{\zeta_1 + \eta}{2\zeta_2}\right\},$$

both of the conditions of Theorem 2.4 are satisfied. Therefore, if $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$ then the power-of-d choices algorithm is throughput optimal.

Now we prove that if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, then the power-of-$d$ choices algorithm is not throughput optimal. In other words, we prove that if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, there exists $\lambda \in \text{Int}\left(\mathcal{C}\right)$ such that $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is not positive recurrent.

First observe that if $\boldsymbol{\mu} \notin \mathcal{M}^{(d)}$, there exists $\ell \in \mathbb{Z}_+$ such that $d \leq \ell \leq n-1$ and $\frac{\sum_{i=1}^{\ell}\mu_i}{\mu_\Sigma} < \frac{\binom{\ell}{d}}{\binom{n}{d}}$. Let ℓ^* be the smallest ℓ satisfying this condition, and $\delta_{\ell^*} > 0$ satisfy

$$\frac{\sum_{i=1}^{\ell^*}\mu_i}{\mu_\Sigma} + \delta_{\ell^*} = \frac{\binom{\ell^*}{d}}{\binom{n}{d}}. \tag{4.9}$$

We use Theorem 2.5 with function $V_{\ell^*}(\boldsymbol{q}) = \sum_{i=1}^{\ell^*} q_i$. We have

$$\mathbb{E}_{\boldsymbol{q}}\left[V_{\ell^*}(\boldsymbol{q}(k+1)) - V_{\ell^*}(\boldsymbol{q}(k))\right]$$
$$= \sum_{i=1}^{\ell^*} \mathbb{E}_{\boldsymbol{q}}\left[a_i(k) - s_i(k) + u_i(k)\right]$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \sum_{i=1}^{\ell^*} \mathbb{E}_{\boldsymbol{q}}\left[a_i(k)\right] - \sum_{i=1}^{\ell^*} \mu_i \\
&\overset{(b)}{\geq} \sum_{i=1}^{\ell^*} \mathbb{E}_{\boldsymbol{q}}\left[a_{\tilde{\phi}(i)}(k)\right] - \sum_{i=1}^{\ell^*} \mu_i \\
&\overset{(c)}{=} \sum_{i=d}^{\ell^*} \lambda \frac{\binom{i-1}{d-1}}{\binom{n}{d}} - \mu_\Sigma \left(\frac{\binom{\ell^*}{d}}{\binom{n}{d}} - \delta_{\ell^*} \right) \\
&\overset{(d)}{=} \mu_\Sigma \delta_{\ell^*} - \epsilon \frac{\binom{\ell^*}{d}}{\binom{n}{d}},
\end{aligned}
$$

where (a) holds because $\mathbb{E}\left[s_i(k)\right] = \mu_i$ and $\mathbb{E}\left[u_i(k)\right] \geq 0$ for all $i \in [n]$; (b) holds by letting $\tilde{\phi}(i)$ be the index of the i^{th} longest element of $\boldsymbol{q}$, and because under power-of-d choices the arrivals are routed to the shortest queue among the d selected; (c) holds by Equation 4.9, and because the arrivals are routed to the i^{th} longest queue only if the other $d-1$ selected queues are larger, and this happens with probability $\frac{\binom{i-1}{d-1}}{\binom{n}{d}}$ if $i \geq d$ and with probability 0 otherwise (similarly to the computation of Equation 4.5); and (d) holds because $\sum_{i=d}^{\ell^*} \binom{i-1}{d-1} = \binom{\ell^*}{d}$ and $\lambda = \mu_\Sigma - \epsilon$.

This proves conditions the first two conditions for $\epsilon > 0$ satisfying

$$
\epsilon \leq \mu_\Sigma \min\left\{ 1, \delta_{\ell^*} \frac{\binom{\ell^*}{d}}{\binom{n}{d}} \right\}
$$

To prove the third condition, observe

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[V_{\ell^*}(\boldsymbol{q}(k+1)) - V_{\ell^*}(\boldsymbol{q}(k))\right] \\
&= \sum_{i=1}^{\ell^*} \mathbb{E}_{\boldsymbol{q}}\left[a_i(k) - (s_i(k) - u_i(k))\right] \\
&\overset{(a)}{\leq} \sum_{i=1}^{\ell^*} \mathbb{E}_{\boldsymbol{q}}\left[a_i(k)\right] \\
&\overset{(b)}{\leq} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{q}}\left[a_i(k)\right] \\
&\overset{(c)}{=} \lambda
\end{aligned}
$$

where (a) holds because $u_i(k) \leq s_i(k)$ with probability 1, by definition of unused service; (b) holds because arrivals to each queue are a nonnegative random variable; and (c) holds because $a(k) = \sum_{i=1}^{n} a_i(k)$ and $\lambda = \mathbb{E}[a(k)]$. Since $\lambda < \infty$, this proves the third condition. This completes the proof of the theorem. □

We now present the proof of Claim 4.8.

Proof of Claim 4.8. We prove each of the three properties. We obtain:

1. If $i = n$ we have $\alpha_n = -\mu_{\phi(n)} \leq -\mu_1$, because $\mu_1 = \min_{i \in [n]} \mu_i$.

2. The total sum of α_i's satisfies

$$\sum_{i=1}^{n} \alpha_i = \frac{\lambda}{\binom{n}{d}} \sum_{i=1}^{n-d+1} \binom{n-i}{d-1} - \mu_\Sigma \overset{(a)}{=} \lambda - \mu_\Sigma = -\epsilon,$$

   where (a) holds because $\sum_{i=1}^{n-d+1} \binom{n-i}{d-1} = \binom{n}{d}$.

3. If $2 \leq \ell \leq n - d + 1$ we have that the tail sums are

$$\begin{aligned}
\sum_{i=\ell}^{n} \alpha_i =& \frac{\lambda}{\binom{n}{d}} \sum_{i=\ell}^{n-d+1} \binom{n-i}{d-1} - \sum_{i=\ell}^{n} \mu_{\phi(i)} \\
\overset{(a)}{=}& \lambda \frac{\binom{n+1-\ell}{d}}{\binom{n}{d}} - \sum_{i=\ell}^{n} \mu_{\phi(i)} \\
\overset{(b)}{=}& (\mu_\Sigma - \epsilon) \frac{\binom{n+1-\ell}{d}}{\binom{n}{d}} - \sum_{i=\ell}^{n} \mu_{\phi(i)} \\
\overset{(c)}{\leq}& \sum_{i=1}^{n+1-\ell} \mu_i - \frac{\binom{n+1-\ell}{d}}{\binom{n}{d}} \epsilon - \sum_{i=\ell}^{n} \mu_{\phi(i)} \\
\overset{(d)}{\leq}& -\frac{\epsilon}{\binom{n}{d}},
\end{aligned}$$

   where (a) holds because $\sum_{i=\ell}^{n-d+1} \binom{n-i}{d-1} = \binom{n+1-\ell}{d}$; (b) holds by definition of ϵ; (c) holds because $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$; and (d) holds because $\binom{n+1-\ell}{d} \geq 1$, and because $\sum_{i=1}^{n+1-\ell} \mu_i - \sum_{i=\ell}^{n} \mu_{\phi(i)} \leq 0$, since $\sum_{i=1}^{n+1-\ell} \mu_i$ is the sum of the $n + \ell - 1$ smallest

elements of $\boldsymbol{\mu}$, and $\sum_{i=\ell}^{n} \mu_{\phi(i)}$ is the sum of $n+\ell-1$ of the elements of $\boldsymbol{\mu}$ which are not necessarily the smallest.

If $n-d+1 < \ell \leq n-1$ we have

$$\sum_{i=\ell}^{n} \alpha_i = -\sum_{i=\ell}^{n} \mu_{\phi(i)} \leq -\mu_1.$$

where the inequality holds because $\mu_1 = \min_{i\in[n]} \mu_i$. Then, for all $2 \leq \ell \leq n-1$ we have

$$\sum_{i=\ell}^{n} \alpha_i \leq -\zeta_2 \stackrel{\triangle}{=} -\min\left\{ \frac{\epsilon}{\binom{n}{d}},\, \mu_1 \right\}.$$

□

4.4 Heavy-traffic optimality

In this section we perform heavy-traffic analysis of a heterogeneous load balancing system operating under power-of-d choices. Specifically, we prove that in the heavy-traffic limit, the load balancing system operating under power-of-d choices behaves as a single server queue and show that the scaled vector of queue lengths converges to a vector of exponential random variables. This result is similar to Corollary 3.7 and Corollary 3.8. The main difference is that in Corollary 3.7 we used the JSQ routing policy, and in Corollary 3.8 we studied power-of-2 choices under homogeneous servers. Here we show that the same result can be proved for power-of-d choices with heterogeneous servers, under similar conditions to Theorem 4.3.

We parametrize the system similarly to section 3.5. Specifically, we fix a sequence of service rate vectors $\{\boldsymbol{s}(k) : k \in \mathbb{Z}_+\}$ and take $\epsilon \in (0, \mu_\Sigma)$. The arrival process to the system parametrized by ϵ is an i.i.d. sequence $\{a^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ that satisfies $\lambda^{(\epsilon)} \stackrel{\triangle}{=} \mathbb{E}\left[a^{(\epsilon)}(1)\right] = \mu_\Sigma - \epsilon$. Then, the heavy-traffic limit is obtained by taking $\epsilon \downarrow 0$. We add a

superscript (ϵ) to the queue length, arrival and unused service variables when we refer to the load balancing system parametrized by ϵ.

In the next proposition we show SSC to the cone $\mathcal{K} \stackrel{\triangle}{=} \left\{\boldsymbol{x} \in \mathbb{R}^n_+ : x_i = x_\ell \; \forall i, \ell \in [n]\right\}$. For any vector $\boldsymbol{x} \in \mathbb{R}^n$, define

$$\boldsymbol{x}_\parallel = \mathbf{1}\left(\frac{\sum_{i=1}^n x_i}{n}\right) \; , \; \boldsymbol{x}_\perp \stackrel{\triangle}{=} \boldsymbol{x} - \boldsymbol{x}_\parallel. \tag{4.10}$$

Then, $\boldsymbol{x}_\parallel$ is the projection of $\boldsymbol{x}$ on $\mathcal{K}$ and $\boldsymbol{x}_\perp$ is the error of approximating $\boldsymbol{x}$ by $\boldsymbol{x}_\parallel$. Now we present the result.

Proposition 4.9. *Given a sequence $\{\boldsymbol{s}(k) : k \in \mathbb{Z}_+\}$ of i.i.d. random vectors, and $\epsilon \in (0, \mu_\Sigma)$, consider a load balancing system operating under power-of-d choices, parametrized by ϵ as described above. Suppose $d \geq 2$, and that the number of arrivals and the potential service in each time slot are bounded. Let $\boldsymbol{\mu} \in$ Int $\left(\mathcal{M}^{(d)}\right)$ and let $\overline{\boldsymbol{q}}^{(\epsilon)}$ be a steady-state vector such that $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ converges in distribution to $\overline{\boldsymbol{q}}^{(\epsilon)}$ as $k \uparrow \infty$. Let $\delta > 0$ be such that for all $\ell \in \mathbb{Z}_+$ satisfying $d \leq \ell \leq n-1$ we have*

$$\frac{\sum_{i=1}^\ell \mu_i}{\mu_\Sigma} - \delta \geq \frac{\binom{\ell}{d}}{\binom{n}{d}}. \tag{4.11}$$

If $\epsilon < \delta\mu_\Sigma$, then $\mathbb{E}\left[\|\overline{\boldsymbol{q}}^{(\epsilon)}_\perp\|^j\right] \leq J_j$ for each $j \in \mathbb{Z}_+$ with $j \geq 1$, where J_j is a finite constant (independent of ϵ).

Proposition 4.9 says that the error of approximating $\overline{\boldsymbol{q}}^{(\epsilon)}$ by $\overline{\boldsymbol{q}}^{(\epsilon)}_\parallel$ is negligible in heavy traffic because, as ϵ gets smaller, the arrival rate to the system increases and, therefore, the vector of queue lengths $\overline{\boldsymbol{q}}^{(\epsilon)}$ becomes larger. Then, the projection $\overline{\boldsymbol{q}}^{(\epsilon)}_\parallel$ also becomes larger. However, the error of approximating $\overline{\boldsymbol{q}}^{(\epsilon)}$ by $\overline{\boldsymbol{q}}^{(\epsilon)}_\parallel$, denoted as $\overline{\boldsymbol{q}}^{(\epsilon)}_\perp$, has bounded moments. Then, as ϵ goes to zero it becomes negligible.

Observe that the vector $\overline{\boldsymbol{q}}^{(\epsilon)}$ is well defined, because $\boldsymbol{\mu} \in$ Int $\left(\mathcal{M}^{(d)}\right) \subset \mathcal{M}^{(d)}$. Then, from Theorem 4.3 we know that the DTMC $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent for all

$\epsilon \in (0, \mu_\Sigma)$.

Proof of 4.9. For ease of exposition, we omit the dependence on ϵ of the variables. Define

$$V(\boldsymbol{q}) \triangleq \|\boldsymbol{q}\|^2, \quad V_\parallel(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_\parallel\|^2, \quad W_\perp(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_\perp\|.$$

We use Lemma 2.7 with $Z(\boldsymbol{q}) = W_\perp(\boldsymbol{q})$. We start with a fact first used in [34]. Observe that $\|\boldsymbol{q}_\perp\| = \sqrt{\|\boldsymbol{q}_\perp\|^2}$ by definition of square root, and $f(x) = \sqrt{x}$ is a concave function. Then, by definition of concavity and the Pythagoras theorem,

$$\Delta W_\perp(\boldsymbol{q}) \leq \frac{1}{2\|\boldsymbol{q}_\perp\|} \left(\Delta V(\boldsymbol{q}) - \Delta V_\parallel(\boldsymbol{q})\right). \tag{4.12}$$

Then, to prove condition (C1), it suffices to upper bound $\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q})]$ and lower bound $\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_\parallel(\boldsymbol{q})\right]$. We start with $\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q})]$. From the proof of Theorem 4.3, we know Equation 4.4 is satisfied, i.e.,

$$\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q}(k))] \leq \zeta_1 + 2\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle].$$

We analyze the last term differently here. Defining $\phi(i)$ as in the proof of Theorem 4.3, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle] \\
=&\lambda \sum_{i=1}^{n-d+1} q_{(i)} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} - \sum_{i=1}^{n} q_{(i)} \mu_{\phi(i)} \\
\overset{(a)}{=}& -\epsilon \left(\frac{\sum_{i=1}^n q_i}{n}\right) + \sum_{i=1}^{n-d+1} q_{(i)} \frac{\lambda\binom{n-i}{d-1}}{\binom{n}{d}} + \sum_{i=1}^{n} q_{(i)} \left(\frac{\epsilon}{n} - \mu_{\phi(i)}\right) \\
\overset{(b)}{=}& -\epsilon \left(\frac{\sum_{i=1}^n q_i}{n}\right) + \sum_{i=1}^{n} q_{(i)} \beta_i
\end{aligned}$$

where (a) holds by adding and subtracting $\frac{\epsilon}{n}\left(\sum_{i=1}^n q_i\right)$, and reorganizing terms; and (b)

holds defining for each $i \in [n]$

$$\beta_i \triangleq \begin{cases} \frac{\binom{n-i}{d-1}}{\binom{n}{d}}\lambda + \frac{\epsilon}{n} - \mu_{\phi(i)} & \text{, if } 1 \leq i \leq n-d+1 \\ \frac{\epsilon}{n} - \mu_{\phi(i)} & \text{, if } n-d+1 < i \leq n \end{cases} \tag{4.13}$$

Observe $\beta_i = \alpha_i + \frac{\epsilon}{n}$ for each $i \in [n]$, where α_i is defined in Equation 4.7.

Claim 4.10. *The parameters β_i defined in Equation 4.13 satisfy*

1. $\beta_n \leq -\mu_{(1)} + \frac{\epsilon}{n}$.
2. $\sum_{i=1}^{n} \beta_i = 0$.
3. *For any $\ell \in \mathbb{Z}_+$ satisfying $2 \leq \ell \leq n-1$ we have $\sum_{i=\ell}^{n} \beta_i \leq -\delta\mu_\Sigma + \epsilon$.*

We prove Claim 4.10 in subsection 4.6.1. Observe that if $d = 1$, the second property is not satisfied. Using Claim 4.10 we obtain

$$\begin{aligned} \sum_{i=1}^{n} q_{(i)}\beta_i =& q_{(1)} \sum_{i=1}^{n} \beta_i + \sum_{\ell=2}^{n} \left(\sum_{i=\ell}^{n} \beta_i \right) \left(q_{(\ell)} - q_{(\ell-1)} \right) \\ \leq& \left(-\delta\mu_\Sigma + \epsilon \right) \left(q_{(n)} - q_{(1)} \right). \end{aligned} \tag{4.14}$$

Observe that, by definition of $\boldsymbol{q}_\perp$, we have

$$\|\boldsymbol{q}_\perp\|^2 = \sum_{i=1}^{n} \left(q_i - \frac{\sum_{j=1}^{n} q_j}{n} \right) \overset{(a)}{\leq} n \left(q_{(n)} - q_{(1)} \right),$$

where (a) holds because $q_i \leq q_{(n)}$ for all $i \in [n]$ and $\frac{1}{n}\sum_{\ell=1}^{n} q_\ell \geq q_{(1)}$ by definition of $q_{(1)}$ and $q_{(n)}$. Using this result in Equation 4.14 we obtain that

$$\sum_{i=1}^{n} q_{(i)}\beta_i \leq \left(\frac{-\delta\mu_\Sigma + \epsilon}{\sqrt{n}} \right) \|\boldsymbol{q}_\perp\| \leq \left(\frac{-\delta\mu_\Sigma + \epsilon_0}{\sqrt{n}} \right) \|\boldsymbol{q}_\perp\|,$$

for any $\epsilon_0 \in (0, \delta\mu_\Sigma)$. Therefore,

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V(\boldsymbol{q}(k))\right] \leq \zeta_1 - 2\epsilon\left(\frac{\sum_{i=1}^n q_i}{n}\right) + 2\left(\frac{-\delta\mu_\Sigma + \epsilon_0}{\sqrt{n}}\right)\|\boldsymbol{q}_\perp\|. \tag{4.15}$$

To lower bound $\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_\parallel(\boldsymbol{q})\right]$ we only use properties of the norm and the unused service. We obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_\parallel(\boldsymbol{q}(k))\right] \geq -2\epsilon\left(\frac{\sum_{i=1}^n q_i}{n}\right) - \zeta_3, \tag{4.16}$$

where $\zeta_3 \stackrel{\triangle}{=} 2nS_{\max}^2$, and $S_{\max}$ is a finite constant such that $s_i(1) \leq S_{\max}$ for all $i \in [n]$ with probability 1. Using Equation 4.15 and Equation 4.16 in Equation 4.12 we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta W_\perp(\boldsymbol{q}(k))\right] \leq \frac{\zeta_1 + \zeta_3}{2\|\boldsymbol{q}_\perp\|} + \left(\frac{-\delta\mu_\Sigma + \epsilon_0}{\sqrt{n}}\right),$$

which satisfies condition (C1) for $\eta > 0$ and

$$\kappa = \left(\frac{\zeta_1 + \zeta_3}{2}\right)\left(-\eta + \frac{\delta\mu_\Sigma - \epsilon_0}{\sqrt{n}}\right)^{-1}.$$

Condition (C2) is trivially satisfied because potential service and arrivals in one time slot are bounded random variables. □

Using SSC, we can completely determine the behavior of the vector of queue lengths in heavy traffic. In the next proposition we provide this result.

Theorem 4.11. *Consider a set of load balancing systems operating under power-of-d as described in Proposition 4.9. Let $\sigma_a^{(\epsilon)}$ be the standard deviation of $a^{(\epsilon)}(1)$ and assume $\sigma_a = \lim_{\epsilon\downarrow 0}\sigma_a^{(\epsilon)}$. Then, $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \Longrightarrow \Upsilon\mathbf{1}$ as $\epsilon \downarrow 0$, where Υ is an exponential random variable with mean $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right)$.*

Remark 4.12. *In Proposition 4.9 and Theorem 4.11 we assume that the set $\mathcal{M}^{(d)}$ has nonempty interior. This can be proved by observing that, for $d \geq 2$, a vector of homoge-*

neous service rates $\boldsymbol{\mu} = \xi\mathbf{1}$ (with $\xi > 0$) satisfies all the inequalities in Equation 4.2, and none of them is tight. Then, such $\boldsymbol{\mu} = \xi\mathbf{1} \in$ Int $\left(\mathcal{M}^{(d)}\right)$. On the other hand, when $d = 1$, the set $\mathcal{M}^{(d)}$ only contains the homogeneous service rate vectors, which has an empty interior. Then, our heavy-traffic results are not applicable. This is consistent with the fact that random routing is not heavy-traffic optimal.

Proof of Theorem 4.11. We use the MGF method introduced in section 3.3. In fact, our theorem is a corollary of Theorem 3.5. We only verify that three conditions are satisfied.

We first verify that the routing algorithm is throughput optimal, which holds from Theorem 4.3 because we assume $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$. The second condition is SSC to a one-dimensional subspace, which is satisfied by Proposition 4.9. In fact, in Theorem 3.5 we require a weaker notion of SSC, which is trivially satisfied after proving Proposition 4.9. The last condition is existence of the MGF of $\epsilon \sum_{i=1}^{n} \overline{q}_i$, which we formalize in Claim 4.13. We omit the proof, since it is equivalent to the proof of Lemma 3.13.

Claim 4.13. *For the load balancing system described in Theorem 4.11, there exists $\Theta > 0$ such that $\mathbb{E}\left[e^{\theta\epsilon\sum_{i=1}^{n}\overline{q}_i^{(\epsilon)}}\right]$ is finite for all $\theta \in [-\Theta, \Theta]$.*

□

4.5 Generalization to other routing policies

In this section we generalize the sufficient conditions in Theorem 4.3 to a larger class of routing policies. Instead of using power-of-d choices, suppose the router randomly selects an arbitrary subset of servers, and then the arrivals are routed to the server with the shortest queue among these. Let $\pi : 2^{[n]} \to [0, 1]$ be the probability mass function that governs the set of servers that are randomly selected in each time slot. We call $\mathcal{R}^{\pi}$ the routing algorithm described above.

Theorem 4.14. *Given $\pi : 2^{[n]} \to [0, 1]$, consider a load balancing system as described in section 3.4, operating under $\mathcal{R}^{\pi}$. For each subset $\mathcal{S} \subseteq [n]$, let $\pi(\mathcal{S})$ be the probability of*

sampling the servers in the set $\mathcal{S}$. Let $\mathcal{P}([n])$ be the set of permutations of the elements of the set $[n]$, and for each $\tau \in \mathcal{P}([n])$ define

$$\mathcal{M}_\tau \stackrel{\triangle}{=} \left\{ \boldsymbol{y} \in \mathbb{R}_+^n : \frac{\sum_{i=1}^{\ell} y_{(i)}}{y_\Sigma} \leq \sum_{i=1}^{\ell} \sum_{\mathcal{S} \in \mathcal{S}_i^\tau} \pi(\mathcal{S}) \ \forall \ell \in [n-1] \right\},$$

where

$$\mathcal{S}_i^\tau \stackrel{\triangle}{=} \{\mathcal{S} \subseteq [n] : \tau(n-i+1) \in \mathcal{S}, \tau(\ell) \notin \mathcal{S} \quad \forall \ell < n-i+1\}.$$

Then, the routing algorithm $\mathcal{R}^\pi$ is throughput optimal if $\boldsymbol{\mu} \in \mathcal{M}_\tau$ for all $\tau \in \mathcal{P}([n])$.

The proof is similar to the proof of Theorem 4.3, and we present a sketch below.

Proof sketch of Theorem 4.14. The proof is very similar to Theorem 4.3. In fact, the only difference is the computation of $\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k) \rangle]$. Since the sampling scheme in power-of-d choices is symmetric, in Theorem 4.3 we obtain the simple expression presented in Equation 4.5. In this case, we obtain

$$\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k) \rangle] = \sum_{i=1}^{n} q_{(i)} \lambda \left(\sum_{\substack{\mathcal{S} \subseteq [n]: \\ \phi(i) \in \arg\min_{\ell \in \mathcal{S}} q_\ell}} \pi(\mathcal{S}) \right).$$

We omit the rest of the proof for brevity. □

4.6 Details of the proofs in section 4.4

4.6.1 Proof of Claim 4.10

Proof of Claim 4.10. We prove each of the three properties. We have:

1. If $i = n$ we have

$$\beta_n = \alpha_n + \frac{\epsilon}{n} \leq -\mu_{(1)} + \frac{\epsilon}{n},$$

where we used property item 1 from Claim 4.8.

2. The total sum of β_i's satisfies

$$\sum_{i=1}^{n} \beta_i = \sum_{i=1}^{n} \alpha_i + \epsilon = 0,$$

where we used property item 2 from Claim 4.8.

3. To prove this property we divide in 2 cases. If $\ell \leq n - d + 1$ we have

$$\begin{aligned}
\sum_{i=\ell}^{n} \beta_i &= \lambda \sum_{i=\ell}^{n-d+1} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} + \sum_{i=\ell}^{n} \left(\frac{\epsilon}{n} - \mu_{\phi(i)} \right) \\
&= (\mu_\Sigma - \epsilon) \frac{\binom{n+1-\ell}{d}}{\binom{n}{d}} + \left(\frac{n-\ell+1}{n} \right) \epsilon - \sum_{i=\ell}^{n} \mu_{\phi(i)} \\
&\overset{(a)}{\leq} \frac{\binom{n+1-\ell}{d}}{\binom{n}{d}} \mu_\Sigma + \epsilon - \sum_{i=1}^{n-\ell+1} \mu_{(i)} \\
&\overset{(b)}{\leq} \epsilon - \delta \mu_\Sigma
\end{aligned}$$

where (a) holds because $\epsilon > 0$, $\frac{n-\ell+1}{n} \leq 1$ and because $\sum_{i=1}^{n-\ell+1} \mu_i$ is the sum of the smallest $(n - \ell + 1)$ elements of $\boldsymbol{\mu}$; and (b) holds by Equation 4.11 and reorganizing terms.

If $\ell > n - d + 1$ we have

$$\begin{aligned}
\sum_{i=\ell}^{n} \beta_i &= \sum_{i=\ell}^{n} \left(\frac{\epsilon}{n} - \mu_{\phi(i)} \right) \\
&\overset{(a)}{\leq} \frac{n-\ell+1}{n} \epsilon - \sum_{i=1}^{n-\ell+1} \mu_i \\
&\overset{(b)}{\leq} \epsilon - \mu_\Sigma \left(\frac{\binom{n-\ell+1}{d}}{\binom{n}{d}} + \delta \right) \\
&\overset{(c)}{\leq} \epsilon - \delta \mu_\Sigma
\end{aligned}$$

where (a) holds because $\sum_{i=1}^{n-\ell+1} \mu_i$ is the sum of the smallest $(n-\ell+1)$ elements of $\boldsymbol{\mu}$; (b) holds because $\frac{n-\ell+1}{n} \leq 1$ and by Equation 4.11; and (c) because $\frac{\binom{n-\ell+1}{d}}{\binom{n}{d}} \geq 0$.

□

4.7 Conclusion and future work

In this chapter we study performance of power-of-d choices in inhomogeneous load balancing systems. We find necessary and sufficient conditions for throughput optimality and we show that almost under the same conditions, we have heavy-traffic optimality. The conditions we obtain formalize the intuition that power-of-d choices is a good routing algorithm when the service rates are not too imbalanced. However, they do not need to be equal.

Future work is to explore routing policies as described in section 4.5 and obtain heavy-traffic results for them.

CHAPTER 5

LOAD BALANCING UNDER MANY-SERVER HEAVY-TRAFFIC REGIME

Based on:

D. Hurtado-Lange and S. T. Maguluri, "Load balancing system under join the shortest queue: Many-server-heavy-traffic asymptotics," *arXiv preprint arXiv:2004.04826v2*, 2020

5.1 Introduction

So far, we have been studying the heavy-traffic asymptotics of the load balancing system. In this regime, we keep the number of jobs constant and we increase the load to the maximum capacity. In this chapter we work in the many-server heavy-traffic regime, where both, the load and the number of servers, increase together. Specifically, we let n be the number of servers and we parametrize the arrival process so that the mean arrival rate *per server* is $1 - n^{-\alpha}$, where $\alpha > 0$. Then, the *total* arrival rate to the system is $n(1 - n^{-\alpha})$. In the many-server heavy-traffic regime, there are different phases depending on the value of α. As $\alpha \downarrow 0$, we approximately approach the mean-field regime, $\alpha = \frac{1}{2}$ represents the Halfin-Whitt regime [75], $\alpha = 1$ represents the nondegenerate-slowdown regime (NDS) [76], and $\alpha \to \infty$ can be thought of as the classical heavy-traffic regime. In this chapter, we look at all super-NDS regimes, i.e., the regimes with $\alpha > 1$ and, hence, that are more heavily loaded than NDS. The main contributions of this chapter are summarized below:

(i) We show that the total queue length scaled by $n^{-\alpha}$ (or, equivalently, the average queue length scaled by $n^{1-\alpha}$) converges in distribution to an exponential random variable if the load grows 'fast enough' with respect to the number of servers. In particular, under power-of-d choices with constant d, we show that this result is valid if $\alpha > 3$ (see Corollary 5.11); and under JSQ the same result holds for $\alpha > 2$ (see

Corollary 5.12). Further, we show the condition that α must satisfy under power-of-d choices when d is a function of the number of servers. Specifically, we show that if $d \triangleq cn^{\beta}$ for some $c > 0$ and $\beta \geq 0$ such that $d \in [n]$, the convergence to the exponential random variable is valid if $\alpha + \beta > 3$ (see Theorem 5.10). We provide two proofs to our result, which we explain in the next two contributions.

(ii) We first show the result using one-sided Laplace transform (see section 5.3 and section 5.7). Specifically, we generalize the transform method introduced in chapter 3 for discrete-time systems in heavy traffic. In subsection 5.3.2 we generalize it to the many-server heavy-traffic asymptotics, and in section 5.9 we further generalize it to continuous-time model.

(iii) We compute the rate of convergence of the scaled total queue length to the exponential random variable in Wasserstein's distance (see Theorem 5.6 and Theorem 5.19). These are stronger versions of Theorem 5.10 and Theorem 5.1, respectively, where we actually obtain the convergence in distribution as a consequence of the error bound. To show this result we use Stein's method (see section 5.10).

(iv) All these proofs are powered by multiplicative SSC results that we show in Proposition 5.3 and Proposition 5.14. We show SSC to the line generated by the vector $\mathbf{1}$, i.e., we show that all the queue lengths are similar. Specifically, we compute bounds for the moments of the norm of the difference between the queue length vector and its projection on the line generated by the vector $\mathbf{1}$. Further, we compute a bound for the MGF of its norm. These bounds grow to infinity as the number of servers increase. However, after scaling the total queue length by $n^{-\alpha}$ they become negligible, hence the name multiplicative.

(v) We compute the rate of convergence in expected value of the total average queue length scaled by $n^{-\alpha}$. Specifically, we show that the rate of convergence is of order $\log(n)n^{3-\beta}$. As a consequence, we prove the convergence of the expectation under

the same conditions established in Theorem 5.10 (see Theorem 5.21). Similarly to Theorem 5.10, we explicitly show that the power-of-d choices algorithm with constant d and the JSQ algorithm are immediate consequences of Theorem 5.21. To prove this result we use the drift method.

5.2 Related work

In previous chapters, we have focused on heavy-traffic analysis, where the number of servers is kept constant and the load is increased to maximum capacity. Another popular regime is mean-field , where we keep the load constant and we increase the number of servers.

The mean-field regime has become popular after it was used to show that the power-of-2 choices algorithm yields queue lengths that are considerably smaller than random routing [18, 19, 20]. It was later proved that the JSQ system behaves as an $M/M/\infty$ system in the mean-field regime [77]. In [78], it was shown that under power-of-d choices with d growing with n, the fluid limit does not depend on the growth rate and, hence, power-of-d and JSQ have the same fluid limit. More recently, it has been shown that, in this regime, there must always be a proportion of empty queues and, hence, any routing policy that prioritizes empty queues yields queue lengths of at most one job [79]. Under the same logic, the join the idle queue (JIQ) policy has become popular. It was proposed in [59] and the idea is that, whenever a server idles, it communicates its status to the dispatcher. Then, the arrivals are routed randomly to one of the empty queues. If none of the queues is empty, then a server is selected uniformly at random. This policy has been rigorously analyzed in [60] under exponential job sizes, and in [80] for general job-size distributions. In both cases, the authors show that the steady-state probability that an arriving job waits in line vanishes as the number of servers grows to infinity.

Among the many-server heavy-traffic regimes (where the number of servers and the load increase together), one of the most popular is the Halfin-Whit regime, where the dif-

ference between the service and arrival rate *per server* is $n^{-1/2}$, i.e., $\alpha = \frac{1}{2}$. This regime was introduced in [75], where the authors present the classical analysis of the $M/M/n$ queue. More recently, [58] shows that the number of empty queues and the number of queues with one customer in line are of order $O(\sqrt{n})$. The authors use the diffusion limits approach, but interchange of limits is not proved. This step is completed in [26]. In [81, 82] the work of [58] is continued. Specifically, in [81] the authors study tail asymptotics of the stationary distribution, and in [82] they study the moments of the stationary distribution. In [83], the authors show that JIQ routing yields diffusion-level optimality in the Halfin-Whitt regime.

In [29], load balancing systems under several routing policies in the sub-Halfin-Whitt regime are studied, and in [84] the analysis is extended to the super-Halfin-Whitt regime, i.e., when $\alpha \in \left[\frac{1}{2}, 1\right)$. In [85], the authors also focus on the super-Halfin-Whitt regime, and they compute the asymptotic distribution of the (centered and scaled) queue lengths. In [86] a load balancing system operating under power-of-d, where jobs are batches of tasks, is analyzed. Specifically, the authors find conditions on the value of d (as a function of the number of servers, the load and the number of tasks per job) such that power-of-d choices achieves zero delay in sub-Halfin-Whitt regime.

The NDS regime was introduced in [76] in the context of an $M/M/n$ queue, and the author shows that the regime yields new diffusion processes. More recently, it has been used to compare routing policies in the load balancing system [79]. Specifically, the authors in [79] characterize the diffusion approximation of JSQ and propose a new policy with less communication overhead, and that achieves JSQ optimality. This policy is called idle-one-first, and prioritizes routing to servers that are idling or have one job.

Multiplicative SSC has been used in a variety of contexts in the literature [87, 88, 89, 90, 64, 22, 9]. The most relevant work in our context are the results in [88, 22]. In [88] the authors study a parallel-server system in the Halfin-Whitt regime, and they propose a framework for establishing SSC in queueing systems with multiple server pools in parallel

Table 5.1: Literature review for asymptotic regimes depending on the value of α.

Value of α	Regime	References
$\alpha \downarrow 0$ (intuitively)	Mean-field	[77, 80, 19, 20, 78, 60, 18]
$\alpha \in \left(0, \frac{1}{2}\right)$	Sub-Halfin-Whitt	[29, 86]
$\alpha = \frac{1}{2}$	Halfin-Whitt	[81, 82, 26, 58, 75, 83]
$\alpha \in \left(\frac{1}{2}, 1\right)$	Super-Halfin-Whitt	[84, 85]
$\alpha = 1$	Nondegenerate Slow-down (NDS)	[76, 79]
$\alpha \in (1, 2]$	Super-NDS	Open question
$\alpha \in (2, \infty)$	Super-NDS	This chapter
$\alpha \to \infty$ (intuitively)	Classical heavy-traffic	[6, 7, 9, 3, 2, 8, 5, 4, 44, 16, 34, 63, 67, 91, 24]

and different customer classes. They use the fluid dynamics to establish their result. In [22] the multiplicative SSC result is used in the context of the heavy-traffic analysis of a bandwidth sharing network, and they use the drift method to analyze it. Their proof is based on bounding the drift of the error of approximating the actual vector of flows by its projection on the subspace where SSC occurs, which is traditional in the drift method [34, 15, 14]. In the traditional drift method technique, the SSC bounds are independent of the heavy-traffic parameter. However, in [22], the bounds depend on the heavy-traffic parameter and the authors show that they become negligible after scaling. In this chapter, we adopt their technique to show SSC and we use it in the context of a load balancing system in the many-server heavy-traffic regime.

In Table 5.1 we show a summary of the related work presented above, classified according to the value of α. We also include the literature on heavy-traffic analysis introduced in previous chapters.

5.3 Load balancing under JSQ

Consider the load balancing model introduced in chapter 3. We start introducing the details of the many-server heavy-traffic parametrization. For each $i \in [n]$, assume $\mathbb{E}[s_i(1)] = 1$

and $\text{Var}\left[s_i(1)\right] = \sigma_s^2$. We are interested in the many-server heavy-traffic limit, so we parametrize the system by the number of servers n in the following way. We add a superscript (n) to the variables when we refer to the parametrized system. Let $\lambda^{(n)} \triangleq \mathbb{E}\left[a^{(n)}(1)\right] = n\left(1 - n^{-\alpha}\right)$, where $\alpha > 0$ and $\text{Var}\left[a^{(n)}(1)\right] = n\sigma_a^2$. Observe that the mean and variance of the arrival rate increases linearly with n. Hence, the upper bound $A_{\max}$ also needs to be a function of n. Let $\tilde{A}_{\max} > 0$ be a finite constant such that $a^{(n)}(1) \leq n\tilde{A}_{\max}$ with probability 1 for each n.

Note that

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

which is positive. Therefore, the Markov Chain $\left\{\boldsymbol{q}^{(n)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent [34, Lemma 2]. Assume $\mathbb{P}\left[a(k) - s_i(k) = 0\right] > 0$ for some $i \in [n]$. Then, $\left\{\boldsymbol{q}^{(n)}(k) : k \in \mathbb{Z}_+\right\}$ is also aperiodic because $\mathbb{P}\left[\boldsymbol{q}(k) = \boldsymbol{q}(k+1)\right] > 0$. Then, the vector of queue lengths converges in distribution to a steady-state random vector, that we denote $\overline{\boldsymbol{q}}^{(n)}$. Let $\overline{a}^{(n)}$ be a steady-state random variable with the same distribution as $a^{(n)}(1)$ and $\overline{\boldsymbol{s}}$ be a steady-state random vector with the same distribution as $\boldsymbol{s}(1)$. Let $\overline{\boldsymbol{a}}^{(n)}$ be the vector of arrivals after routing in steady state, given that the queue lengths are $\overline{\boldsymbol{q}}^{(n)}$ and $\overline{a}^{(n)}$ jobs arrive to the system, and let $\overline{\boldsymbol{u}}^{(n)}$ be the vector of unused service in steady-state, given $\overline{\boldsymbol{q}}^{(n)}$, $\overline{a}^{(n)}$ and $\overline{\boldsymbol{s}}^{(n)}$. Define $\left(\overline{\boldsymbol{q}}^{(n)}\right)^+ \triangleq \overline{\boldsymbol{q}}^{(n)} + \overline{\boldsymbol{a}}^{(n)} - \overline{\boldsymbol{s}} + \overline{\boldsymbol{u}}^{(n)}$ as the vector of queue lengths one time slot after $\overline{\boldsymbol{q}}^{(n)}$ is observed, given $\overline{a}^{(n)}$ and $\overline{\boldsymbol{s}}$. Then, we have $\left(\overline{q}_i^{(n)}\right)^+ \overline{u}_i^{(n)} = 0$ with probability 1 for all $i \in [n]$.

In the next subsections we prove Theorem 5.1, using two different approaches.

Theorem 5.1. *Consider a sequence of load balancing systems operating under JSQ, parametrized by n as described above. If $\alpha > 4$, then $n^{-\alpha}\sum_{i=1}^{n} \overline{q}_i^{(n)} \Rightarrow \Upsilon$ as $n \to \infty$, where Υ is an exponential random variable with mean $\frac{\sigma_a^2+\sigma_s^2}{2}$.*

Similarly to the classical heavy-traffic regime, proving SSC and bounding the unused

service are essential in the proof of Theorem 5.1. Further computing the expected total unused service in steady state is essential in the proof of SSC and in the proof of Theorem 5.1. We state the result below, as it will be repeatedly in the rest of this section.

Lemma 5.2. *Consider a load balancing system operating under JSQ, parametrized by n as described above. Then,*

$$\mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i^{(n)}\right] = n^{1-\alpha}.$$

Proof of Lemma 5.2. In this proof we omit the dependence on n of the variables for ease of exposition. We set to zero the drift of $V_\ell(\boldsymbol{q}) = \sum_{i=1}^{n} q_i$. Before doing it, we should show that $\mathbb{E}\left[V_\ell(\overline{\boldsymbol{q}})\right] < \infty$. This result is a direct consequence of [34, Proposition 3], so we omit the proof. Setting the drift to zero we obtain

$$\begin{aligned}
0 &= \mathbb{E}\left[V_\ell(\overline{\boldsymbol{q}}^+) - V_\ell(\overline{\boldsymbol{q}})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left(\overline{q}_i^+ - \overline{q}_i\right)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{i=1}^{n} \left(\overline{a}_i - \overline{s}_i + \overline{u}_i\right)\right],
\end{aligned}$$

where (a) holds by the dynamics of the queues presented in Equation 1.2 and by definition of $\overline{\boldsymbol{q}}^+$. Rearranging terms we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \left(\overline{s}_i - \overline{a}_i\right)\right] \\
&\overset{(a)}{=} \sum_{i=1}^{n} \mathbb{E}\left[\overline{s}_i\right] - \mathbb{E}\left[\overline{a}\right] \\
&= n^{1-\alpha},
\end{aligned}$$

where (a) holds because $\overline{a} = \sum_{i=1}^{n} \overline{a}_i$ by definition. □

Observe that in the proof of Lemma 5.2 we do not specifically use that the routing

policy is JSQ. In fact, all we need is a routing policy such that $\mathbb{E}\left[\sum_{i=1}^{n} \overline{q}_i^{(n)}\right] < \infty$, and this condition is usually satisfied by throughput optimal policies. For example, power-of-d choices satisfies this condition as well.

5.3.1 State space collapse

The goal of this section is to show that when $\alpha > 4$, the average queue length in the load balancing system in the many-server heavy-traffic limit behaves similarly to the classical heavy-traffic regime. It is known that the load balancing system operating under JSQ exhibits one-dimensional SSC in classical heavy traffic, i.e., in the limit it behaves as a single-server queue. In [34], SSC is proved by showing that the error of approximating the actual vector of queue lengths by its projection on the line where SSC occurs is bounded, and the bound does not depend on the traffic intensity. Therefore, as the traffic intensity increases, this error becomes negligible. In this case, the traffic intensity depends on the number of servers, so we need to show that the bound becomes negligible as n increases. Before stating the result we introduce notation. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, let

$$\boldsymbol{x}_{\parallel} \stackrel{\triangle}{=} \left(\sum_{i=1}^{n} \frac{x_i}{n}\right) \mathbf{1} \quad \text{and} \quad \boldsymbol{x}_{\perp} \stackrel{\triangle}{=} \boldsymbol{x} - \boldsymbol{x}_{\parallel}. \tag{5.1}$$

Then, $\boldsymbol{x}_{\parallel}$ is the projection of $\boldsymbol{x}$ on the line generated by the vector $\mathbf{1}$, and $\boldsymbol{x}_{\perp}$ is the error of approximating $\boldsymbol{x}$ by $\boldsymbol{x}_{\parallel}$. In this section we prove the following proposition.

Proposition 5.3. *Consider a load balancing system operating under JSQ, parametrized by n as described above. Let $\delta \in (0,1)$ and $n_0 \in \mathbb{Z}_+$ be such that $\delta \leq 1 - n^{-\alpha}$ for all $n \geq n_0$. Then, there exists a finite constant C such that for any $j \in \mathbb{Z}_+$ with $j \geq 1$, we have*

$$\mathbb{E}\left[\left\|\overline{\boldsymbol{q}}_{\perp}^{(n)}\right\|^j\right]^{\frac{1}{j}} \leq Cjn^3. \tag{5.2}$$

Additionally, if $\theta \leq \frac{1}{4A_{\max}n^{\frac{5}{2}-\alpha}} \log\left(1+\frac{\delta}{4A_{\max}n^{\frac{3}{2}}}\right)$ *we have*

$$\mathbb{E}\left[\exp\left(\theta n^{1-\alpha}\left\|\overline{\boldsymbol{q}}_{\perp}^{(n)}\right\|\right)\right] \leq \frac{\delta \exp\left(\frac{2\theta n^{3-\alpha}}{\delta}\right)}{\delta + 4A_{\max}n^{\frac{3}{2}}\left(1-\exp\left(4\theta A_{\max}n^{\frac{5}{2}-\alpha}\right)\right)}. \tag{5.3}$$

In the proof of Proposition 5.3 we use the moment and exponential bounds based on drift arguments presented in Lemma 2.7 and Lemma 2.8.

5.3.2 Proof of Theorem 5.1 using transform method

In this section we prove Theorem 5.1 using the Transform method introduced in chapter 3. We use one-sided Laplace transform.

Proof of Theorem 5.1. For ease of exposition, in this proof we omit the dependence on n of the variables. The first step is to prove an 'exponential version' of the key property of the unused service presented in Equation 1.3. We state the result below, and we prove it in subsection 5.6.1.

Lemma 5.4. *Consider a load balancing system operating under JSQ, parametrized by* n *as described in Theorem 5.1, and suppose* $\alpha > 4$*. Let* $\hat{\theta} \stackrel{\triangle}{=} \theta\left(\frac{\sigma_a^2+\sigma_s^2}{2}\right)$*, where* $\theta < 0$ *is a finite parameter. Then, there exists* $\tilde{\Theta} > 0$ *such that for all* $|\hat{\theta}| < \tilde{\Theta}$ *we have*

$$\left|\mathbb{E}\left[\left(\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\left(\overline{q}_i^{(n)}\right)^{+}\right)-1\right)\left(\exp\left(-\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i^{(n)}\right)-1\right)\right]\right| \quad \textit{is } o\left(n^{1-2\alpha}\right),$$

where $\overline{q}_{\Sigma}^{+}$ *represents the total queue length one time slot after observing* $\overline{q}_{\Sigma}$*.*

Rearranging terms in the expression of Lemma 5.4 we obtain

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{q}_i^{+}\right)\right]-\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{q}_i\right)\right]\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)\right)\right]\\&=1-\mathbb{E}\left[\exp\left(-\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i\right)\right]+o\left(n^{1-2\alpha}\right),\end{aligned}$$

where we used the dynamics of the queues presented in Equation 1.2, the fact that $\overline{a} =$

$\sum_{i=1}^{n} \overline{a}_i$, and that the arrival process to the system and the potential service processes are independent of the queue lengths.

Since $\hat{\theta} < 0$, we know $\mathbb{E}\left[\exp\left(\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{q}_i\right)\right] \leq 1 < \infty$. Then, we can set its drift to zero, i.e., we set $\mathbb{E}\left[\exp\left(\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{q}_i\right)\right] = \mathbb{E}\left[\exp\left(\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{q}_i^{+}\right)\right]$. Using this property and rearranging terms we obtain

$$\mathbb{E}\left[\exp\left(\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{q}_i\right)\right] = \frac{1 - \mathbb{E}\left[\exp\left(-\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{u}_i\right)\right] + o\left(n^{1-2\alpha}\right)}{1 - \mathbb{E}\left[\exp\left(\hat{\theta} n^{-\alpha} \left(\overline{a} - \sum_{i=1}^{n} \overline{s}_i\right)\right)\right]} \tag{5.4}$$

Our goal is to take the limit as $n \to \infty$. Observe that the right-hand side of Equation 5.4 yields a $\frac{0}{0}$ form in the limit. Then, we take the Taylor expansion with respect to $\hat{\theta}$ of the numerator and denominator. For the numerator we obtain

$$1 - \mathbb{E}\left[\exp\left(-\hat{\theta} n^{-\alpha} \sum_{i=1}^{n} \overline{u}_i\right)\right] = \hat{\theta} n^{-\alpha} \mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] + o\left(n^{1-2\alpha}\right) = \hat{\theta} n^{1-2\alpha} + o\left(n^{1-2\alpha}\right),$$

where the last equality holds by Lemma 5.2. The $o\left(n^{1-2\alpha}\right)$ term arises because for all $j \geq 2$ we have

$$\begin{aligned}
\left| \frac{\hat{\theta}^j n^{-j\alpha}}{j!} \mathbb{E}\left[\left(\sum_{i=1}^{n} \overline{u}_i\right)^j\right] \right| &\overset{(a)}{=} \frac{|\hat{\theta}|^j n^{-j\alpha}}{j!} \mathbb{E}\left[\left(\sum_{i=1}^{n} \overline{u}_i\right)^{j-1} \left(\sum_{i=1}^{n} \overline{u}_i\right)\right] \\
&\overset{(b)}{\leq} \frac{|\hat{\theta}|^j S_{\max}^{j-1}}{j!} n^{j(1-\alpha)-1} \mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] \\
&\overset{(c)}{=} \frac{|\hat{\theta}|^j S_{\max}^{j-1}}{j!} n^{j-\alpha(j+1)},
\end{aligned}$$

where (a) holds because $\overline{u}_i \geq 0$ with probability 1 for all $i \in [n]$ by definition of unused service; (b) holds because $0 \leq \overline{u}_i \leq \overline{s}_i \leq S_{\max}$; and (c) holds by Lemma 5.2. Also, $j - \alpha(j+1) - (1-2\alpha) = (j-1)(1-\alpha)$, which is negative for all $\alpha > 1$. Then, it is negative for $\alpha > 4$.

For the denominator we obtain

$$
\begin{aligned}
&1-\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)\right)\right] \\
&=-\hat{\theta}n^{-\alpha}\mathbb{E}\left[\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right]-\frac{\hat{\theta}^2n^{-2\alpha}}{2}\mathbb{E}\left[\left(\overline{a}-\sum_{i=1}^{n}\overline{s}_i\right)^2\right]+o\left(n^{1-2\alpha}\right) \\
&\overset{(a)}{=}\hat{\theta}n^{1-2\alpha}-\frac{\hat{\theta}^2n^{1-2\alpha}}{2}\left(\sigma_a^2+\sigma_s^2\right)-\frac{\hat{\theta}^2n^{2-4\alpha}}{2}+o\left(n^{1-2\alpha}\right),
\end{aligned}
$$

where (a) holds by definition of variance and because $\mathbb{E}\left[\sum_{i=1}^{n}\overline{s}_i-\overline{a}\right]=n^{1-\alpha}$. The $o\left(n^{1-2\alpha}\right)$ arises similarly to the case of the numerator. We omit the details for brevity.

Putting everything together and canceling $\hat{\theta}n^{1-2\alpha}$ from the numerator and the denominator we obtain

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{q}_i\right)\right]&=\frac{1+o(1)}{1-\hat{\theta}\left(\frac{\sigma_a^2+\sigma_s^2}{2}\right)+o(1)} \\
&=\frac{1+o(1)}{1-\theta+o(1)},
\end{aligned}
$$

where the last equality holds because $\hat{\theta}=\frac{2\theta}{\sigma_a^2+\sigma_s^2}$. Therefore,

$$
\lim_{n\to\infty}\mathbb{E}\left[\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{q}_i\right)\right]=\frac{1}{1-\theta},
$$

which is the one-sided Laplace transform of an exponential random variable with mean 1. This completes the proof. □

5.4 Rate of convergence in Wasserstein's distance

A different approach to show Theorem 5.1 is using Stein's method, where one additionally obtains the rate of convergence in Wasserstein's distance. We present a definition of this metric and the result below.

Definition 5.5. *For two probability measures ν_1 and ν_2, the Wasserstein's distance between them is*

$$d_W(\nu_1, \nu_2) \stackrel{\triangle}{=} \sup_{h \in \text{Lip(1)}} \left| \int h(x)\, d\nu_1(x) - \int h(x)\, d\nu_2(x) \right|,$$

where $\text{Lip(1)} = \{h : \mathbb{R} \to \mathbb{R}$ *such that* $|h(x) - h(y)| \leq |x - y|\}$ *is the set of Lipschitz functions with constant 1.*

For random variables X and Y with laws ν_1 and ν_2, respectively, we write $d_W(X, Y)$ instead of $d_W(\nu_1, \nu_2)$, and when the measures are clear from the context we write

$$d_W(X, Y) = \sup_{h \in \text{Lip(1)}} \left| \mathbb{E}\left[h(X)\right] - \mathbb{E}\left[h(Y)\right] \right|. \tag{5.5}$$

Theorem 5.6. *Consider a load balancing system operating under JSQ, parametrized by n as described in Section 5.3. Let Z be an exponential random variable with mean 1. Then, we have*

$$\begin{aligned} &d_W\left(\frac{2n^{-\alpha}}{\sigma_a^2 + \sigma_s^2} \sum_{i=1}^{n} \overline{q}_i^{(n)}, Z \right) \\ &\leq \frac{1}{\sigma_a^2 + \sigma_s^2} \Big(5S_{\max} n^{1-\alpha} + n^{1-2\alpha} + CS_{\max}(\alpha - 1) n^{4-\alpha} \left\lceil \log(n) \right\rceil n^{\frac{1}{\lceil \log(n) \rceil}} \\ &\qquad + \frac{4n^{2-3\alpha}}{3(\sigma_a^2 + \sigma_s^2)} \left(A_{\max} + 2S_{\max} \right)^3 \Big). \end{aligned}$$

where $\overline{C} \stackrel{\triangle}{=} CS_{\max} e^{\frac{1}{2e}+1}$ and C is the constant from Proposition 5.3.

It is known that convergence to zero of the Wasserstein's distance implies convergence in distribution [92, Theorem 2]. Hence, Theorem 5.1 is an immediate consequence of Theorem 5.6.

We omit the proof of Theorem 5.6, since it is very similar to the proof of Theorem 5.19.

5.5 Load balancing under power-of-d choices

As mentioned in chapter 4, JSQ is a special case of the power-of-d choices routing algorithm. In this section, we present the many-server heavy-traffic asymptotics of the average queue length in a load balancing system operating under this routing algorithm. We omit the proof, since it is similar to the proofs in chapter 4 and section 5.3.

Theorem 5.7. *Consider a sequence of load balancing systems operating under power-of-d choices, parametrized by n as described in section 5.3. Let $d = cn^{\beta}$, where $c, \beta > 0$ are constants. If $\alpha + \beta > \frac{11}{2}$, then $n^{-\alpha} \sum_{i=1}^{n} \overline{q}_i^{(n)} \Rightarrow \Upsilon$ as $n \to \infty$, where Υ is an exponential random variable with mean $\frac{\sigma_a^2+\sigma_s^2}{2}$.*

Below we state the SSC result that we use in the proof of Theorem 5.7.

Proposition 5.8. *Consider a load balancing system operating under power-of-d choices, parametrized by n as described above. Let $\delta \in (0,1)$ and $n_0 \in \mathbb{Z}_+$ be such that $\delta \leq 1 - n^{-\alpha}$ for all $n \geq n_0$. Then, there exists a finite constant C such that for any $j \in \mathbb{Z}_+$ with $j \geq 1$, we have*

$$\mathbb{E}\left[\left\|\overline{\boldsymbol{q}}_{\perp}^{(n)}\right\|^{j}\right]^{\frac{1}{j}} \leq C\left(\frac{jn^{\frac{9}{2}}}{d-1}\right).$$

Additionally, if $\theta \leq \frac{1}{4\tilde{A}_{\max} n^{\frac{5}{2}-\alpha}} \log\left(1 + \frac{\delta}{4\tilde{A}_{\max} n^5}\right)$ we have

$$\mathbb{E}\left[\exp\left(\theta n^{1-\alpha}\left\|\overline{\boldsymbol{q}}_{\perp}^{(n)}\right\|\right)\right] \leq \frac{(1-\delta)(d-1)\exp\left(\frac{2\theta n^{\frac{9}{2}-\alpha}}{(1-\delta)(d-1)}\right)}{(1-\delta)(d-1) + 4\tilde{A}_{\max} n^5\left(1 - \exp\left(4\theta\tilde{A}_{\max} n^{\frac{5}{2}-\alpha}\right)\right)}.$$

Recall that if we set $d = n$, power-of-d choices is equivalent to JSQ. However, if we use $d = n$ in Theorem 5.7 and Proposition 5.8, we do not recover the results from section 5.3. Exploring this gap is future work.

5.6 Details of the proofs of section 5.3

5.6.1 Proof of Lemma 5.4

We use the following lemma, which was proved in chapter 3 (see Lemma 3.14). We state it here for completeness.

Lemma 5.9. *Consider a load balancing system operating under JSQ, parametrized by n as described in section 5.3. Then, for any $\zeta \in \mathbb{R}$ and $k \in \mathbb{Z}_+$ we have*

$$\sum_{i=1}^{n} u_i^{(n)}(k) \left(\exp\left(\tfrac{\zeta}{n}\sum_{j=1}^{n} q_j^{(n)}(k+1)\right) - 1\right) = \sum_{i=1}^{n} u_i^{(n)}(k) \left(\exp\left(-\zeta q_{\perp i}^{(n)}(k+1)\right) - 1\right),$$

where $q_{\perp i}^{(n)}(k+1)$ is the i^{th} element of $\boldsymbol{q}_{\perp}^{(n)}(k+1)$.

Now we prove Lemma 5.4.

Proof of Lemma 5.4. We have

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\exp\left(\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^{n}\overline{q}_i^{+}\right) - 1\right)\left(\exp\left(-\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^{n}\overline{u}_i\right) - 1\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|\left(\exp\left(\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^{n}\overline{q}_i^{+}\right) - 1\right)\left(\exp\left(-\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^{n}\overline{u}_i\right) - 1\right)\right|\right] \\
&\overset{(a)}{=} |\hat{\theta}|n^{-\alpha}\mathbb{E}\left[\left(\sum_{i=1}^{n}\overline{u}_i\right)\left|\exp\left(\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^{n}\overline{q}_i^{+}\right) - 1\right|\left(\frac{\exp\left(-\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i\right) - 1}{-\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i}\right)\mathbb{1}_{\left\{\sum_{i=1}^{n}\overline{u}_i \neq 0\right\}}\right] \\
&\overset{(b)}{\leq} |\hat{\theta}|n^{-\alpha}\left(\frac{\exp\left(-\hat{\theta}n^{-\alpha}S_{\max}\right) - 1}{-\hat{\theta}n^{-\alpha}S_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i\left|\exp\left(\hat{\theta}n^{-\alpha}\textstyle\sum_{j=1}^{n}\overline{q}_j^{+}\right) - 1\right|\right] \\
&\overset{(c)}{\leq} |\hat{\theta}|n^{-\alpha}\left(\frac{\exp\left(-\hat{\theta}n^{-\alpha}S_{\max}\right) - 1}{-\hat{\theta}n^{-\alpha}S_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i\left|\exp\left(-\hat{\theta}n^{1-\alpha}\overline{q}_{\perp i}\right) - 1\right|\right] \\
&\overset{(d)}{\leq} |\hat{\theta}|n^{-\alpha}\left(\frac{\exp\left(-\hat{\theta}n^{-\alpha}S_{\max}\right) - 1}{-\hat{\theta}n^{-\alpha}S_{\max}}\right)\mathbb{E}\left[\sum_{i=1}^{n}\overline{u}_i^{j}\right]^{\frac{1}{j}}\mathbb{E}\left[\sum_{i=1}^{n}\left|\exp\left(-\hat{\theta}n^{1-\alpha}\overline{q}_{\perp i}\right) - 1\right|^{\frac{j}{j-1}}\right]^{\frac{j-1}{j}}
\end{aligned} \tag{5.6}$$

where $j > 1$ is an integer number. Here (a) holds after multiplying and dividing by $|\hat{\theta}|n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i$, because if $\sum_{i=1}^{n}\overline{u}_i = 0$ then $\exp\left(\hat{\theta}n^{-\alpha}\sum_{i=1}^{n}\overline{u}_i\right) - 1 = 0$, and because

the function $f(x) = \frac{e^x - 1}{x}$ is nonnegative; (b) holds because the function $f(x) = \frac{e^x - 1}{x}$ is nonnegative and increasing, and because $0 \leq \overline{u}_i \leq S_{\max}$ with probability 1 for all $i \in [n]$; (c) holds by Lemma 5.9 and the triangle inequality; and (d) holds by Hölder's inequality.

We analyze each expression in Equation 5.6 separately. First observe that

$$\lim_{n\to\infty} \frac{\exp\left(-\hat{\theta} n^{-\alpha} S_{\max}\right) - 1}{-\hat{\theta} n^{-\alpha} S_{\max}} = 1.$$

The unused service is nonnegative by definition, then

$$0 \leq \mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i^j\right] \overset{(a)}{\leq} S_{\max}^{j-1}\, \mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] \overset{(b)}{=} S_{\max}^{j-1} n^{1-\alpha}$$

where (a) holds because $\overline{u}_i \leq S_{\max}$ with probability 1 for all $i \in [n]$; and (b) holds by Lemma 5.2.

For the last term we use Hölder's inequality again. Let $j' > 1$ be an integer number. Then, for each $i \in [n]$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\exp\left(-\hat{\theta} n^{1-\alpha} \overline{q}_{\perp i}\right) - 1\right|^{\frac{j}{j-1}}\right] \\
&= |\hat{\theta}|^{\frac{j}{j-1}} n^{\frac{j}{j-1}(1-\alpha)} \mathbb{E}\left[\left(\frac{\exp\left(-\hat{\theta} n^{1-\alpha} \overline{q}_{\perp i}\right) - 1}{-\hat{\theta} n^{1-\alpha} \overline{q}_{\perp i}}\right)^{\frac{j}{j-1}} |\overline{q}_{\perp i}|^{\frac{j}{j-1}}\, \mathbb{1}_{\{\overline{q}_{\perp i} \neq 0\}}\right] \\
&\overset{(a)}{\leq} \hat{\theta}^{\frac{j}{j-1}} n^{\frac{j}{j-1}(1-\alpha)} \left(\mathbb{E}\left[\left|\frac{\exp\left(-\hat{\theta} n^{1-\alpha} \overline{q}_{\perp i}\right) - 1}{-\hat{\theta} n^{1-\alpha} \overline{q}_{\perp i}}\right|^{\left(\frac{j}{j-1}\right)\left(\frac{j'}{j'-1}\right)} \mathbb{1}_{\{\overline{q}_{\perp i} \neq 0\}}\right]\right)^{\frac{j'-1}{j'}} \left(\mathbb{E}\left[|\overline{q}_{\perp i}|^{\frac{j}{j-1} j'}\right]\right)^{\frac{1}{j'}},
\end{aligned}$$

where $j' > 1$ is an integer number. Here, (a) holds by Hölder's inequality. We bound each of these terms.

Using Proposition 5.3 for the last term we obtain

$$0 \leq \mathbb{E}\left[|\overline{q}_{\perp i}|^{\frac{j}{j-1} j'}\right]^{\frac{j-1}{jj'}} \overset{(a)}{\leq} \mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|^{\frac{j}{j-1} j'}\right]^{\frac{j-1}{jj'}} \leq C n^3 \left(\frac{jj'}{j-1}\right), \tag{5.7}$$

where (a) holds if $\frac{j}{j-1}j' \geq 2$, by the inequalities between norms.

On the other hand, since $\frac{e^x-1}{x} \leq e^{|x|}$, we obtain

$$\left(\mathbb{E}\left[\left|\frac{\exp\left(-\hat{\theta}n^{1-\alpha}\overline{q}_{\perp i}\right)-1}{-\hat{\theta}n^{1-\alpha}\overline{q}_{\perp i}}\right|^{\left(\frac{j}{j-1}\right)\left(\frac{j'}{j'-1}\right)}\mathbb{1}_{\{\overline{q}_{\perp i}\neq 0\}}\right]\right)^{\frac{j'-1}{j'}}$$
$$\leq \left(\mathbb{E}\left[\exp\left(|\hat{\theta}|n^{1-\alpha}\left(\tfrac{j'-1}{j'}\right)\left(\tfrac{j-1}{j}\right)|\overline{q}_{\perp i}|\right)\right]\right)^{\frac{j'-1}{j'}}, \tag{5.8}$$

and the last term can be bounded similarly to Equation 5.7, using Lemma 2.8 for

$$|\hat{\theta}|\left(\frac{j'-1}{j'}\right)\left(\frac{j-1}{j}\right) \leq \frac{1}{4A_{\max}n^{\frac{5}{2}-\alpha}}\log\left(1+\frac{\delta}{4A_{\max}n^{\frac{3}{2}}}\right).$$

Observe that the right-hand side grows to infinity as $n \to \infty$. Then, there exists $n_0^* \in \mathbb{Z}_+$ such that the lemma is satisfied with

$$\tilde{\Theta} \triangleq \left(\frac{j'}{j'-1}\right)\left(\frac{j}{j-1}\right)\left(\frac{1}{4A_{\max}(n_0^*)^{\frac{5}{2}-\alpha}}\right)\log\left(1+\frac{\delta}{4A_{\max}(n_0^*)^{\frac{3}{2}}}\right).$$

Further, the upper bound in Equation 5.8 converges to a constant as $n \to \infty$. We omit the details for brevity.

Putting everything together in Equation 5.6 we obtain

$$\left|\mathbb{E}\left[\left(\exp\left(\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^n \overline{q}_i^+\right)-1\right)\left(\exp\left(-\hat{\theta}n^{-\alpha}\textstyle\sum_{i=1}^n \overline{u}_i\right)-1\right)\right]\right| \leq L(n)\, n^{4-2\alpha+\frac{1-\alpha}{p}},$$

where $L(n)$ is of order $O(1)$. Finally, observe that $4-2\alpha+\frac{1-\alpha}{j} < 1-2\alpha$ if and only if $\alpha > 3j+1$, and j can be taken as close to one as desired. Therefore, the lemma holds for all $\alpha > 4$. $\square$

5.7 Load balancing in continuous time model and asymptotic result

In the load balancing model that we studied in the previous sections of this chapter, we have that the expected number of arrivals per time slot is $n(1-n^{-\alpha})$, and all the arrivals of each time slot are routed to the same server. Then, as n grows, the number of jobs that are routed in each time slot tends to grow and the number of jobs in the server that receives the new arrivals dramatically grows. Hence, the SSC result that we proved is weak. Even if we have $\boldsymbol{q}^{(n)}(k) = \boldsymbol{q}_{\parallel}^{(n)}(k)$ for a certain $k \in \mathbb{Z}_+$ (i.e., perfect SSC), we will have a large $\boldsymbol{q}_{\perp}^{(n)}(k+1)$ if n is large. Intuitively, if we were able to route the jobs one by one, we would obtain a better SSC result and, hence, we would obtain that the average queue lengths behaves as in the classical heavy-traffic regime for smaller values of α. In this section, we propose to work with a continuous-time model, and we observe that the value of α indeed improves and, additionally, we no longer have the gap between power-of-d choices with $d = n$ and JSQ. We start specifying the model.

Consider a load balancing system operating in continuous time, that is, a queueing system with n parallel servers, each of them with an infinite buffer. Arrivals to the system occur according to a Poisson process at rate λn, where $\lambda \in (0,1)$. Upon arrival, a dispatcher immediately routes the new job to one of the servers, where they wait in line until the server can process them. All the servers are identical, and all the arriving jobs have exponential size with mean 1. Routing occurs according to power-of-d choices, where $d \in \mathbb{Z}_+$ is of the form $d \stackrel{\triangle}{=} cn^{\beta}$ for constants $c > 0$ and $\beta \in [0,1]$ such that $d \in [n]$. Specifically, upon arrival, d servers are sampled uniformly at random and the new job is routed to the server with the shortest queue among those d. Ties are broken uniformly at random. Observe that if $c = \beta = 1$, then $d = n$ and power-of-d choices is equivalent to JSQ.

For each $t \in \mathbb{R}_+$ and each $i \in [n]$, let $q_i(t)$ be the number of jobs in queue i at time t, including the job in service (if any). Then, the queue length process $\{\boldsymbol{q}(t) : t \in \mathbb{R}_+\}$ is

a continuous-time Markov chain (CTMC) with the generator matrix G defined in Equation 5.9. Let $\boldsymbol{q} \in \mathbb{Z}_+^n$, and for each $i \in [n]$ let $\psi_{\boldsymbol{q}}(i)$ be the index of the i^{th} smallest element of $\boldsymbol{q}$. Then, for any $\boldsymbol{q}' \in \mathbb{Z}_+^n$ we have that the transition rate from state $\boldsymbol{q}$ to state $\boldsymbol{q}'$ is

$$G_{\boldsymbol{q},\boldsymbol{q}'} \triangleq \begin{cases} -\left(\lambda n + \sum_{i=1}^n \mathbb{1}_{\{q_i>0\}}\right) & \text{if } \boldsymbol{q} = \boldsymbol{q}', \\ 1 & \text{if } q_i > 0 \text{ and } \boldsymbol{q}' = \boldsymbol{q} - \boldsymbol{e}^{(i)}, \text{ with } i \in [n], \\ \lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} & \text{if } \boldsymbol{q}' = \boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

The first case is the additive inverse of the sum of the other cases; the second case corresponds to a departure from queue i, which occurs at rate 1 and it can only happen if the queue is nonempty; and the third case corresponds to an arrival to the i^{th} shortest queue, and the rate holds because of the following reason. Arrivals occur at rate λn, and the new arrival is routed to the i^{th} shortest queue (which has index $\psi_{\boldsymbol{q}}(i)$) if it is the shortest among the d randomly selected queues. There are $\binom{n}{d}$ ways to select d queues uniformly at random and, out of those selections, only $\binom{n-i}{d-1}$ would result in routing to the i^{th} shortest queue.

We are interested in the steady-state analysis of the load balancing system described above. First observe that the Markov chain is irreducible and nonexplosive. Additionally, the total arrival rate to the system (λn) is strictly smaller than the total service rate (n) for any $\lambda \in (0, 1)$. Then, the queue length process is also positive recurrent. Hence, stationary distribution exists and it is unique [37, Proposition 6.9b]. Let $\overline{\boldsymbol{q}}$ be a steady-state random vector which is limit in distribution of $\{\boldsymbol{q}(t) : t \in \mathbb{R}_+\}$, and define $\overline{q}_\Sigma \triangleq \sum_{i=1}^n \overline{q}_i$.

We parametrize the load balancing system as follows. Consider $\alpha > 0$ and let $\lambda^{(n)} \triangleq 1 - n^{-\alpha}$ be the arrival rate *per server* to the system. Then, the *total* arrival rate is $\lambda^{(n)} n$. Let $\left\{\boldsymbol{q}^{(n)}(t) : t \in \mathbb{R}_+\right\}$ be the queue length process of the n^{th} system and $\overline{\boldsymbol{q}}^{(n)}$ a steady-state random vector which is limit in distribution of $\left\{\boldsymbol{q}^{(n)}(t) : t \in \mathbb{R}_+\right\}$. In the next theorem we present the main result of this chapter.

Theorem 5.10. *Consider a sequence of load balancing systems operating under power-of-d, parametrized by n as described above. If $\alpha + \beta > 3$, then $n^{-\alpha}\overline{q}_{\Sigma}^{(n)} \Rightarrow \Upsilon$ as $n \to \infty$, where Υ is an exponential random variable with mean 1.*

Immediate corollaries of Theorem 5.10 are the cases of power-of-d with constant d, and JSQ. We formally present these results below.

Corollary 5.11. *Consider a sequence of load balancing systems operating under power-of-d choices, parametrized by n as described Theorem 5.10. Suppose $d = c$, where $c \in \mathbb{Z}_+$ is a fixed parameter. If $\alpha > 3$, then $n^{-\alpha}\overline{q}_{\Sigma}^{(n)} \Rightarrow \Upsilon$ as $n \to \infty$, where Υ is an exponential random variable with mean 1.*

The proof of Corollary 5.11 holds easily after setting $\beta = 0$. Now we present a result for the load balancing system under JSQ.

Corollary 5.12. *Consider a sequence of load balancing systems operating under JSQ, parametrized by n as described in Theorem 5.10. If $\alpha > 2$, then $n^{-\alpha}\overline{q}_{\Sigma}^{(n)} \Rightarrow \Upsilon$ as $n \to \infty$, where Υ is an exponential random variable with mean 1.*

The proof of Corollary 5.12 holds after letting $c = \beta = 1$ in Theorem 5.10.

Remark 5.13. *In the classical heavy-traffic regime, one defines the heavy-traffic parameter as $\epsilon \stackrel{\triangle}{=} \mu_\Sigma - \lambda_\Sigma$, where μ_Σ is the total service rate (the sum of the mean service rate of each server) and λ_Σ is the total arrival rate. Then, one parametrizes the vector of queue lengths by ϵ and can show that $\epsilon\frac{1}{n}\sum_{i=1}^{n}\overline{q}_i^{(\epsilon)} \Rightarrow \tilde{\Upsilon}$, where $\tilde{\Upsilon}$ is an exponential random variable whose mean depends on the variance of the arrival and service processes. Further, one can show that $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \Rightarrow \mathbf{1}\tilde{\Upsilon}$. We presented the details of this analysis in chapter 3.*

In this chapter, we study the many-server heavy-traffic regime, and our goal is to find the value of α such that the scaled average queue length converges in distribution to an exponential random variable. In Theorem 5.10 we show convergence in distribution of the total queue length scaled by $n^{-\alpha}$, which is equivalent to the average queue length scaled

by $n^{1-\alpha}$. Additionally, observe that the difference between the total service and arrival rate in this chapter is $n^{1-\alpha}$. In other words, in the many-server heavy-traffic regime, $n^{1-\alpha}$ plays the role of the heavy-traffic parameter ϵ. Further, in the classical heavy-traffic regime there is an analogous result to Lemma 5.15 (proved in subsection 5.8.1), which is key to bound the unused service.

5.8 Multiplicative state space collapse

Before stating the result formally, we introduce some notation. Given a vector $\boldsymbol{x} \in \mathbb{R}^n_+$, define

$$\boldsymbol{x}_{\parallel} \triangleq \mathbf{1}\left(\frac{\sum_{i=1}^n x_i}{n}\right), \text{ and } \boldsymbol{x}_{\perp} \triangleq \boldsymbol{x} - \boldsymbol{x}_{\parallel}. \tag{5.10}$$

Then, $\boldsymbol{x}_{\parallel}$ is the projection of $\boldsymbol{x}$ on the line generated by $\mathbf{1}$, and $\boldsymbol{x}_{\perp}$ represents the error of approximating $\boldsymbol{x}$ by $\boldsymbol{x}_{\parallel}$. Now we introduce the result.

Proposition 5.14. *Consider a load balancing system operating under power-of-d choices, as described in section 5.7, and let $\lambda_0 \in (0,1)$. If c and β are such that $d = cn^{\beta} \geq 2$, then for any $\lambda \in (\lambda_0, 1)$:*

1. *There exists a finite constant C, which is independent of λ, β and n, such that for any positive integer j we have*

$$\mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|^j\right] \leq C\left(\frac{n^2}{cn^{\beta}-1}\right)^j j^{j+\frac{1}{2}}. \tag{5.11}$$

2. *Let e be Euler's constant and $\overline{C} \triangleq C\exp\left(\frac{1}{2e}\right)$. Then, for any positive integer j we have*

$$\mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|^j\right]^{\frac{1}{j}} \leq \overline{C} j\left(\frac{n^2}{cn^{\beta}-1}\right). \tag{5.12}$$

3. Let $\theta^* \in \mathbb{R}$ be such that $|\theta^*| < \frac{1}{4}\log\left(1+\frac{\lambda_0(cn^\beta-1)}{8n}\right)$. Then,

$$\mathbb{E}\left[\exp\left(\theta^*\|\overline{\boldsymbol{q}}_\perp\|\right)\right] \leq \frac{\lambda_0(cn^\beta-1)\exp\left(\frac{\theta^* n^2}{\lambda_0(cn^\beta-1)}\right)}{\lambda_0(cn^\beta-1)+2n\left(1-\exp\left(4\theta^*\right)\right)}. \tag{5.13}$$

In Proposition 5.14 we give an upper bound for the j^{th} moment, the j^{th} norm and the MGF of $\|\boldsymbol{q}_\perp\|$. Observe that these upper bounds are independent of λ, but they depend on n and d. Our SSC result is multiplicative because the bounds grow to infinity as $n \to \infty$. However, after scaling, they become negligible.

Before presenting the proof of Proposition 5.14 we present a preliminary result that will be repeatedly used in the rest of this section.

5.8.1 Preliminary result

A key challenge when studying the load balancing system described in section 5.7 is to handle the indicator function that arises from the second case in Equation 5.9. Specifically, when one computes the drift of any function, we get a term of the form $\mathbb{1}_{\{q_i>0\}}$ for each $i \in [n]$. This term represents that service cannot occur at empty queues and, therefore, the queue lengths cannot be negative. In this chapter, we handle this indicator function using the property $\mathbb{1}_{\{q_i>0\}} = 1 - \mathbb{1}_{\{q_i=0\}}$ for every $i \in [n]$ and the following lemma. In fact, Lemma 5.15 is repeatedly used in the proof of SSC, in the two proofs that we provide for Theorem 5.10 and in the proof of Theorem 5.21.

Lemma 5.15. *Consider a load balancing system as described above, where the routing policy is throughput optimal. Let $\lambda \in (0,1)$ be the arrival rate per server, and let $\overline{\boldsymbol{q}}$ be a steady-state random vector which is limit in distribution of $\{\boldsymbol{q}(t) : t \in \mathbb{R}_+\}$. Then, if $\mathbb{E}\left[\overline{q}_\Sigma\right] < \infty$ we have*

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right] = n(1-\lambda).$$

Note that if we use $\lambda^{(n)}$ in this lemma, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\left\{\overline{q}_i^{(n)}=0\right\}}\right] = n^{1-\alpha}.$$

Now we prove the result.

Proof of 5.15. We consider the test function $V_\ell(\boldsymbol{q}) = \sum_{i=1}^{n} q_i$. Since $\mathbb{E}\left[\overline{q}_\Sigma\right] < \infty$, we can set to zero the drift of $V_\ell(\boldsymbol{q})$ in steady state. Note that we use the definition of drift in continuous time, provided in Definition 2.3. We obtain

$$\begin{aligned}\Delta V_\ell(\boldsymbol{q}) &= \lambda n \left(\left(\sum_{i=1}^{n} q_i + 1\right) - \sum_{i=1}^{n} q_i\right) + \sum_{j=1}^{n} \left(1 - \mathbb{1}_{\{q_j=0\}}\right) \left(\left(\sum_{i=1}^{n} q_i - 1\right) - \sum_{i=1}^{n} q_i\right) \\ &= \lambda n - n + \sum_{i=1}^{n} \mathbb{1}_{\{q_i=0\}}.\end{aligned}$$

Taking expected value with respect to the stationary distribution, and reorganizing terms we obtain the result. □

Observe that in Lemma 5.15 we do not need to assume that routing occurs according to power-of-d choices. In fact, this proof has only two steps: (i) Determine the drift of the function $V_\ell(\boldsymbol{q})$; and (ii) Set the drift to zero in steady state. In the first step, we do not use the details of the generator matrix. We only use that, if there is an arrival, the total queue length increases by 1 and, if there is a departure (provided that the system is not empty), the total queue length decreases by 1. In the second step we need $\mathbb{E}\left[\overline{q}_\Sigma\right] < \infty$ because, otherwise, we cannot set the drift to zero. The last property can be frequently concluded from the proof of throughput optimality. We prove that the load balancing system operating under power-of-d choices satisfies this condition below.

Proposition 5.16. *Consider a load balancing system operating under power-of-d choices, parametrized by n as described in section 5.7. Then, $\mathbb{E}\left[\overline{q}_\Sigma^{(n)}\right] < \infty$.*

Proof of Proposition 5.16. We use Theorem 2.6 with $Z(\boldsymbol{q}) = \|\boldsymbol{q}\|^2 = \sum_{i=1}^{n} q_i^2$. Using the

definition of drift, we obtain

$$\begin{aligned}
\Delta Z(\boldsymbol{q}) &= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(\left\| \boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))} \right\|^2 - \|\boldsymbol{q}\|^2 \right) + \sum_{i=1}^{n} \left(1 - \mathbb{1}_{\{q_i=0\}}\right) \left(\left\| \boldsymbol{q} - \boldsymbol{e}^{(i)} \right\|^2 - \|\boldsymbol{q}\|^2 \right) \\
&= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(1 + 2q_{(i)}\right) + \sum_{i=1}^{n} (1 - \mathbb{1}_{\{q_i=0\}}) \left(1 - 2q_i\right) \\
&\overset{(a)}{=} n(\lambda + 1) + 2 \sum_{i=1}^{n} q_{(i)} \left(\lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} - 1 \right) - \sum_{i=1}^{n} \mathbb{1}_{\{q_i=0\}} \\
&\overset{(b)}{\leq} n(\lambda + 1) + 2 \sum_{i=1}^{n} q_{(i)} \left(\lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} - 1 \right),
\end{aligned}$$

where (a) holds because $\mathbb{1}_{\{q_i=0\}} q_i = 0$ for all $i \in [n]$; and (b) holds because $\mathbb{1}_{\{q_i=0\}} \geq 0$ for all $i \in [n]$.

Now we bound the last term. For each $i \in [n]$ define

$$\overline{\gamma}_i \triangleq \lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} - 1.$$

Observing the analogy of $\overline{\gamma}_i$ and the parameters α_i defined in Equation 4.7. Note that:

(i) $\overline{\gamma}_i = -1$ for all $i > n - d + 1$,

(ii) The total sum satisfies:

$$\sum_{i=1}^{n} \overline{\gamma}_i = -n^{1-\alpha},$$

(iii) For every $\ell \in [n]$, the partial sums satisfy

$$\sum_{i=\ell}^{n} \overline{\gamma}_i \leq -\zeta_1,$$

where $\zeta_1 \triangleq \min \left\{ d - 1, \frac{n^{1-\alpha}}{\binom{n}{d}} \right\}$.

Using these three facts we obtain

$$\begin{aligned}\sum_{i=1}^{n}\overline{\gamma}_i q_{(i)} &= q_{(1)}\sum_{i=1}^{n}\overline{\gamma}_i + \sum_{\ell=2}^{n}\left(\sum_{i=\ell}^{n}\overline{\gamma}_i\right)\left(q_{(\ell)} - q_{(\ell-1)}\right)\\ &\leq -n^{1-\alpha}q_{(1)} - \zeta_1\left(q_{(n)} - q_{(1)}\right)\\ &\overset{(a)}{\leq} -\zeta_1 q_{(n)},\end{aligned}$$

where (a) holds because $n^{1-\alpha} \geq \zeta_1$ by definition of ζ_1. Therefore, we obtain

$$\Delta Z(\boldsymbol{q}) \leq n(\lambda+1) - 2\zeta_1 q_{(n)}.$$

Hence, conditions 1 and 2 from Theorem 2.6 are satisfied with

$$f(\boldsymbol{q}) = 2\zeta_1 q_{(n)}, \quad \text{and} \quad g(\boldsymbol{q}) = n(\lambda+1).$$

Condition 3 is trivially satisfied. Therefore, we obtain

$$\mathbb{E}\left[\overline{q}_{(n)}\right] \leq \frac{n(\lambda+1)}{2\zeta_1},$$

and since $q_{(n)} \overset{\triangle}{=} \max_{i\in[n]} q_i$ we obtain

$$\mathbb{E}\left[\overline{q}_{\Sigma}\right] \leq \frac{n^2(\lambda+1)}{2\zeta_1},$$

which is finite for every $n \in \mathbb{Z}_+$. □

5.8.2 Proof of Proposition 5.14

In the proof of Proposition 5.14 we use the moment bounds from Lemma 2.9 and the bound on the MGF from Lemma 2.10.

Proof of Proposition 5.14. We prove the proposition using d instead of cn^{β}, for ease of

exposition. Before providing the proof recall the following notation. Given a vector $\boldsymbol{q} \in \mathbb{Z}_+^n$, let

$$V(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}\|^2, \; V_{\|}(\boldsymbol{q}) \stackrel{\triangle}{=} \left\|\boldsymbol{q}_{\|}\right\|^2, \text{ and } W_{\perp}(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_{\perp}\|. \tag{5.14}$$

We verify the conditions of Lemma 2.9 with $Z(\boldsymbol{q}) = W_{\perp}(\boldsymbol{q})$. Then, Equation 2.4 and Stirling's approximation yield Equation 5.11. Equation 5.12 is a direct consequence of Equation 5.11, and Equation 5.13 is a consequence of Lemma 2.10.

To verify the first condition, observe that for any $\boldsymbol{q} \in \mathbb{Z}_+^n$ we have

$$\Delta W_{\perp}(\boldsymbol{q}) \leq \frac{1}{2\|\boldsymbol{q}_{\perp}\|}\left(\Delta V(\boldsymbol{q}) - \Delta V_{\|}(\boldsymbol{q})\right). \tag{5.15}$$

The proof of Equation 5.15 holds by concavity of the function $g(x) = \sqrt{x}$, and is presented in subsection 5.8.3. Therefore, it suffices to compute an upper bound for $\Delta V(\boldsymbol{q})$ and a lower bound for $\Delta V_{\|}(\boldsymbol{q})$. We first find an upper bound for $\Delta V(\boldsymbol{q})$. We have

$$\begin{aligned}
&\Delta V(\boldsymbol{q}) \\
&= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(\left\|\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right\|^2 - \|\boldsymbol{q}\|^2\right) + \sum_{i=1}^{n} \left(1 - \mathbb{1}_{\{q_i=0\}}\right) \left(\left\|\boldsymbol{q} - \boldsymbol{e}^{(i)}\right\|^2 - \|\boldsymbol{q}\|^2\right) \\
&= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(1 + 2q_{(i)}\right) + \sum_{i=1}^{n} (1 - \mathbb{1}_{\{q_i=0\}}) \left(1 - 2q_i\right) \\
&\stackrel{(a)}{\leq} n(\lambda+1) - 2(1-\lambda)\sum_{i=1}^{n} q_i + 2\lambda \sum_{i=1}^{n} \left(\frac{n\binom{n-i}{d-1}}{\binom{n}{d}} - 1\right) q_{(i)},
\end{aligned} \tag{5.16}$$

where (a) holds because $\mathbb{1}_{\{q_i=0\}} q_i = 0$ for all $i \in [n]$, because $\sum_{i=1}^{n} \mathbb{1}_{\{q_i=0\}} \geq 0$ and reorganizing terms.

We compute an upper bound for the last term of Equation 5.16. For each $i \in [n]$ define

$$\gamma_i = \frac{n\binom{n-i}{d-1}}{\binom{n}{d}}, \tag{5.17}$$

and observe that $\gamma_i = 0$ for $i \geq n - d + 1$. Then,

$$\sum_{i=1}^{n} \left(\frac{n\binom{n-i}{d-1}}{\binom{n}{d}} - 1 \right) q_{(i)} = \sum_{i=1}^{n} (\gamma_i - 1)\, q_{(i)}.$$

Observe that $\gamma_1 = d$. Then,

$$\begin{aligned}
&\sum_{i=1}^{n} (\gamma_i - 1)\, q_{(i)} \\
&= (d-1) q_{(1)} + \sum_{i=2}^{n} (\gamma_i - 1)\, q_{(i)} \\
&\overset{(a)}{=} \left(\frac{d-1}{n} \right) \sum_{i=1}^{n} \left(q_{(1)} - q_i \right) + \sum_{i=1}^{n} \left(\gamma_i - \frac{n-d+1}{n} \right) q_{(i)} - (d-1) q_{(1)} \\
&\overset{(b)}{\leq} -\frac{d-1}{n} \left\| \boldsymbol{q}_{\perp} \right\|, \qquad (5.18)
\end{aligned}$$

where (a) holds after reorganizing terms; and (b) holds by the claim below.

Claim 5.17. *Consider a load balancing system as described in Proposition 5.14, and let γ_i be as defined in Equation 5.17 for each $i \in [n]$. Then,*

$$\left(\frac{d-1}{n} \right) \sum_{i=1}^{n} \left(q_{(1)} - q_i \right) + \sum_{i=1}^{n} \left(\gamma_i - \frac{n-d+1}{n} \right) q_{(i)} - (d-1) q_{(1)} \leq -\left(\frac{d-1}{n} \right) \left\| \boldsymbol{q}_{\perp} \right\|.$$

We prove Claim 5.17 in subsection 5.12.1. Using Equation 5.18 in Equation 5.16 we obtain

$$\Delta V(\boldsymbol{q}) \leq n(\lambda + 1) - 2(1 - \lambda) \sum_{i=1}^{n} q_i - 2\lambda \left(\frac{d-1}{n} \right) \left\| \boldsymbol{q}_{\perp} \right\|. \qquad (5.19)$$

Now we compute a lower bound for $\Delta V_{\|}(\boldsymbol{q})$. We obtain

$$\begin{aligned}
&\Delta V_{\|}(\boldsymbol{q}) \\
&= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(\left\| \left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))} \right)_{\|} \right\|^2 - \left\| \boldsymbol{q}_{\|} \right\|^2 \right)
\end{aligned}$$

$$+\sum_{i=1}^{n}\left(1-\mathbb{1}_{\{q_i=0\}}\right)\left(\left\|\left(\boldsymbol{q}-\boldsymbol{e}^{(i)}\right)_{\|}\right\|^2-\left\|\boldsymbol{q}_{\|}\right\|^2\right)$$

$$\overset{(a)}{=}\lambda\sum_{i=1}^{n}\frac{\binom{n-i}{d-1}}{\binom{n}{d}}\left(1+2\sum_{\ell=1}^{n}q_\ell\right)+\frac{1}{n}\sum_{i=1}^{n}\left(1-\mathbb{1}_{\{q_i=0\}}\right)\left(1-2\sum_{j=1}^{n}q_j\right) \tag{5.20}$$

$$=\lambda-2(1-\lambda)\sum_{i=1}^{n}q_i+\frac{1}{n}\sum_{i=1}^{n}\left(1-\mathbb{1}_{\{q_i=0\}}\right)+\frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\left(\sum_{i=1}^{n}q_i\right)$$

$$\overset{(b)}{\geq}-2(1-\lambda)\sum_{i=1}^{n}q_i, \tag{5.21}$$

where (a) holds by the definition of $\boldsymbol{x}_{\|}$ given a vector $\boldsymbol{x}$ in Equation 5.10, and computing the norms; and (b) holds because $\lambda \geq 0$, $1-\mathbb{1}_{\{q_i=0\}} \geq 0$ for all $i \in [n]$, and $\left(\sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\left(\sum_{i=1}^{n}q_i\right) \geq 0$ since every term is nonnegative. Using Equation 5.19 and Equation 5.21 in Equation 5.15, and reorganizing terms we obtain

$$\begin{aligned}\Delta W_{\perp}(\boldsymbol{q}) &\leq \frac{n(\lambda+1)}{2\left\|\boldsymbol{q}_{\perp}\right\|}-\frac{\lambda(d-1)}{n}\\ &\overset{(a)}{\leq}\frac{n}{\left\|\boldsymbol{q}_{\perp}\right\|}-\frac{\lambda_0(d-1)}{n},\end{aligned}$$

where (a) holds because $\lambda \in (\lambda_0, 1)$. Therefore, the first condition of Lemma 2.9 is satisfied with

$$\eta=\frac{\lambda_0(d-1)}{n},\ \text{and}\ \kappa=\frac{n^2}{\lambda_0(d-1)}. \tag{5.22}$$

Now we verify the second condition. From Equation 5.9 observe that if $\boldsymbol{q}, \boldsymbol{q}' \in \mathbb{Z}_{+}^{n}$ are such that $G_{\boldsymbol{q},\boldsymbol{q}'} > 0$, then there are only two options: either $\boldsymbol{q}' = \boldsymbol{q}+\boldsymbol{e}^{(i)}$ or $\boldsymbol{q}' = \boldsymbol{q}-\boldsymbol{e}^{(i)}$ for some $i \in [n]$. Then, by definition of $\boldsymbol{q}_{\perp}$, we have, $\boldsymbol{q}'_{\perp} = \boldsymbol{q}_{\perp} \pm \left(\boldsymbol{e}^{(i)}-\frac{1}{n}\mathbf{1}\right)$. Hence,

$$\nu_{\max} \leq 2. \tag{5.23}$$

To verify the third condition we observe from Equation 5.9 that

$$\overline{G} = n\left(\lambda + 1\right), \tag{5.24}$$

where the maximum is attained when none of the queues is empty. Finally, observe that $G_{\max} \leq \overline{G}$ by definition, and the upper bound is indeed attained when $\boldsymbol{q} = \boldsymbol{q}_{\|}$. Therefore,

$$G_{\max} = n\left(\lambda + 1\right). \tag{5.25}$$

We verified that all the conditions are satisfied. Then, using Equation 5.22, Equation 5.23 and Equation 5.25 in Equation 2.4 from Lemma 2.9, we obtain that for any positive integer j

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^{j}\right] &\leq \left(\frac{2n^2}{\lambda_0(d-1)}\right)^{j} + \left(\frac{64n^2 + 8\lambda_0(d-1)}{\lambda_0(d-1)}\right)^{j} j! \\
&\overset{(a)}{\leq} C\left(\frac{n^2}{d-1}\right)^{j} j! \\
&\overset{(b)}{\leq} C\left(\frac{n^2}{d-1}\right)^{j} j^{j+\frac{1}{2}},
\end{aligned}$$

where (a) holds for some constant $C \geq 64$ which is independent of λ, d and n; and (b) holds by Stirling's approximation and because $e^{1-j} \leq 1$. This proves Equation 5.11.

From Equation 5.11, observe that $C > 1$. Then, $C^{\frac{1}{j}} \leq C$. Also, $j^{\frac{1}{2j}}$ is maximized at $j = e$, so $j^{\frac{1}{2j}} \leq \exp\left(\frac{1}{2e}\right)$. This completes the proof of Equation 5.12.

To prove Equation 5.13 we use Lemma 2.10. Then, replacing $d = cn^{\beta}$ we obtain the results. □

5.8.3 Proof of Equation 2.5 for a load balancing system in continuous time

.

Proof of Equation 2.5 for load balancing system in continuous time. First observe that if $g(x)$

is a differentiable concave function on $\mathbb{R}_+$, we have that for any $x, y \in \mathbb{R}_+$

$$g(x) - g(y) \leq g'(y)(x - y). \tag{5.26}$$

Now, observe that $W_\perp(\boldsymbol{q}) = \|\boldsymbol{q}_\perp\| = \sqrt{\|\boldsymbol{q}_\perp\|^2}$ and $g(x) = \sqrt{x}$ is a concave function. Therefore, by definition of drift in Definition 2.3, and the generator matrix in Equation 5.9, we have

$$\begin{aligned}
&\Delta W_\perp(\boldsymbol{q}) \\
&= \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(W_\perp\left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right) - W_\perp\left(\boldsymbol{q}\right)\right) \\
&\quad + \sum_{i=1}^{n} \left(1 - \mathbb{1}_{\{q_i=0\}}\right) \left(W_\perp\left(\boldsymbol{q} + \boldsymbol{e}^{(i)}\right) - W_\perp(\boldsymbol{q})\right) \\
&\overset{(a)}{\leq} \lambda n \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(\frac{\left\|\left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right)_\perp\right\|^2 - \|\boldsymbol{q}_\perp\|^2}{2\|\boldsymbol{q}_\perp\|}\right) \\
&\quad + \sum_{i=1}^{n} \left(1 - \mathbb{1}_{\{q_i=0\}}\right) \left(\frac{\left\|\left(\boldsymbol{q} + \boldsymbol{e}^{(i)}\right)_\perp\right\|^2 - \|\boldsymbol{q}_\perp\|^2}{2\|\boldsymbol{q}_\perp\|}\right) \\
&\overset{(b)}{=} \frac{\lambda n}{2\|\boldsymbol{q}_\perp\|} \sum_{i=1}^{n} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(V\left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right) - V(\boldsymbol{q}) - \left(V_\parallel\left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right) - V_\parallel(\boldsymbol{q})\right)\right) \\
&\quad + \sum_{i=1}^{n} \left(\frac{1 - \mathbb{1}_{\{q_i=0\}}}{2\|\boldsymbol{q}_\perp\|}\right) \left(V\left(\boldsymbol{q} + \boldsymbol{e}^{(i)}\right) - V(\boldsymbol{q}) - \left(V_\parallel\left(\boldsymbol{q} + \boldsymbol{e}^{(i)}\right) - V_\parallel(\boldsymbol{q})\right)\right) \\
&\overset{(c)}{=} \frac{1}{2\|\boldsymbol{q}_\perp\|} \left(\Delta V(\boldsymbol{q}) - \Delta V_\parallel(\boldsymbol{q})\right)
\end{aligned}$$

where (a) holds by Equation 5.26 applied in the first and the second term in the following way. In the first term we use $x = \left\|\left(\boldsymbol{q} + \boldsymbol{e}^{(\psi_{\boldsymbol{q}}(i))}\right)_\perp\right\|^2$ and $y = \|\boldsymbol{q}_\perp\|^2$, and in the second term we use $x = \left\|\left(\boldsymbol{q} + \boldsymbol{e}^{(i)}\right)_\perp\right\|^2$ and $y = \|\boldsymbol{q}_\perp\|^2$. Equality (b) holds by the definition of $V(\cdot)$ and $V_\parallel(\cdot)$ in Equation 5.14 and because for any vector $\boldsymbol{x} \in \mathbb{R}^n$, we have $\|\boldsymbol{x}_\perp\|^2 = \|\boldsymbol{x}\|^2 - \|\boldsymbol{x}_\parallel\|^2$; and (c) holds by reorganizing terms and by definition of drift. $\square$

5.9 Transform method: Proof of Theorem 5.10

The first proof of Theorem 5.10 that we present is motivated by the transform method introduced in section 3.3. We use the key ideas and we extend them to use them in the context of the load balancing system modeled in continuous time in the many-server heavy-traffic regime, as described in section 5.7. In this case, the role of the unused service is played by the indicator functions $\mathbb{1}_{\{q_i(t)=0\}}$ for $i \in [n]$, since no job can be served from the i^{th} queue if $q_i(t) = 0$. In fact, this indicator function satisfies the key property $\mathbb{1}_{\{q_i=0\}} q_i = 0$ with probability 1 for all $i \in [n]$, which written in 'exponential form' yields $\mathbb{1}_{\{q_i=0\}} = \mathbb{1}_{\{q_i=0\}} \exp(\tilde{\theta} q_i)$, where $\tilde{\theta}$ is any real number. We present the complete proof below.

5.9.1 Proof of Theorem 5.10 using the transform method

In this proof we use one-sided Laplace transform to illustrate a different transform that falls in the scope of the transform method introduced in chapter 3. The exponential equation required for Step 1 is accomplished by the following lemma.

Lemma 5.18. *Consider a load balancing system operating under power-of-d with $d = cn^{\beta}$, as described in Theorem 5.10. Then, if c and β are such that $d \geq 2$, there exists $n_0^* \in \mathbb{Z}_+$ such that for any θ satisfying*

$$|\theta| < \frac{1}{8\lceil \alpha - 1 \rceil \lceil \log(n_0^*) \rceil (n_0^*)^{1-\alpha}} \log\left(1 + \frac{\lambda_0 (c(n_0^*)^{\beta} - 1)}{8 n_0^*}\right),$$

we have

$$\left(\exp\left(\theta n^{-\alpha} q_{\Sigma}^{(n)}(t)\right) - 1\right) \left(\sum_{i=1}^{n} \mathbb{1}_{\left\{q_i^{(n)}(t)=0\right\}}\right) = \phi(\boldsymbol{q}^{(n)}(t), n),$$

where $q_{\Sigma}^{(n)}(t) \stackrel{\triangle}{=} \sum_{i=1}^{n} q_i^{(n)}(t)$. The function $\phi(\boldsymbol{q}^{(n)}(t), n)$ satisfies the following property. If $\alpha + \beta > 3$, then $\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}^{(n)}, n)\right]$ is of order $o(n^{1-\alpha})$, where the expectation is taken with respect to the stationary distribution of the queue lengths.

The proof of Lemma 5.18 is presented in section 5.13. Observe that Proposition 5.14 plays a key role in bounding $\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}^{(n)}, n)\right]$. Now we prove the theorem.

Proof of Theorem 5.10 using Transform method. We omit the dependence on n of the variables, and we work with d instead of cn^{β} for ease of exposition. This proof is based on the use of the test function $V_{\exp}(\boldsymbol{q}) \stackrel{\triangle}{=} \exp\left(\theta n^{-\alpha} q_{\Sigma}\right)$, where $\theta < 0$. Using the definition of drift from Definition 2.3, we obtain that for any $\boldsymbol{q} \in \mathbb{Z}_{+}^{n}$

$$
\begin{aligned}
&\Delta V_{\exp}(\boldsymbol{q}) \\
&= \exp\left(\theta n^{-\alpha} q_{\Sigma}\right) \left(\sum_{i=1}^{n} \lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \left(\exp\left(\theta n^{-\alpha}\right) - 1\right) + n\left(\exp\left(-\theta n^{-\alpha}\right) - 1\right) \right) \\
&\quad - \left(\sum_{i=1}^{n} \mathbb{1}_{\{q_i = 0\}} \right) \exp\left(\theta n^{-\alpha} q_{\Sigma}\right) \left(\exp\left(-\theta n^{-\alpha}\right) - 1\right) \\
&\stackrel{(a)}{=} \left(\exp\left(-\theta n^{-\alpha}\right) - 1\right) \exp\left(\theta n^{-\alpha} \overline{q}_{\Sigma}\right) \left(n\left(1 - \lambda \exp\left(\theta n^{-\alpha}\right)\right) - \sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i = 0\}} \right) \\
&\stackrel{(b)}{=} \left(\exp\left(-\theta n^{-\alpha}\right) - 1\right) \left(\exp\left(\theta n^{-\alpha} \overline{q}_{\Sigma}\right) n\left(1 - \lambda \exp\left(\theta n^{-\alpha}\right)\right) - \sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i = 0\}} - \phi(\boldsymbol{q}, n) \right),
\end{aligned}
$$

where (a) holds because $\sum_{i=1}^{n} \binom{n-i}{d-1} = \binom{n}{d}$ and rearranging terms; and (b) holds by Lemma 5.18.

Now we set to zero the drift of $V_{\exp}(\boldsymbol{q})$ in steady state. Observe that, since $\theta < 0$, we have $\mathbb{E}\left[\exp\left(\theta n^{-\alpha} \overline{q}_{\Sigma}\right)\right] \leq 1$. Then, we know $\mathbb{E}\left[\Delta V_{\exp}(\overline{\boldsymbol{q}})\right] = 0$. Therefore, taking expected value with respect to stationary distribution in the expression above, replacing $\lambda = 1 - n^{-\alpha}$, using Lemma 5.15 and rearranging terms we obtain

$$
\mathbb{E}\left[\theta n^{-\alpha} \overline{q}_{\Sigma}\right] = \frac{n^{1-\alpha} + \mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]}{n\left(1 - (1 - n^{-\alpha}) \exp\left(\theta n^{-\alpha}\right)\right)}. \tag{5.27}
$$

This completes Step 1. Observe that Equation 5.27 gives an expression for the one-sided Laplace transform of $n^{-\alpha} \overline{q}_{\Sigma}$ that is valid for all n. However, the numerator depends on $\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]$, which is not an explicit expression.

Now we move to the second step, where the goal is to take the many-server heavy-traffic limit. The fraction in Equation 5.27 is of the form $\frac{0}{0}$ in the limit as $n \to \infty$, so we take Taylor expansion of the exponential function in the denominator. Expanding up to second order and canceling the factor $n^{1-\alpha}$ from the numerator and the denominator we obtain

$$\mathbb{E}\left[\exp\left(\theta n^{-\alpha}\overline{q}_{\Sigma}\right)\right] = \frac{1 + n^{\alpha-1}\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]}{1 - \theta + O(n^{-\alpha})}.$$

Finally, taking the limit as $n \to \infty$ we obtain

$$\lim_{n\to\infty} \mathbb{E}\left[\exp\left(\theta n^{-\alpha}\overline{q}_{\Sigma}\right)\right] = \frac{1}{1-\theta},$$

which is the one-sided Laplace transform of an exponential random variable with mean 1. This completes the proof. □

Observe that Equation 5.27 is valid for all n. Hence, it can be used to obtain an error bound between $\mathbb{E}\left[\exp\left(\theta n^{-\alpha}\overline{q}_{\Sigma}\right)\right]$ and $\frac{1}{1-\theta}$, which is the limiting one-sided Laplace transform. We do not perform this step for brevity.

5.10 Rate of convergence in Wasserstein's distance

The proof we provide in this section is based on bounding the Wasserstein's distance between the scaled total queue length and an exponential random variable, similarly to section 5.4. Specifically, in the rest of this section we prove the following theorem.

Theorem 5.19. *Consider a load balancing system operating under power-of-d choices with $d = cn^{\beta}$, as described in Theorem 5.10. Let Υ be an exponential random variable with mean 1. Then,*

$$d_W\left((1-\lambda)\overline{q}_{\Sigma}, \Upsilon\right) \leq \overline{C}e\left(\frac{n^3(1-\lambda)}{cn^{\beta}-1}\right)\left\lceil \log\left(\tfrac{1}{n(1-\lambda)}\right)\right\rceil + \frac{5}{3}(1-\lambda), \tag{5.28}$$

where $\overline{C}$ is the constant from Proposition 5.14.

Note that if we let $\lambda = 1 - n^{-\alpha}$ and $\alpha + \beta > 3$, the right-hand side of Equation 5.28 converges to zero as $n \to \infty$. Therefore, we can prove Theorem 5.10 as a direct consequence of Theorem 5.19. Now we prove Theorem 5.19. We start with a result presented in [93, Theorem 5.2 part 2].

Lemma 5.20. *Let Y be a random variable with $\mathbb{E}[Y] < \infty$, and let Υ be an exponential random variable with mean 1. Define*

$$\mathcal{F}_W \stackrel{\triangle}{=} \{g : \mathbb{R} \to \mathbb{R} \text{ such that } g(0) = 0,\ \|g'\| \leq 1,\ \|g''\| \leq 2\}.$$

Then,

$$d_W(Y, \Upsilon) \leq \sup_{g \in \mathcal{F}_W} |\mathbb{E}[g'(Y) - g(Y)]|.$$

Now we prove the theorem.

Proof of Theorem 5.19. Similarly to all our previous proofs, for ease of exposition we omit the dependence on n of the variables and we use d instead of cn^β. We use Lemma 5.20 with $Y = (1-\lambda)\overline{q}_\Sigma$. Let f be a differentiable function such that $g = f' \in \mathcal{F}_W$. By assuming differentiability we do not lose generality, for the following reason. Observe that $f' \in \mathcal{F}_W$ implies that $f \in \text{Lip}(1)$ and, hence, it implies that f is integrable. Therefore, if $f' \in \mathcal{F}_W$, then f is well defined [94, Theorem 7.2].

By definition of drift, for any vector $\boldsymbol{q} \in \mathbb{Z}_+^n$, we have

$$\begin{aligned}
&\Delta f\left((1-\lambda)q_\Sigma\right) \\
&= \sum_{i=1}^{n} \lambda n \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \Big(f\big((1-\lambda)q_\Sigma + 1 - \lambda\big) - f\big((1-\lambda)q_\Sigma\big) \Big) \\
&\quad + \sum_{i=1}^{n} \big(1 - \mathbb{1}_{\{q_i=0\}}\big) \Big(f\big((1-\lambda)q_\Sigma - 1 + \lambda\big) - f\big((1-\lambda)q_\Sigma\big) \Big)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \lambda n\Big(f\big((1-\lambda)q_\Sigma+1-\lambda\big)-f\big((1-\lambda)q_\Sigma\big)\Big)\\
&\quad+\left(n-\sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\Big(f\big((1-\lambda)q_\Sigma-(1-\lambda)\big)-f\big((1-\lambda)q_\Sigma\big)\Big)\\
&\overset{(b)}{=} \lambda n\left((1-\lambda)f'\big((1-\lambda)q_\Sigma\big)+\frac{(1-\lambda)^2}{2}f''\big((1-\lambda)q_\Sigma\big)+\frac{(1-\lambda)^3}{6}f'''\left(\xi_1\right)\right)\\
&\quad+n\left(-(1-\lambda)f'\big((1-\lambda)q_\Sigma\big)+\frac{(1-\lambda)^2}{2}f''\big((1-\lambda)q_\Sigma\big)-\frac{(1-\lambda)^3}{6}f'''(\xi_2)\right)\\
&\quad+\left(\sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\left((1-\lambda)f'\big((1-\lambda)q_\Sigma\big)-\frac{(1-\lambda)^2}{2}f''\big((1-\lambda)q_\Sigma\big)\right)\\
&\quad+\left(\sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\frac{(1-\lambda)^3}{6}f'''(\xi_3)
\end{aligned}
$$

where ξ_1 is between $(1-\lambda)q_\Sigma$ and $(1-\lambda)\left(q_\Sigma+1\right)$, and ξ_2,ξ_3 are between $(1-\lambda)$ and $(1-\lambda)\left(q_\Sigma-1\right)$. Here, (a) holds because $\sum_{i=1}^{n}\binom{n-i}{d-1}=\binom{n}{d}$; and (b) holds by taking Taylor approximation.

Since $f'\in\mathcal{F}_W$, we know that f is integrable. Then, we can set its drift to zero in steady state. Taking expectation with respect to stationary distribution and reorganizing terms we obtain

$$
\begin{aligned}
&\mathbb{E}\left[f'\big((1-\lambda)\overline{q}_\Sigma\big)\right]\\
&=\frac{1}{n(1-\lambda)}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)f'\big((1-\lambda)\overline{q}_\Sigma\big)\right]+\left(\frac{1+\lambda}{2}\right)\mathbb{E}\left[f''\big((1-\lambda)\overline{q}_\Sigma\big)\right]\\
&\quad-\frac{1}{2n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)f''\big((1-\lambda)\overline{q}_\Sigma\big)\right]+\frac{\lambda(1-\lambda)}{6}\mathbb{E}\left[f'''(\xi_1)\right]\\
&\quad-\left(\frac{1-\lambda}{6}\right)\mathbb{E}\left[f'''\left(\xi_2\right)\right]+\left(\frac{1-\lambda}{6n}\right)\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)f'''(\xi_3)\right].
\end{aligned}
$$

Using the last expression and the triangle inequality we have

$$
\begin{aligned}
&\left|\mathbb{E}\left[f'\big((1-\lambda)\overline{q}_\Sigma\big)-f''\big((1-\lambda)\overline{q}_\Sigma\big)\right]\right| \qquad (5.29)\\
&\leq\frac{1}{n(1-\lambda)}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left|f'\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right]+\left|\frac{\lambda-1}{2}\right|\mathbb{E}\left[\left|f''\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right]
\end{aligned}
$$

$$+\frac{1}{2n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left|f''\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right]+\frac{\lambda(1-\lambda)}{6}\mathbb{E}\left[|f'''(\xi_1)|\right]$$
$$+\left(\frac{1-\lambda}{6}\right)\mathbb{E}\left[|f'''(\xi_2)|\right]+\left(\frac{1-\lambda}{6n}\right)\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)|f'''(\xi_3)|\right].$$

We bound term by term. For the first term we expand $f'\big((1-\lambda)\overline{q}_\Sigma\big)$ in Taylor series up to first order, around 0. Since $f' \in \mathcal{F}_W$, we know that $f'(0) = 0$. Then, $f'\big(1-\lambda)\overline{q}_\Sigma\big) = (1-\lambda)\overline{q}_\Sigma f''(\xi_4)$, where $|\xi_4| \in (0, \overline{q}_\Sigma)$. Therefore, we obtain

$$\begin{aligned}
\frac{1}{n(1-\lambda)}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left|f'\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right] &= \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\overline{q}_\Sigma\left|f''(\xi_4)\right|\right] \\
&\overset{(a)}{\leq} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\overline{q}_\Sigma\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\left(\frac{\overline{q}_\Sigma}{n}-\overline{q}_i\right)\right] \\
&\overset{(c)}{=} -\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\overline{q}_{\perp i}\right] \\
&\overset{(d)}{\leq} \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right]^{1-\frac{1}{j}}\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp\|_j^j\right]^{\frac{1}{j}} \\
&\overset{(e)}{\leq} \overline{C}(1-\lambda)^{1-\frac{1}{j}}j\left(\frac{n^{3-\frac{1}{j}}}{d-1}\right),
\end{aligned}$$

where $j > 1$. Here, (a) holds because $f' \in \mathcal{F}_W$ and, hence, $|f''(\xi_4)| \leq 1$; (b) holds because $\mathbb{1}_{\{\overline{q}_i=0\}}\overline{q}_i = 0$ for all $i \in [n]$; (c) holds by definition of $\overline{\boldsymbol{q}}_\perp$ in Equation 5.10; (d) holds by Hölder's inequality; and (e) holds by Lemma 5.15, because for any $j \geq 2$ the j^{th} norm lower bounds the Euclidean norm, and by Proposition 5.14. Taking $j = \left\lceil \log\left(\frac{1}{n(1-\lambda)}\right)\right\rceil$ we obtain

$$\frac{1}{n(1-\lambda)}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left|f'\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right] \leq \overline{C}e\left(\frac{n^3(1-\lambda)}{d-1}\right)\left\lceil \log\left(\frac{1}{n(1-\lambda)}\right)\right\rceil. \tag{5.30}$$

For the second term, since $f' \in \mathcal{F}_W$ we obtain

$$\left(\frac{1-\lambda}{2}\right) \mathbb{E}\left[\left|f''\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right] \leq \frac{1-\lambda}{2}. \tag{5.31}$$

For the third term we obtain

$$\begin{aligned}\frac{1}{2n}\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right)\left|f''\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right] &\overset{(a)}{\leq} \frac{1}{2n}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right] \\ &\overset{(b)}{=} \frac{1-\lambda}{2},\end{aligned} \tag{5.32}$$

where (a) holds because $f' \in \mathcal{F}_W$; and (b) holds by Lemma 5.15.

For the fourth term, since $f' \in \mathcal{F}_W$, we obtain

$$\frac{\lambda(1-\lambda)}{6}\mathbb{E}\left[|f'''(\xi_1)|\right] \leq \frac{\lambda(1-\lambda)}{3}. \tag{5.33}$$

Similarly, for the fifth term we have

$$\left(\frac{1-\lambda}{6}\right)\mathbb{E}\left[|f'''(\xi_2)|\right] \leq \frac{1-\lambda}{3}. \tag{5.34}$$

For the last term, we obtain

$$\begin{aligned}\left(\frac{1-\lambda}{6n}\right)\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right)|f'''(\xi_3)|\right] &\overset{(a)}{\leq} \left(\frac{1-\lambda}{3n}\right)\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right)\right] \\ &\overset{(b)}{=} \frac{(1-\lambda)^2}{3},\end{aligned} \tag{5.35}$$

where (a) holds because $f' \in \mathcal{F}_W$; and (b) holds by Lemma 5.15.

Using Equation 5.30–Equation 5.35 in Equation 5.29, and rearranging terms we obtain

$$\begin{aligned}&\mathbb{E}\left[\left|f'\big((1-\lambda)\overline{q}_\Sigma\big) - f''\big((1-\lambda)\overline{q}_\Sigma\big)\right|\right] \\ &\leq \overline{C}e\left(\frac{n^3(1-\lambda)}{d-1}\right)\left\lceil \log\left(\tfrac{1}{n(1-\lambda)}\right)\right\rceil + \frac{5}{3}(1-\lambda).\end{aligned}$$

Replacing $d = cn^\beta$, we complete the proof. $\square$

5.11 Rate of convergence of the first moment

In Theorem 5.10 we showed convergence in distribution of the average queue length scaled by $n^{1-\alpha}$ (or, equivalently, the total queue length scaled by $n^{-\alpha}$). However, convergence in distribution is not a sufficient condition to conclude convergence of the expected value. In other words, in Theorem 5.10 we showed $n^{-\alpha}\overline{q}_\Sigma^{(n)} \Rightarrow \Upsilon$, where Υ is an exponential random variable with mean 1. However, from this statement we cannot directly conclude that $\lim_{n\to\infty} \mathbb{E}\left[n^{-\alpha}\overline{q}_\Sigma^{(n)}\right] = 1$. In this section we show that the last result holds using the drift method [34, 14, 15, 22]. We first state the result formally.

Theorem 5.21. *Consider a sequence of load balancing systems operating under power-of-d with $d = cn^\beta$, parametrized by n as described in section 5.7. If c and β are such that $cn^\beta \geq 2$, then*

$$\left|\mathbb{E}\left[\sum_{i=1}^{n} \overline{q}_i\right] - n^\alpha\right| \leq 1 + \left(\frac{\overline{C}e}{c}\right)\lceil\alpha - 1\rceil\lceil\log(n)\rceil\left(\frac{cn^\beta}{cn^\beta - 1}\right)n^{3-\beta}, \tag{5.36}$$

where $\overline{C}$ is the constant from Proposition 5.14. Additionally, if $\alpha + \beta > 3$, then

$$\lim_{n\to\infty} n^{-\alpha}\mathbb{E}\left[\overline{q}_\Sigma^{(n)}\right] = 1.$$

Note that the second part of the theorem is an immediate consequence of the error bound because, after multiplying everything by $n^{-\alpha}$, the right-hand side of Equation 5.36 converges to zero as $n \to \infty$.

Similarly to Theorem 5.10, the case of power-of-d choices with constant d and JSQ are immediate consequences of Theorem 5.21. We formally state the results below.

Corollary 5.22. *Consider a sequence of load balancing systems operating under power-of-d choices with constant d, parametrized by n as described in section 5.7. If $d \geq 2$ and*

$\alpha > 3$*, then*

$$\lim_{n\to\infty} n^{-\alpha}\mathbb{E}\left[\overline{q}_{\Sigma}^{(n)}\right] = 1.$$

The proof of Corollary 5.22 holds easily after letting $\beta = 0$ in Theorem 5.21. Now we present the formal result for JSQ routing.

Corollary 5.23. *Consider a sequence of load balancing systems operating under JSQ, parametrized by n as described in section 5.7. If $\alpha > 2$, then*

$$\lim_{n\to\infty} n^{-\alpha}\mathbb{E}\left[\overline{q}_{\Sigma}^{(n)}\right] = 1.$$

The proof of Corollary 5.23 holds after realizing that JSQ is equivalent to power-of-d choices with $d = n$. Hence, it suffices to replace $c = \beta = 1$ in Theorem 5.21.

In the rest of this section we prove Theorem 5.21 using the drift method. Recall that in the drift method there are two main steps. First, one shows SSC (which we did in Proposition 5.14), and secondly, one sets to zero the drift of $V_{\|}(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$ in steady state (provided that its expectation is finite).

Proof of Theorem 5.21. Similarly to our previous proofs, we omit the dependence on n of the variables and we work with d instead of cn^{β} for ease of exposition. We start computing the drift of $V_{\|}(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$. From Equation 5.20 in the proof of SSC, and since $\sum_{i=1}^{n}\binom{n-i}{d-1} = \binom{n}{d}$, we obtain

$$\Delta V_{\|}(\boldsymbol{q}) = \lambda\left(1 + 2\sum_{i=1}^{n} q_i\right) + \frac{1}{n}\left(n - \sum_{i=1}^{n}\mathbb{1}_{\{q_i=0\}}\right)\left(1 - 2\sum_{i=1}^{n} q_i\right).$$

Now we set to zero the drift of $V_{\|}(\boldsymbol{q})$. We skip the proof of $\mathbb{E}\left[V_{\|}(\overline{\boldsymbol{q}})\right] < \infty$ for ease of exposition. Taking expectation with respect to the stationary distribution, replacing $\lambda = 1 - n^{-\alpha}$, using Lemma 5.15 to replace $\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right] = n^{1-\alpha}$ and reorganizing terms

we obtain:

$$n^{-\alpha}\mathbb{E}\left[\sum_{i=1}^{n}\overline{q}_i\right] = 1 - n^{-\alpha} + \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left(\sum_{\ell=1}^{n}\overline{q}_\ell\right)\right]. \tag{5.37}$$

We bound the last term of Equation 5.37 using SSC. First, note $\mathbb{1}_{\{\overline{q}_i=0\}}\overline{q}_i = 0$ with probability 1 for all $i \in [n]$. Then,

$$\begin{aligned}\frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left(\sum_{\ell=1}^{n}\overline{q}_\ell\right) &= \sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\left(\frac{1}{n}\sum_{\ell=1}^{n}\overline{q}_\ell - \overline{q}_i\right)\\ &\overset{(a)}{=} -\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\overline{q}_{\perp i},\end{aligned}$$

where $\overline{q}_{\perp i}$ is the i^{th} element of $\overline{\boldsymbol{q}}_\perp$ and (a) holds by the definition of $\overline{\boldsymbol{q}}_\perp$ in Equation 5.20. Then,

$$\begin{aligned}\left|\left(\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right)\left(\sum_{\ell=1}^{n}\overline{q}_\ell\right)\right| &= \left|\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\overline{q}_{\perp i}\right]\right|\\ &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{\{\overline{q}_i=0\}}\right]^{1-\frac{1}{j}}\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp\|^j\right]^{\frac{1}{j}}\\ &\overset{(b)}{\leq} n^{(1-\alpha)\left(1-\frac{1}{j}\right)}\overline{C}j\left(\frac{n^2}{cn^\beta - 1}\right)\\ &= n^{(1-\alpha)\left(1-\frac{1}{j}\right)}\frac{\overline{C}}{c}jn^{2-\beta}\left(\frac{cn^\beta}{cn^\beta - 1}\right)\\ &= \left(\frac{\overline{C}}{c}\right)jn^{3-\alpha-\beta}n^{\frac{\alpha-1}{j}}\left(\frac{cn^\beta}{cn^\beta - 1}\right)\\ &\overset{(c)}{\leq} \left(\frac{\overline{C}e}{c}\right)\lceil\alpha-1\rceil\lceil\log(n)\rceil\left(\frac{cn^\beta}{cn^\beta - 1}\right)n^{3-\alpha-\beta},\end{aligned}$$

where j is a positive integer. Here, (a) holds by Hölder's inequality; (b) holds by Lemma 5.15 and by Proposition 5.14 for $j \geq 2$ because of the inequalities of norms; and (c) holds by setting $j = \lceil\alpha - 1\rceil\lceil\log(n)\rceil$ and because $n^{\frac{\alpha-1}{\lceil\alpha-1\rceil\lceil\log(n)\rceil}} \leq e$. Using this result in Equation 5.37

we obtain

$$\left| n^{-\alpha}\mathbb{E}\left[\sum_{i=1}^{n}\overline{q}_i\right] - 1\right| \leq n^{-\alpha} + \left(\frac{\overline{C}e}{c}\right)\lceil\alpha-1\rceil\lceil\log(n)\rceil\left(\frac{cn^{\beta}}{cn^{\beta}-1}\right)n^{3-\alpha-\beta}.$$

This proves the theorem. □

5.12 Details of proofs of Section section 5.8

5.12.1 Proof of Claim 5.17

Proof. We first prove that the first term of Claim 5.17 satisfies

$$\left(\frac{d-1}{n}\right)\sum_{i=1}^{n}\left(q_{(1)}-q_i\right) \leq -\left(\frac{d-1}{n}\right)\left\|\boldsymbol{q}_{\perp}\right\|. \tag{5.38}$$

We have:

$$\begin{aligned}\left(\frac{d-1}{n}\right)\sum_{i=1}^{n}\left(q_{(1)}-q_i\right) &\overset{(a)}{=} -\left(\frac{d-1}{n}\right)\sum_{i=1}^{n}\left|q_i-q_{(1)}\right| \\ &= -\left(\frac{d-1}{n}\right)\left\|\boldsymbol{q}-q_{(1)}\mathbf{1}\right\|_1 \\ &\overset{(b)}{\leq} -\left(\frac{d-1}{n}\right)\left\|\boldsymbol{q}-q_{(1)}\mathbf{1}\right\| \\ &\overset{(c)}{\leq} -\left(\frac{d-1}{n}\right)\left\|\boldsymbol{q}_{\perp}\right\|,\end{aligned}$$

where (a) holds because $q_{(1)} = \min_{i\in[n]} q_i$; (b) holds because norm-1 upper bounds the Euclidean norm; and (c) holds because, by definition of projection, the function $g(x) = \|\boldsymbol{q} - x\mathbf{1}\|$ is minimized at $x = \frac{1}{n}\sum_{i=1}^{n} q_i$, and by definition of $\boldsymbol{q}_{\|}$ and $\boldsymbol{q}_{\perp}$ in Equation 5.10.

Now we only need to show that

$$\sum_{i=1}^{n}\left(\gamma_i - \frac{n-d+1}{n}\right)q_{(i)} - (d-1)q_{(1)} \leq 0.$$

As shown in [71, Section A.2.a], it suffices to show that

(i) $\sum_{i=1}^{n} \gamma_i - (d-1) = n - d + 1$, and

(ii) $\sum_{i=1}^{j} \gamma_i - (d-1) \geq \frac{j\,(n-d+1)}{n}\ \forall j \in [n-1]$.

Indeed, we have

$$\sum_{i=1}^{n} \gamma_i - (d-1) = \sum_{i=1}^{n} \frac{n\binom{n-i}{d-1}}{\binom{n}{d}} - (d-1) = n - d + 1.$$

To prove the second condition observe

$$\sum_{i=1}^{\ell} \gamma_i - (d-1) = n - d + 1 - \sum_{i=\ell+1}^{n} \gamma_i.$$

Then, it suffices to show that for any $\ell \in [n]$ with $\ell \geq 2$ we have

$$\sum_{i=\ell}^{n} \gamma_i \leq \frac{(n-\ell+1)(n-d+1)}{n}. \tag{5.39}$$

Since $\gamma_i = 0$ for any $i > n - d + 1$, the condition is trivially satisfied for $\ell \geq n - d + 2$. Now, if $\ell \leq n - d + 1$ we have

$$\begin{aligned}
\sum_{i=\ell}^{n} \gamma_i &= \sum_{i=\ell}^{n-d+1} \frac{n\binom{n-i}{d-1}}{\binom{n}{d}} \\
&= (n-\ell+1)\frac{\binom{n-j}{d-1}}{\binom{n-1}{d-1}} \\
&\overset{(a)}{\leq} (n-\ell+1)\frac{\binom{n-2}{d-1}}{\binom{n-1}{d-1}} \\
&= (n-\ell+1)\left(\frac{n-d}{n-1}\right)
\end{aligned}$$

where (a) holds because $\binom{n-\ell}{d-1}$ is decreasing in ℓ and $\ell \geq 2$. Then, Equation 5.39 is satisfied

if

$$\frac{n-d+1}{n} \geq \frac{n-d}{n-1},$$

which is satisfied because $d \geq 2$ by definition. This completes the proof. □

5.13 Proof of Lemma 5.18

Proof of Lemma 5.18. We omit the dependence on n and t of the variables, for ease of exposition. By definition of indicator function, for any $i \in [n]$ we have

$$\begin{aligned}\mathbb{1}_{\{q_i=0\}} \exp\left(\theta n^{-\alpha} q_\Sigma\right) &= \mathbb{1}_{\{q_i=0\}} \exp\left(-\theta n^{1-\alpha} q_i\right) \exp\left(\theta n^{-\alpha} q_\Sigma\right) \\ &\overset{(a)}{=} \mathbb{1}_{\{q_i=0\}} + \mathbb{1}_{\{q_i=0\}} \left(\exp\left(-\theta n^{1-\alpha} q_{\perp i}\right) - 1\right),\end{aligned}$$

where $q_{\perp i}$ is the i^{th} component of $\boldsymbol{q}_\perp$. Here, (a) holds by definition of $\boldsymbol{q}_\perp$ according to Equation 5.10. Then, it suffices to show that

$$\phi(\boldsymbol{q}, n) \overset{\triangle}{=} \sum_{i=1}^{n} \mathbb{1}_{\{q_i=0\}} \left(\exp\left(-\theta n^{1-\alpha} q_{\perp i}\right) - 1\right)$$

satisfies

$$\lim_{n\to\infty} \frac{1}{n^{1-\alpha}} \mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right] = 0.$$

We bound $|\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]|$ and we show that the bound goes to zero for the given values of θ. We have

$$\begin{aligned}\left|\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]\right| &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}} \left|\exp\left(-\theta n^{1-\alpha} \overline{q}_{\perp i}\right) - 1\right|\right] \\ &\overset{(b)}{\leq} |\theta| n^{1-\alpha} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}} |\overline{q}_{\perp i}| \exp\left(|\theta| n^{1-\alpha} |\overline{q}_{\perp i}|\right)\right]\end{aligned}$$

$$\overset{(c)}{\leq} |\theta| n^{1-\alpha} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\overline{q}_i=0\}}\right]^{1-\frac{1}{j}} \mathbb{E}\left[\sum_{i=1}^{n} |\overline{q}_{\perp i}|^j \exp\left(|\theta| n^{1-\alpha} j |\overline{q}_{\perp i}|\right)\right]^{\frac{1}{j}}$$

$$\overset{(d)}{=} |\theta| n^{(1-\alpha)\left(2-\frac{1}{j}\right)} \mathbb{E}\left[\sum_{i=1}^{n} |\overline{q}_{\perp i}|^j \exp\left(|\theta| n^{1-\alpha} j |\overline{q}_{\perp i}|\right)\right]^{\frac{1}{j}}, \tag{5.40}$$

where $j > 1$. Here, (a) holds by triangle inequality; (b) holds because $|e^x - 1| \leq |x| e^{|x|}$ for all $x \in \mathbb{R}$; (c) hods by Hölder's inequality; and (d) holds by Lemma 5.15.

Now we bound the expectation in section 5.40 using Cauchy-Schwarz inequality and Proposition 5.14. For $j \geq 2$ we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} |\overline{q}_{\perp i}|^j \exp\left(|\theta| n^{1-\alpha} j |\overline{q}_{\perp i}|\right)\right]^{\frac{1}{j}} \\
&\overset{(a)}{\leq} \mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|_j^j \exp\left(|\theta| n^{1-\alpha} j \|\overline{\boldsymbol{q}}_{\perp}\|\right)\right]^{\frac{1}{j}} \\
&\overset{(b)}{\leq} \mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|^j \exp\left(|\theta| n^{1-\alpha} j \|\overline{\boldsymbol{q}}_{\perp}\|\right)\right]^{\frac{1}{j}} \\
&\overset{(c)}{\leq} \mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}\|^{2j}\right]^{\frac{1}{2j}} \mathbb{E}\left[\exp\left(|\theta| n^{1-\alpha} 2j \|\overline{\boldsymbol{q}}_{\perp}\|\right)\right]^{\frac{1}{2j}} \\
&\overset{(d)}{\leq} 2\overline{C} j \left(\frac{n^2}{cn^\beta - 1}\right) \exp\left(\tfrac{|\theta| n^{3-\alpha}}{\lambda_0(cn^\beta - 1)}\right) \left(\frac{\lambda_0(cn^\beta - 1)}{\lambda_0(cn^\beta - 1) + 2n\left(1 - \exp\left(8j|\theta| n^{1-\alpha}\right)\right)}\right)^{\frac{1}{2j}},
\end{aligned}$$

where (a) holds using that $|\overline{q}_{\perp i}| \leq \|\overline{\boldsymbol{q}}_{\perp}\|$ in the exponent and by definition of the j-norm; (b) holds because the j-norm is smaller than the Euclidean norm for all $j \geq 2$; (c) holds by Cauchy-Schwarz inequality; and (d) holds by Proposition 5.14 for θ satisfying

$$|\theta| \leq \frac{1}{8jn^{1-\alpha}} \log\left(1 + \tfrac{\lambda_0(cn^\beta - 1)}{8n}\right).$$

Taking $j \overset{\triangle}{=} \lceil \alpha - 1 \rceil \lceil \log(n) \rceil$ we obtain

$$|\mathbb{E}\left[\phi(\overline{\boldsymbol{q}}, n)\right]| \leq |\theta| n^{2(1-\alpha)} \varphi(n),$$

where

$$\varphi(n) \triangleq n^{\frac{\alpha-1}{\lceil\alpha-1\rceil\lceil\log(n)\rceil}} \left(\frac{2\overline{C}}{cn^{\beta}-1}\right) n^2 \lceil\alpha-1\rceil\lceil\log(n)\rceil \exp\left(\frac{|\theta|n^{3-\alpha}}{\lambda_0(cn^{\beta}-1)}\right) \times \left(\frac{\lambda_0(cn^{\beta}-1)}{\lambda_0(cn^{\beta}-1)+2n\left(1-\exp\left(8|\theta|n^{1-\alpha}\lceil\alpha-1\rceil\lceil\log(n)\rceil\right)\right)}\right)^{\frac{1}{2\lceil\alpha-1\rceil\lceil\log(n)\rceil}}$$

and it converges to a constant as $n \to \infty$, since $\alpha + \beta > 3$. Using $j = \lceil\alpha-1\rceil\lceil\log(n)\rceil$ in the bound for θ yields

$$|\theta| \leq \frac{1}{8n^{1-\alpha}\lceil\alpha-1\rceil\lceil\log(n)\rceil} \log\left(1+\frac{\lambda_0(cn^{\beta}-1)}{8n}\right).$$

Since the upper bound grows to infinity as $n \to \infty$, we obtain the existence of n_0^* as described in the lemma. This completes the proof. □

5.14 Conclusion and future work

In this chapter we study the load balancing system in the many-server heavy-traffic regime. We parametrize the arrival rate so that the arrival rate *per server* is $n^{-\alpha}$, for $\alpha > 0$ where n is the number of servers. Specifically, we answer the question: how fast should the number of servers grow with respect to the load to observe the classical heavy-traffic behavior of the scaled average queue lengths?

We show that the answer strongly depends on the routing policy. If we model the system in discrete time, then all the arrivals of each time slot are routed to the same server. Then, as n increases, the deviations from the region where SSC occurs are large and, hence, higher values of α are required to observe the heavy-traffic behavior. If we model the system in continuous time, then the arrivals are routed one by one, and the value of α automatically decreases. A line of future work is to explore routing algorithms that do not route all the arrivals to the same server in discrete time.

The case of $\alpha \leq 1$ is well studied in the literature. Then, there is a gap between our

results and the literature. Future work is to explore how the system behaves if $\alpha \in (1, 3]$ for power-of-d choices and if $\alpha \in (1, 2]$ JSQ in continuous time. We believe that there are only two phase transitions for $\alpha \in (0, \infty)$: one at $\alpha = \frac{1}{2}$ which corresponds to the Halfin-Whitt regime; and one at $\alpha = 1$ which corresponds to the NDS regime. Hence, we need to develop new proof techniques to close the gap.

Another line of future work is to create a unifying framework for all $\alpha \in (0, \infty)$. The cases of $\alpha \in (0, 1]$ are well-studied in the literature. However, the proof techniques are different for every phase of α. We believe there is a framework which gives a generic result, where we can obtain the results from the literature by simply plugging in the desired value of α.

CHAPTER 6
HEAVY-TRAFFIC ANALYSIS OF THE GENERALIZED SWITCH UNDER THE CRP CONDITION

Based on:

D. Hurtado-Lange and S. T. Maguluri, "Transform methods for heavy-traffic analysis," *Stochastic Systems*, vol. 10, no. 4, pp. 275–309, 2020

6.1 Introduction

In this chapter we continue to develop the transform method introduced in chapter 3. In this case, we use this approach in the context of one of the most general SPNs with control on the service process: the generalized switch. The main goal of this chapter is to illustrate the flexibility and simplicity of the MGF method, this time in a system that seems very different from the single server queue and the load balancing system.

A secondary contribution, is that we allow the arrivals to different queues to be correlated. We show that this generalization from the popular assumption of independent arrivals does not bring additional difficulties in the proof.

6.2 Related work

In section 3.2 we discussed the use of moment generating functions and characteristic functions in the queueing literature. In this section we discuss the literature on the generalized switch and MaxWeight algorithm.

MaxWeight algorithm was first proposed in [21] in the context of scheduling for downlinks in wireless base stations. This algorithm was later adapted to be used in a variety of systems including ad hoc wireless networks, input-queued switches [95], cloud com-

puting [63], was generalized into the back-pressure algorithm [21] in networks, and was extensively studied in[8, 96, 97]. The generalized switch model subsumes many of these systems, and has been studied under the CRP condition using the diffusion limit method [8], and the drift method [34]. In [4] the authors generalize the results from [8] to SPNs where the jobs can join a queue after being served.

6.3 Generalized switch model

Consider a system with n separate queues, as described in section 1.5. For each $i \in [n]$, let $\{a_i(k) : k \in \mathbb{Z}_+\}$ 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. Let Σ_a be the covariance matrix of the vector $\boldsymbol{a}(1)$.

The servers interfere with each other. Then, the vector of service rates must satisfy feasibility constraints in each time slot. Additionally, there are conditions of the environment that affect these constraints, which we group in a single random variable called channel state. For each $k \in \mathbb{Z}_+$, let $M(k)$ be the channel state in time slot k. The sequence of random variables $\{M(k) : k \in \mathbb{Z}_+\}$ is i.i.d. and it is independent of the queue length and the arrival processes. We assume that the state space of the channel state is a finite set $\mathcal{M}$ and we let $\boldsymbol{\psi}$ be the probability mass function of $M(1)$, i.e., for each $m \in \mathcal{M}$ the probability of observing state m is $\psi_m \stackrel{\triangle}{=} \mathbb{P}[M(1) = m]$. For each $m \in \mathcal{M}$, let $\mathcal{S}^{(m)}$ be the set of feasible service rate vectors under channel state m, i.e., the set of service rate vectors that satisfy the interference constraints in channel state m. We assume that if $\boldsymbol{x} \in \mathcal{S}^{(m)}$ for some $m \in \mathcal{M}$, then all vectors that are strictly dominated by $\boldsymbol{x}$ are feasible. In other words, if $\boldsymbol{y}$ is a nonnegative vector that satisfies $\boldsymbol{y} \leq \boldsymbol{x}$ component-wise, then $\boldsymbol{y}$ is also a feasible service rate vector if the channel state is m. In particular, the projection of $\boldsymbol{x} \in \mathcal{S}^{(m)}$ on each of the coordinate axes is a feasible service rate vector as well. We assume that $\mathcal{S}^{(m)}$ is finite for each $m \in \mathcal{M}$, so we only consider maximal feasible schedules and their projection on the coordinate axes in $\mathcal{S}^{(m)}$. With this assumption we do not lose much generality because the vector $\boldsymbol{s}(k)$ is the potential (not actual) service rate vector and we are interested in the

heavy-traffic limit.

In this queueing system the control problem (which is a scheduling problem), is to select $\boldsymbol{s}(k)$ in each time slot after realizing the channel state. Let $\boldsymbol{s}(k)$ be the solution of the scheduling problem in time slot k. Since $\mathcal{S}^{(m)}$ is finite for each $m \in \mathcal{M}$ and $\mathcal{M}$ is also finite, there exists a constant $S_{\max}$ such that $s_i(k) \leq S_{\max}$ for all $i \in [n]$ and all $k \in \mathbb{Z}_+$.

It is known [34] that the capacity region of the generalized switch is

$$\mathcal{C} = \sum_{m \in \mathcal{M}} \psi_m \, \text{ConvexHull}\left\{\mathcal{S}^{(m)}\right\}. \tag{6.1}$$

Providing a formal proof of Equation 6.1 is beyond the scope of this document, but we intuitively explain why it holds. First suppose that the channel state is fixed and the set of feasible service rate vectors is $\mathcal{S}^{(1)}$. Then, the capacity region should have all vectors $\boldsymbol{x}$ that satisfy $\boldsymbol{x} \leq \boldsymbol{s}$ for all $\boldsymbol{s} \in \mathcal{S}^{(1)}$. Since $\mathcal{S}^{(1)}$ contains the projection of its elements on the coordinate axis, the set of such vectors $\boldsymbol{x}$ is ConvexHull $\left\{\mathcal{S}^{(1)}\right\}$. Now, if we consider the channel state as a random variable, recall that ψ_m is the probability that the channel state is m, and if the channel state is m then the set of feasible service rate vectors is $\mathcal{S}^{(m)}$. Then, Equation 6.1 just gives the capacity region associated to each channel state, weighted by the probability that each channel state is observed. In some sense, it represents the expected capacity region.

Recall that, by assumption, each set $\mathcal{S}^{(m)}$ is finite. Then, for each $m \in \mathcal{M}$ the set ConvexHull $\left\{\mathcal{S}^{(m)}\right\}$ is the convex hull of finitely many points. Therefore, ConvexHull $\left\{\mathcal{S}^{(m)}\right\}$ is a polytope, i.e., a bounded polyhedron. Also, the state space of the channel state $\mathcal{M}$ is finite by assumption. Then, Equation 6.1 is the weighted sum of finitely many polytopes. This implies that $\mathcal{C}$ is also a polytope. In order to exploit this structure, we describe it as the intersection of a finite number of half-spaces, where each half-space defines a facet of $\mathcal{C}$. Let L be the minimal number of half-spaces that are required to describe $\mathcal{C}$, and for each $\ell \in [L]$ let $\boldsymbol{c}^{(\ell)} \in \mathbb{R}^n$ and $b^{(\ell)} \in \mathbb{R}$ be the parameters that define each facet of the polytope.

In other words, we describe $\mathcal{C}$ as follows

$$\mathcal{C} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^n : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \leq b^{(\ell)} \quad \forall \ell \in [L] \right\}. \tag{6.2}$$

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

In this document, we assume that the scheduling problem is solved using MaxWeight algorithm in each time slot, i.e., if the channel state is m, then the selected schedule satisfies

$$\boldsymbol{s}(k) \in \arg \max_{\boldsymbol{x} \in \mathcal{S}^{(m)}} \langle \boldsymbol{x}, \boldsymbol{q}(k) \rangle, \tag{6.3}$$

and ties are broken at random.

For technical reasons that will be apparent in the following sections, we introduce the following definition. For each $\ell \in [L]$ and $m \in \mathcal{M}$ define the *maximum $\boldsymbol{c}^{(\ell)}$-weighted service rate available when channel state is* m as

$$b^{(m,\ell)} = \max_{\boldsymbol{x} \in \mathcal{S}^{(m)}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle. \tag{6.4}$$

Observe that $\boldsymbol{c}^{(\ell)}$ and $b^{(m,\ell)}$ define a half-space that passes through the boundary of $\text{ConvexHull}\big(\mathcal{S}^{(m)}\big)$, but this half-space does not necessarily define a facet of $\text{ConvexHull}\big(\mathcal{S}^{(m)}\big)$. For each $\ell \in [L]$ and $k \in \mathbb{Z}_+$, let $B_\ell(k) \stackrel{\triangle}{=} b^{(M(k),\ell)}$. Notice that $B_\ell(k)$ is an i.i.d. sequence of random variables that satisfies $\mathbb{P}\left[B_\ell(1) = b^{(m,\ell)}\right] = \psi_m$ for each $m \in \mathcal{M}$. Let Σ_B be the covariance matrix of the vector $\boldsymbol{B}(1) \stackrel{\triangle}{=} (B_1(1), \ldots, B_L(1))$, i.e., for each $\ell_1, \ell_2 \in [L]$ we have

$$(\Sigma_B)_{\ell_1,\ell_2} \stackrel{\triangle}{=} \mathbb{E}\left[B_{\ell_1}(k) B_{\ell_2}(k)\right] - \mathbb{E}\left[B_{\ell_1}(k)\right] \mathbb{E}\left[B_{\ell_2}(k)\right].$$

From Equation 6.1 and Equation 6.3, observe that the service rate vector $\boldsymbol{s}(k)$ does not necessarily belong to the capacity region $\mathcal{C}$ because $\psi_m \leq 1$ for all $m \in \mathcal{M}$. However, the expected service rate vector does belong to the capacity region. We end this section proving this result formally.

Lemma 6.1. *Consider a generalized switch operating under MaxWeight as described above, and let* $\mathbb{E}_{\boldsymbol{q}}[\,\cdot\,] \stackrel{\triangle}{=} \mathbb{E}[\,\cdot\,|\boldsymbol{q}(k) = \boldsymbol{q}]$. *Then,* $\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}(k), \boldsymbol{s}(k)\rangle] = \max_{\boldsymbol{x}\in\mathcal{C}}\langle \boldsymbol{q}, \boldsymbol{x}\rangle$.

Proof. Since $\boldsymbol{s}(k)$ is selected using MaxWeight algorithm (see Equation 6.3), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}(k), \boldsymbol{s}(k)\rangle] =& \mathbb{E}_{\boldsymbol{q}}\left[\max_{\boldsymbol{x}\in\mathcal{S}^{(M(k))}}\langle \boldsymbol{q}(k), \boldsymbol{x}\rangle\right] \\ \stackrel{(a)}{=}& \sum_{m\in\mathcal{M}} \psi_m \max_{\boldsymbol{x}\in\mathcal{S}^{(m)}}\langle \boldsymbol{q}, \boldsymbol{x}\rangle \stackrel{(b)}{=} \max_{\boldsymbol{x}\in\mathcal{C}}\langle \boldsymbol{q}, \boldsymbol{x}\rangle,\end{aligned}$$

where (a) holds because the channel state process is independent from the queue lengths process; and (b) holds by definition of the capacity region $\mathcal{C}$ presented in Equation 6.1. □

6.4 Transform method applied to generalized switches

In this section we apply the MGF method in the context of a generalized switch operating under MaxWeight. We compute the distribution of the scaled vector of queue lengths in heavy traffic under the assumption that CRP is satisfied. The generalized switch is a model that was first introduced in [8], and it represents a generalization of a variety of queueing systems, such as the input-queued switch [95], cloud computing [63], down-links in wireless base stations [21], etc.

To perform heavy-traffic analysis, we fix a facet $\mathcal{F}^{(\ell)}$ and we study a set of generalized switches where the vector of arrival rates approaches a fixed point in the relative interior of $\mathcal{F}^{(\ell)}$. Formally, we fix $\boldsymbol{\nu}^{(\ell)}$ in the relative interior of $\mathcal{F}^{(\ell)}$ and we let $\epsilon \in (0, 1)$. Then, the system parametrized by ϵ is such that $\mathbb{E}\left[\boldsymbol{a}^{(\epsilon)}(k)\right] = \boldsymbol{\nu}^{(\ell)} - \epsilon\boldsymbol{c}^{(\ell)}$ and $\Sigma_a^{(\epsilon)}$ is the covariance matrix of $\boldsymbol{a}^{(\epsilon)}(1)$. In this case, since the point $\boldsymbol{\nu} = \boldsymbol{\nu}^{(\ell)}$ of the boundary of the capacity

region $\mathcal{C}$ is in the relative interior of the facet $\mathcal{F}^{(\ell)} = \left\{\boldsymbol{x} \in \mathcal{C} : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle = b^{(\ell)}\right\}$, the unique outer normal vector to the capacity region $\mathcal{C}$ at $\boldsymbol{\nu}^{(\ell)}$ is the outer normal vector to the facet $\mathcal{F}^{(\ell)}$, i.e., it is $\boldsymbol{c}^{(\ell)}$. Therefore, the CRP condition as defined in Definition 2.1 is satisfied. Observe that if we approach a vector $\boldsymbol{\nu}$ that lies at the intersection of two (or more) facets in heavy traffic, then the CRP condition is not satisfied because there is a range of vectors that are normal to $\mathcal{C}$ at $\boldsymbol{\nu}$.

In this subsection we state the main theorem of this section and we provide some examples. We prove the theorem in subsection 6.4.2.

Theorem 6.2. *Let $\epsilon \in (0, 1)$. Given the ℓ^{th} facet of $\mathcal{C}$, $\mathcal{F}^{(\ell)}$, and a vector $\boldsymbol{\nu}^{(\ell)}$ in the relative interior of $\mathcal{F}^{(\ell)}$, consider a set of generalized switches operating under MaxWeight algorithm as described in section 6.3, parametrized by ϵ as described above. For each ϵ, let $\overline{\boldsymbol{q}}^{(\epsilon)}$ be a steady-state vector such that the queue length process $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ converges in distribution to $\overline{\boldsymbol{q}}^{(\epsilon)}$. Further, let $\lim_{\epsilon \downarrow 0} \Sigma_a^{(\epsilon)} = \Sigma_a$ component-wise. Then, $\epsilon \overline{\boldsymbol{q}}^{(\epsilon)} \Rightarrow \overline{\Upsilon} \boldsymbol{c}^{(\ell)}$ as $\epsilon \downarrow 0$, where $\overline{\Upsilon}$ is an exponential random variable with mean $\frac{1}{2}\left((\boldsymbol{c}^{(\ell)})^T \Sigma_a \boldsymbol{c}^{(\ell)} + \sigma^2_{B_\ell}\right)$, where $\sigma^2_{B_\ell} = (\Sigma_B)_{\ell,\ell}$ is the variance of $B_\ell(1)$.*

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

Corollary 6.3. *Consider a set of generalized switches operating under MaxWeight algorithm as described in section 6.3, parametrized by ϵabove. Suppose that $\mathcal{M}$ has only one element, i.e., the channel state is fixed over time. Then, $\epsilon \overline{\boldsymbol{q}}^{(\epsilon)} \Rightarrow \overline{\Upsilon}_2 \boldsymbol{c}^{(\ell)}$, where $\overline{\Upsilon}_2$ is an exponential random variable with mean $\frac{1}{2}(\boldsymbol{c}^{(\ell)})^T \Sigma_a \boldsymbol{c}^{(\ell)}$.*

The queueing system described in Corollary 6.3 is also known as ad hoc wireless network. In an ad hoc wireless network we have $\sigma^2_{B_\ell} = 0$ because the channel state is not a random variable anymore. The input-queued switch or a cross-bar switch [36, 14, 15] is yet another system that is well studied. When only one port of the switch is saturated,

it satisfies the CRP condition [8], and forms a special case of Corollary 6.3. In the next subsection we present the model and we formalize this result.

6.4.1 MGF method applied to the input-queued switch

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

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

$$\mathcal{C}_{\text{switch}} \stackrel{\triangle}{=} \left\{ \boldsymbol{x} \in \mathbb{R}^n_+ : \ \langle \boldsymbol{\chi}^{(i_r)}, \boldsymbol{x} \rangle \leq 1, \ \langle \widetilde{\boldsymbol{\chi}}^{(i_c}, \boldsymbol{x} \rangle \leq 1 \ , \ \forall i_r, i_c \in [n] \right\},$$

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

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

Similar to the case of the load balancing system, we say that an algorithm $\mathcal{A}$ is through-

put optimal for the input-queued switch if $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent for all $\epsilon \in (0,1)$. Also, defining $\boldsymbol{x}_{\|} \stackrel{\triangle}{=} \langle \boldsymbol{\chi}^{(1)}, \boldsymbol{x}\rangle \boldsymbol{\chi}^{(1)}$ and $\boldsymbol{x}_{\perp} \stackrel{\triangle}{=} \boldsymbol{x} - \boldsymbol{x}_{\|}$ for any vector $\boldsymbol{x}$, we say that the switch operating under a scheduling algorithm $\mathcal{A}$ satisfies SSC if

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

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

Proposition 6.4. *Let $\epsilon \in (0,1)$ and consider a set of input-queued switches parametrized by ϵ, as described above. Suppose that the scheduling algorithm is throughput optimal and it satisfies SSC. For each $\epsilon \in (0,1)$, let $\overline{\boldsymbol{q}}^{(\epsilon)}$ be a steady-state random vector such that the queue length process $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ converges in distribution to $\overline{\boldsymbol{q}}^{(\epsilon)}$. Assume the MGF of $\epsilon\langle \boldsymbol{\chi}^{(1)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists, and that $\lim_{\epsilon \downarrow 0} \Sigma_a^{(\epsilon)} = \Sigma_a$ component-wise. Then, $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} \implies \overline{\Upsilon}_s \boldsymbol{\chi}^{(1)}$ as $\epsilon \downarrow 0$, where $\overline{\Upsilon}_s$ is an exponential random variable with mean $\frac{1}{2}\sum_{i_r=1}^{n}\sum_{i_c=1}^{n} \chi_{i_r}^{(1)}\chi_{i_c}^{(1)} Cov\left[a_{i_r}, a_{i_c}\right]$, where $Cov\left[a_{i_r}, a_{i_c}\right] = (\Sigma_a)_{i_r,i_c}$ for each $i_r, i_c \in [n]$.*

Sketch of proof of Proposition 6.4. For ease of exposition we do not write the dependence on ϵ of the variables. We use the MGF method. We only present a sketch of this proof, since it is similar to the proofs of Theorem 3.5 and Theorem 6.2. We only show the main differences.

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

Step 1. Prove an equation of the form of Equation 3.1 and compute an expression for the MGF of $\epsilon\langle \boldsymbol{\chi}^{(1)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$.

Proving an equation of the form of Equation 3.1 is similar to the proof of Lemma 3.9 and Lemma 6.5. Then, following the steps sketched in Step 1 in section 3.3 we obtain

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

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

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

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

This step is equivalent to Step 2 in the proof of Theorem 3.5 and Theorem 6.2, so we omit the details. □

In the case of a generalized switch, one of the difficulties is to handle the dependence on the queue lengths of the potential service vector. In the case of an input-queued switch this difficulty does not arise because, even though $\overline{\boldsymbol{s}}^{(\epsilon)}$ depends on the queue lengths, the projection $\boldsymbol{s}_{\parallel}^{(\epsilon)} \stackrel{\triangle}{=} \langle \boldsymbol{\chi}^{(1)}, \overline{\boldsymbol{s}}^{(\epsilon)} \rangle \boldsymbol{\chi}^{(1)}$ is independent of $\overline{\boldsymbol{q}}^{(\epsilon)}$. Therefore, we do not need to assume that the scheduling problem is solved with MaxWeight. In general, for any special case of the generalized switch such that $\boldsymbol{s}_{\parallel}^{(\epsilon)}$ is independent of the queue lengths, we can obtain a result similar to Proposition 6.4, i.e., where we assume properties of the scheduling algorithm but not a specific algorithm.

6.4.2 Proof of Theorem 6.2

In the rest of this section we prove Theorem 6.2 using the MGF method. Before presenting the proof, we introduce some notation.

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

Proof of Theorem 6.2. For ease of exposition we omit the dependence on ϵ of the variables in this proof. We use the MGF method. Similarly to the proof of Theorem 3.5, we first need to verify that the prerequisites are satisfied.

Prerequisite 1. Positive recurrence.

In fact, MaxWeight algorithm is throughput optimal [8, 34]. Then, for each $\epsilon > 0$ the Markov chain $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent.

Prerequisite 2. SSC.

Let $\mathcal{K} = \left\{\boldsymbol{x} \in \mathbb{R}_+^n : \boldsymbol{x} = \xi \boldsymbol{c}^{(\ell)}\ ,\ \xi \in \mathbb{R}_+\right\}$. Using the notation introduced in Prerequisite 2 in section 3.3, we have $\boldsymbol{c} = \boldsymbol{c}^{(\ell)}$, $\overline{\boldsymbol{q}}_{\|}^{(\epsilon)} = \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle \boldsymbol{c}^{(\ell)}$ and $\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)} = \overline{\boldsymbol{q}}^{(\epsilon)} - \overline{\boldsymbol{q}}_{\|}^{(\epsilon)}$. In [34], the authors prove that $\mathbb{E}\left[e^{\theta^* \|\overline{\boldsymbol{q}}_{\perp}\|}\right]$ is bounded for some finite θ^*. In fact, the exponential moment bound is not part of the SSC statement of [34], but their proof of Proposition 2 implies it.. Then, for each $j \in \mathbb{Z}_+$ with $j \geq 1$, there exists a constant J_j such that $\mathbb{E}\left[\left\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\right\|^j\right] \leq J_j$. Therefore, SSC as defined in section 3.3 is satisfied, and it occurs into the one-dimensional subspace $\mathcal{K}$. In fact, in this case $\mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^j\right]$ is $O(1)$, which is stronger.

Now we go through the steps of the MGF method.

Step 1. Prove an equation of the form of Equation 3.1 and compute an expression for the MGF of $\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle$.

We first prove Lemma 6.5.

Lemma 6.5. *Consider a generalized switch parametrized by ϵ as described in Theorem 6.2. Then, for any real number θ such that $|\theta\epsilon| \leq \theta^*$ we have*

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

We present the proof of Lemma 6.5 in subsection 6.5.1.

Before continuing, we need to prove that the MGF of $\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle$ exists in an interval around 0. The proof is presented in subsection 6.5.2. Then, following the steps sketched in Step 1 in section 3.3 we obtain Equation 3.4.

When we applied the MGF method to the single server queue and to the load balancing system, we used the fact that the service rate vector is independent of the queue length vector to obtain Equation 2.7 and Equation 3.7, respectively. However, in the case of the generalized switch this is no longer true. To overcome this difficulty we use the following lemma.

Lemma 6.6. *Consider a generalized switch operating under MaxWeight algorithm parametrized by ϵ, as described in Theorem 6.2. Then, for any $\theta \in \mathbb{R}$ we have*

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

We present the proof in subsection 6.5.3. Working with the left hand side of Equation 3.4 we obtain

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

where (a) holds after adding and subtracting $\mathbb{E}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle + \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle - \overline{B}\right)}\right]$, and reorganizing terms; (b) holds because $\overline{\boldsymbol{a}}$ and $\overline{B}$ are independent of the queue lengths vector $\overline{\boldsymbol{q}}$ and the potential service vector $\overline{\boldsymbol{s}}$, and after adding and subtracting $\mathbb{E}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle - \overline{B}\right)}\right]\mathbb{E}\left[e^{\theta\epsilon\left(\overline{B} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right)} - 1\right]$; and (c) holds by Lemma 6.6 and because $\overline{\boldsymbol{a}}$ and $\overline{B}$ are bounded. Reorganizing terms we

obtain

$$\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{q}}\rangle}\right] = \frac{1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}\rangle}\right]+\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{a}}\rangle}\right]\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle}-e^{-\theta\epsilon\overline{B}}\right]+o(\epsilon^2)}{1-\mathbb{E}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{a}}\rangle-\overline{B}\right)}\right]}. \tag{6.5}$$

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

The right hand side of Equation 6.5 yields a $\frac{0}{0}$ form in the limit as $\epsilon \downarrow 0$. Then, we take Taylor expansion of each of its terms, using Lemma 3.1. Similar to the case of the load balancing system, in this step we need to obtain bounds on $\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}\rangle\right]$. In this case we use the following lemma.

Lemma 6.7. *Consider a generalized switch parametrized by ϵ as described in Theorem 6.2. Then,*

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

Proof of Lemma 6.7. We set to zero the drift of $V_1(\boldsymbol{q})=\langle \boldsymbol{c}^{(\ell)},\boldsymbol{q}\rangle$. We obtain

$$\begin{aligned}
0=&\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{q}}^{+}\rangle-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{q}}\rangle\right]\\
=&\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{q}}+\overline{\boldsymbol{a}}-\overline{\boldsymbol{s}}+\overline{\boldsymbol{u}}\rangle-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{q}}\rangle\right]\\
=&\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{a}}\rangle\right]-\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right]+\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}\rangle\right].
\end{aligned} \tag{6.6}$$

Now, observe that

$$\begin{aligned}
\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{a}}\rangle\right]=&\langle \boldsymbol{c}^{(\ell)},\boldsymbol{\nu}^{(\ell)}-\epsilon\boldsymbol{c}^{(\ell)}\rangle\\
=&\langle \boldsymbol{c}^{(\ell)},\boldsymbol{\nu}^{(\ell)}\rangle-\epsilon\left\|\boldsymbol{c}^{(\ell)}\right\|^2\\
\overset{(a)}{=}&b^{(\ell)}-\epsilon,
\end{aligned} \tag{6.7}$$

where (a) holds because $\boldsymbol{\nu}^{(\ell)} \in \mathcal{F}^{(\ell)}$ and because $\left\|\boldsymbol{c}^{(\ell)}\right\| = 1$. Then, using Equation 6.7 in Equation 6.6 and rearranging terms we obtain the result. □

Now we expand each term on the right-hand side of Equation 6.5. For the first term in the numerator, we have

$$\begin{aligned}1 - \mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle}\right] =& 1 - \mathbb{E}\left[f_{\epsilon,-\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle}(\theta)\right] \\ =& \theta\epsilon\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] - \frac{(\theta\epsilon)^2}{2}\mathbb{E}\left[\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right)^2\right] + O(\epsilon^3). \end{aligned} \tag{6.8}$$

In this case the numerator has more terms than in the case of the single server queue and the load balancing system, so we will keep the first moment of the unused service in the equation in order to use Lemma 6.7. However, we still need to bound the second moment.

Claim 6.8. *Consider a generalized switch as described in Theorem 6.2. Then,*

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

We present a proof of Claim 6.8 in subsection 6.5.4. Then, using Claim 6.8 in Equation 6.8 we obtain

$$1 - \mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle}\right] = \theta\epsilon\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] + O(\epsilon^3). \tag{6.9}$$

For the second term in the numerator, we have

$$\begin{aligned}&\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle}\right]\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle} - e^{-\theta\epsilon\overline{B}}\right] \\ &= \mathbb{E}\left[e^{\theta\epsilon\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle - \overline{B}\right)}\right]\mathbb{E}\left[e^{\theta\epsilon\left(\overline{B} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right)} - 1\right] \\ &= \mathbb{E}\left[f_{\epsilon,\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle - \overline{B}\right)}(\theta)\right]\mathbb{E}\left[f_{\epsilon,\left(\overline{B} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right)}(\theta) - 1\right]. \end{aligned} \tag{6.10}$$

Claim 6.9. *Consider a generalized switch as described in Theorem 6.2 and the notation*

introduced in Lemma 3.1. Then,

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

We prove the claim in subsection 6.5.5. Using Claim 6.9 in Equation 6.10, reorganizing terms and using that $\overline{B}$ and $\overline{s}_i$ are bounded for all $i \in [n]$, we obtain

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

Then, the numerator of Equation 6.5 yields

$$\begin{aligned}
&1-\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}^{(\epsilon)}\rangle}\right]+\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{a}}\rangle}\right]\mathbb{E}\left[e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle}-e^{-\theta\epsilon\overline{B}}\right]+o(\epsilon^2)\\
&=\left(\theta\epsilon\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}\rangle\right]+\theta\epsilon\mathbb{E}\left[\overline{B}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right]+O(\epsilon^3)\right)+o(\epsilon^2)\\
&\overset{(a)}{=}\theta\epsilon\left(\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{u}}\rangle+\overline{B}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right]\right)+o(\epsilon^2)\\
&\overset{(b)}{=}\theta\epsilon^2+o(\epsilon^2),
\end{aligned} \tag{6.11}$$

where (a) holds because $O(\epsilon^3)$ is $o(\epsilon^2)$; and (b) holds because $\mathbb{E}\left[\overline{B}\right] = b^{(\ell)}$ (we prove this result in Lemma 7.1) and Lemma 6.7.

For the denominator, we obtain

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

$$\overset{(c)}{=}\theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\mathbf{1}^T\Sigma_a^{(\epsilon)}\mathbf{1} + \sigma_{B_\ell}^2 + \epsilon^2\right) + O(\epsilon^3) \tag{6.12}$$

where (a) holds by Equation 6.7 and expanding the square; (b) holds by definition of variance and covariance, because $\overline{\boldsymbol{a}}$ and $\overline{B}$ are independent, and reorganizing terms; and (c) holds by Equation 6.7 and noticing that $\sum_{i=1}^n \sum_{i'=1}^n \text{Cov}\left[a_i^{(\epsilon)}, a_{i'}^{(\epsilon)}\right] = \mathbf{1}^T\Sigma_a^{(\epsilon)}\mathbf{1}$.

Using Equation 6.11 and Equation 6.12 in Equation 6.5 we obtain

$$\begin{aligned}\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle}\right] =& \frac{\theta\epsilon^2 + o(\epsilon^2)}{\theta\epsilon^2 - \frac{(\theta\epsilon)^2}{2}\left(\mathbf{1}^T\Sigma_a^{(\epsilon)}\mathbf{1} + \sigma_{B_\ell}^2 + \epsilon^2\right) + O(\epsilon^3)} \\ =& \frac{1 + o(1)}{1 - \frac{\theta}{2}\left(\mathbf{1}^T\Sigma_a^{(\epsilon)}\mathbf{1} + \sigma_{B_\ell}^2 + \epsilon^2\right) + O(\epsilon)}.\end{aligned}$$

Then, taking the heavy-traffic limit yields

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}, \overline{\boldsymbol{q}}\rangle}\right] = \frac{1}{1 - \frac{\theta}{2}\left(\mathbf{1}^T\Sigma_a\mathbf{1} + \sigma_{B_\ell}^2\right)},$$

which is the MGF of an exponential random variable with mean $\frac{1}{2}\left(\mathbf{1}^T\Sigma_a\mathbf{1} + \sigma_{B_\ell}^2\right)$. This implies that $\overline{\boldsymbol{q}}_{\parallel}^{(\epsilon)} = \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle \boldsymbol{c}^{(\ell)} \Rightarrow \overline{\Upsilon}\boldsymbol{c}^{(\ell)}$, where $\overline{\Upsilon}$ is an exponential random variable with mean $\frac{1}{2}\left(\mathbf{1}^T\Sigma_a\mathbf{1} + \sigma_{B_\ell}^2\right)$.

Then, we conclude that $\epsilon\overline{\boldsymbol{q}}^{(\epsilon)} = \epsilon\overline{\boldsymbol{q}}_{\parallel}^{(\epsilon)} + \epsilon\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}$ converges in distribution to $\overline{\Upsilon}\boldsymbol{c}^{(\ell)}$ as $\epsilon \downarrow 0$. This proves Theorem 6.2. □

6.5 Details of the proofs of Section 6.4

In this section we present the details of the proofs of Section 6.4.

6.5.1 Proof of Lemma 6.5

To prove Lemma 6.5 we use the following lemma, which is similar to Lemma 3.14.

Lemma 6.10. *Consider a generalized switch parametrized by ϵ, as described in Theorem*

6.2. Then, for any $\alpha \in \mathbb{R}$ and all $k \in \mathbb{Z}_+$ we have

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

Proof of Lemma 6.10. First observe that if $\alpha = 0$ the lemma trivially holds. Now we prove the lemma for $\alpha \neq 0$. From Equation 1.3 we know that $q_i(k+1)u_i(k) = 0$ for all $i \in [n]$. Then, for all $\beta \in \mathbb{R}$ we have

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

and this equation implies

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

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

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

Summing over all $i \in [n]$ we obtain

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

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

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

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

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

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

which proves the lemma. □

Now we prove Lemma 6.5.

Proof of Lemma 6.5. First observe that if $\theta = 0$ the lemma holds trivially. Now assume $\theta \neq 0$. Since $\boldsymbol{c}^{(\ell)} \geq 0$ and $\overline{u}_i \leq \overline{s}_i \leq S_{\max}$ for all $i \in [n]$, we have

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

Then, from facts item (i) and item (ii) stated in subsection 3.6.3, we have

$$\left| e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle} \right| \leq \left| \theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle \right| \left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \mathbf{1}\rangle} - 1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \mathbf{1}\rangle} \right). \tag{6.13}$$

Now, by properties of expected value, we have

$$\begin{aligned}
&\left| \mathbb{E}\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{+}\rangle} - 1 \right) \left(e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle} - 1 \right) \right] \right| \\
\leq &\mathbb{E}\left[\left| e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{+}\rangle} - 1 \right| \left| e^{-\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle} - 1 \right| \right] \\
\overset{(a)}{\leq} &|\theta\epsilon| \left(\frac{e^{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \mathbf{1}\rangle} - 1}{-\theta\epsilon S_{\max}\langle \boldsymbol{c}^{(\ell)}, \mathbf{1}\rangle} \right) \mathbb{E}\left[\left| \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle \left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{+}\rangle} - 1 \right) \right| \right]
\end{aligned}$$

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

where $j\in\mathbb{Z}_+$ satisfies $j>1$. Here (a) holds by Equation 6.13; (b) holds by Lemma 6.10 with $\alpha=\theta\epsilon$; (c) holds by triangle inequality; and (d) holds by Hölder's inequality.

But

$$
\mathbb{E}\left[\sum_{i=1}^{n}\left(c_i^{(\ell)}\overline{u}_i\right)^{j}\right]\leq (c_{\max}S_{\max})^{j-1}\mathbb{E}\left[\sum_{i=1}^{n} c_i^{(\ell)}\overline{u}_i\right]\leq (c_{\max}S_{\max})^{j-1}\epsilon,
$$

where $c_{\max}=\max_{i\in[n]} c_i^{(\ell)}$ and the last equality holds by the following reason. By Lemma 6.7 we have

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

Also, recall that $\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right]\in\mathcal{C}$, by Lemma 6.1. Then,

$$
-b^{(\ell)}+\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right]\leq 0. \tag{6.14}
$$

Therefore,

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

The rest of the argument is similar to the last steps in the proof of Lemma 3.9. However, in this case we do not need to use existence of the MGF of $\epsilon \sum_{i=1}^{n} \overline{q}_i$ because we know $\mathbb{E}\left[e^{\theta\epsilon\|\boldsymbol{q}_\perp\|}\right]$ is bounded for $\theta\epsilon \leq \theta^*$ from SSC.

□

6.5.2 Existence of MGF of $\epsilon \|\overline{\boldsymbol{q}}\|$ in the generalized switch

We prove the following lemma.

Lemma 6.11. *Consider a generalized switch parametrized by ϵ as described in Theorem 6.2. Then, there exists $\Theta > 0$ (which is independent of ϵ) such that $\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle}\right] < \infty$ for all $\theta \in [-\Theta, \Theta]$.*

Lemma 6.11 can be proved using Foster-Lyapunov theorem (Theorem 2.4), following similar steps to the proof of Lemma 3.13. However, here we propose a different approach. We use the Lemma 6.12, which presents explicit bounds for the setting of Lemma 2.7.

Lemma 6.12. *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 nonnegative valued Lyapunov function and consider its drift $\Delta Z(x)$ as defined in Definition 2.2. Suppose the following conditions are satisfied:*

(C1) There exists $\eta > 0$ and $\kappa < \infty$ such that $\mathbb{E}\left[\Delta Z(x) \mid X(k) = x\right] \leq -\eta$ for all $x \in \mathcal{X}$ with $Z(x) \geq \kappa$,

(C2) There exists $D < \infty$ such that $|\Delta Z(x)| \leq D$ with probability 1 for all $x \in \mathcal{X}$

and $\eta^2 \leq 4(e^D - 1 - D)$. Define

$$\Theta^* \stackrel{\triangle}{=} \min\left\{1,\ \frac{\eta}{2(e^D - 1 - D)}\right\}.$$

Then, for any $\theta^* \in (0, \Theta^*)$ we have

$$\lim_{k\to\infty} \mathbb{E}\left[e^{\theta^* V(X_k)}\right] \leq 2\left(1 + \frac{D}{\eta}\right) e^{\theta^* \kappa}.$$

We present the proof of the lemma at the end of this section. Now we prove existence of MGF for the generalized switch.

Proof of Lemma 6.11. First observe that if $\theta \leq 0$ the lemma holds trivially. Therefore, in this proof we assume $\theta > 0$.

We use Lemma 6.12 with $Z(\boldsymbol{q}) = \|\boldsymbol{q}\|$.

To show that condition (C1) is satisfied, we first observe that $f(x) = \sqrt{x}$ is a concave function, and $Z(\boldsymbol{q}) = \sqrt{\|\boldsymbol{q}\|^2}$. Then,

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta Z(\boldsymbol{q})\right] \leq \frac{1}{2\|\boldsymbol{q}\|}\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k+1)\|^2 - \|\boldsymbol{q}\|^2\right]. \tag{6.15}$$

We now compute an upper bound on $\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k+1)\|^2 - \|\boldsymbol{q}\|^2\right]$. For every $k \in \mathbb{Z}_+$, we have

$$\begin{aligned}
&\|\boldsymbol{q}(k+1)\|^2 - \|\boldsymbol{q}(k)\|^2 \\
&= \|\boldsymbol{q}(k+1) - \boldsymbol{u}(k) + \boldsymbol{u}(k)\|^2 - \|\boldsymbol{q}(k)\|^2 \\
&\overset{(a)}{=} \|\boldsymbol{q}(k) + \boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2 + \|\boldsymbol{u}(k)\|^2 + 2\langle \boldsymbol{q}(k+1) - \boldsymbol{u}(k), \boldsymbol{u}(k)\rangle - \|\boldsymbol{q}(k)\|^2 \\
&\overset{(b)}{\leq} \|\boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2 + 2\langle \boldsymbol{q}(k), \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle,
\end{aligned} \tag{6.16}$$

where (a) holds by Equation 1.2; and (b) holds by Equation 1.3, because $\|\boldsymbol{u}(k)\|^2 \geq 0$, and expanding the first term. Next we bound the conditional expectation of each of the terms in Equation 6.16. For the first term we have,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2\right] &\overset{(a)}{\leq} \mathbb{E}\left[\|\boldsymbol{a}(k)\|^2\right] + \mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{s}(k)\|^2\right] && (6.17)\\
&\overset{(b)}{\leq} \sum_{i=1}^{n}\left((\lambda_i^{(\epsilon)})^2 + \sigma_{a_i}^2\right) + nS_{\max}^2, && (6.18)
\end{aligned}$$

where (a) holds by triangle inequality and because the arrival process is independent of the queue length process; and (b) holds by definition of variance and because the potential service to each of the queues is upper bounded by $S_{\max}$. Define $\zeta_1 \stackrel{\triangle}{=} \sum_{i=1}^{n} \left((\lambda_i^{(\epsilon)})^2 + \sigma_{a_i}^2 \right) + nS_{\max}^2$.

For the second term in Equation 6.16 we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{q}} \left[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k) \rangle \right] &\stackrel{(a)}{=} \langle \boldsymbol{q}, \boldsymbol{\nu}^{(\ell)} - \epsilon \boldsymbol{c}^{(\ell)} \rangle - \langle \boldsymbol{q}, \mathbb{E}_{\boldsymbol{q}} \left[\boldsymbol{s}(k) \right] \rangle \\ &\stackrel{(b)}{\leq} \langle \boldsymbol{q}, \boldsymbol{\nu}^{(\ell)} - \epsilon \boldsymbol{c}^{(\ell)} \rangle - \langle \boldsymbol{q}, \boldsymbol{s}^* \rangle, \end{aligned}$$

where $\boldsymbol{s}^* \in \mathcal{C}$. Here, (a) holds because the arrival processes are independent of the queue lengths, and by definition of $\boldsymbol{\lambda}^{(\epsilon)}$; and (b) holds for any $\boldsymbol{s}^* \in \mathcal{C}$ because scheduling occurs according to MaxWeight algorithm, and by Lemma 6.1.

We now compute a vector $\boldsymbol{s}^* \in \mathcal{C}$. Define a vector $\boldsymbol{d}^{(\ell)}$ with elements $d_i^{(\ell)} \stackrel{\triangle}{=} \mathbb{1}_{\left\{c_i^{(\ell)} = 0\right\}}$ for each $i \in [n]$, and observe that $\boldsymbol{d}^{(\ell)}$ is orthogonal to $\boldsymbol{c}^{(\ell)}$. Since $\boldsymbol{\nu}^{(\ell)}$ is in the relative interior of the facet $\mathcal{F}^{(\ell)}$, there exists a positive number $\delta \in (0, 1)$ such that

$$\boldsymbol{\nu}^{(\ell)} + \epsilon \delta \boldsymbol{d}^{(\ell)} \in RelativeInterior(\mathcal{F}^{(\ell)}).$$

Set $\boldsymbol{s}^* = \boldsymbol{\nu}^{(\ell)} + \epsilon \delta \boldsymbol{d}^{(\ell)}$, and note that if all the components of $\boldsymbol{c}^{(\ell)}$ are strictly positive, then $\boldsymbol{d}^{(\ell)} = \boldsymbol{0}$ and $\boldsymbol{s}^* = \boldsymbol{\nu}^{(\ell)}$. Then, we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{q}} \left[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k) \rangle \right] &\leq -\epsilon \langle \boldsymbol{q}, \boldsymbol{c}^{(\ell)} + \delta \boldsymbol{d}^{(\ell)} \rangle \\ &\stackrel{(a)}{\leq} -\epsilon \rho^{(\ell)} \sqrt{n} \|\boldsymbol{q}\|, \end{aligned} \tag{6.19}$$

where $\rho^{(\ell)} \stackrel{\triangle}{=} \min_{i \in [n]} \{ c_i^{(\ell)} + \delta d_i^{(\ell)} \}$. Here, (a) holds by the Cauchy-Schwarz inequality. Observe that $\rho^{(\ell)}$ is a constant independent of ϵ and that $\rho^{(\ell)} > 0$ by definition of δ and $\boldsymbol{d}^{(\ell)}$.

Using Equation 6.16, Equation 6.17 and Equation 6.19 in Equation 6.15, and rearrang-

ing terms we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta Z(\boldsymbol{q})\right] \leq \frac{\zeta_1}{2\|\boldsymbol{q}\|} - \epsilon\rho^{(\ell)}\sqrt{n}.$$

Therefore, condition $(C1)$ is satisfied with

$$\eta \stackrel{\triangle}{=} \frac{\epsilon\rho^{(\ell)}\sqrt{n}}{2}, \text{ and } \kappa \stackrel{\triangle}{=} \frac{\zeta_1}{\epsilon\rho^{(\ell)}\sqrt{n}}.$$

Now we show condition (C2). We have

$$\begin{aligned}
|\Delta Z(\boldsymbol{q})| &= |\|\boldsymbol{q}(k+1)\| - \|\boldsymbol{q}(k)\|| \, \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\stackrel{(a)}{\leq} \|\boldsymbol{q}(k+1) - \boldsymbol{q}(k)\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\stackrel{(b)}{\leq} \|\boldsymbol{a}(k) - \boldsymbol{s}(k) + \boldsymbol{u}(k)\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\stackrel{(c)}{\leq} \left(\|\boldsymbol{a}(k)\| + \|\boldsymbol{s}(k) - \boldsymbol{u}(k)\|\right) \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\stackrel{(d)}{\leq} \sqrt{n}(A_{\max} + S_{\max}),
\end{aligned}$$

where (a) holds by triangle inequality; (b) holds by Equation 1.2; (c) holds by triangle inequality; and (d) because for each $i \in [n]$ we have $a_i(k) \leq A_{\max}$ and $s_i(k) - u_i(k) \leq s_i(k) \leq S_{\max}$ with probability 1. Hence, condition (C2) is satisfied with $D \stackrel{\triangle}{=} \sqrt{n}(A_{\max} + S_{\max})$. Observe that $\eta \leq \frac{1}{2}$ and $D \geq 2$ because we must have $A_{\max} \geq 1$ and $S_{\max} \geq 1$. Then, $\eta^2 \leq 4(e^D - 1 - D)$ holds trivially.

Therefore, setting $\theta^* \stackrel{\triangle}{=} \theta\epsilon$ in Lemma 6.12 we obtain

$$\mathbb{E}\left[e^{\theta\epsilon\|\overline{\boldsymbol{q}}\|}\right] \leq 2\left(1 + \frac{2\sqrt{n}(A_{\max} + S_{\max})}{\epsilon\rho^{(\ell)}\sqrt{n}}\right) e^{\frac{\theta\zeta_1}{\rho^{(\ell)}\sqrt{n}}}$$

for any $\theta \leq \Theta$, where, using that $\epsilon < 1$ we obtain

$$\Theta = \min\left\{1,\ \frac{\rho^{(\ell)}\sqrt{n}}{2\left(e^{\sqrt{n}(A_{\max}+S_{\max})} - 1 - \sqrt{n}(A_{\max}+S_{\max})\right)}\right\}.$$

Since $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle = \|\overline{\boldsymbol{q}}_{\|}\|$ and the projection is nonexpansive, we have $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle \leq \|\overline{\boldsymbol{q}}\|$. Hence, for every $\theta \in (0, \Theta]$ we have

$$\mathbb{E}\left[e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle}\right] \leq 2\left(1 + \frac{2\sqrt{n}(A_{\max}+S_{\max})}{\epsilon\rho^{(\ell)}\sqrt{n}}\right) e^{\frac{\theta\zeta_1}{\rho^{(\ell)}\sqrt{n}}}.$$

□

We finish this section with the proof of Lemma 6.12, which is based on [38, Lemma 2.2]. We state this lemma for completeness.

Lemma 6.13. *[38, Lemma 2.2] Suppose Y_1 and Y_2 are random variables such that $|Y_1|$ is stochastically dominated by Y_2 and $\mathbb{E}\left[e^{\theta_{\max}Y_2}\right] < \infty$ for some $\theta_{\max} > 0$. Then, for $\theta^* \in [0, \theta_{\max}]$*

$$\mathbb{E}\left[e^{\theta^* Y_1}\right] \leq 1 + \theta^*\mathbb{E}\left[Y_1\right] + (\theta^*)^2 \sum_{i=2}^{\infty} \frac{(\theta_{\max})^{i-2}}{i!}\mathbb{E}\left[Y_2^i\right].$$

Now we prove Lemma 6.12.

Proof of Lemma 6.12. We use Lemma 6.13 with $Y_1 = \Delta Z(x)$ and $Y_2 = D$. Since D is a constant, we have $\mathbb{E}\left[e^{\theta_{\max}D}\right] < \infty$ for any finite $\theta_{\max}$. For simplicity, we pick $\theta_{\max} = 1$. Then, for any $\theta^* \in [0, 1]$ we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\theta^*\Delta Z(X_k)}|X_k = x\right] \\
&\leq 1 + \theta^*\mathbb{E}\left[\Delta Z(X_k)|X_k = x\right] + (\theta^*)^2 \sum_{i=2}^{\infty} \frac{D^i}{i!} \\
&\overset{(a)}{\leq} 1 + \theta^*\mathbb{E}\left[-\eta\mathbb{1}_{\{Z(X_k)\geq\kappa\}} + D\mathbb{1}_{\{Z(X_k)<\kappa\}}\middle|\, X_k = x\right] + (\theta^*)^2\left(e^D - 1 - D\right)
\end{aligned}$$

$$\overset{(b)}{=} 1 - \theta^*\eta + \theta^*(\eta + D)\mathbb{1}_{\{Z(x)<\kappa\}} + (\theta^*)^2\left(e^D - 1 - D\right). \tag{6.20}$$

where (a) holds by conditions (C1) and (C2), and solving the series in the last term; and (b) holds reorganizing terms.

Now we compute a bound for $\mathbb{E}\left[e^{\theta^* Z(X_{k+1})}\right]$ as follows. Observe

$$\begin{aligned}
&\mathbb{E}\left[e^{\theta^* Z(X_{k+1})}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{\theta^* Z(X_k)} e^{\theta^* \Delta Z(X_k)} \middle| X_k\right]\right] \\
&\overset{(a)}{\leq} \left(1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right)\right)\mathbb{E}\left[e^{\theta^* Z(X_k)}\right] + \theta^*(\eta + D)e^{\theta^*\kappa} \\
&\overset{(b)}{\leq} \left(1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right)\right)^{k+1}\mathbb{E}\left[e^{\theta^* Z(X_0)}\right] \\
&\quad + \theta^*(\eta + D)e^{\theta^*\kappa}\sum_{i=0}^{k}\left(1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right)\right)^i
\end{aligned}$$

where (a) holds by Equation (6.20) and reorganizing terms; and (b) holds using the recursion from the previous line $k+1$ times because the condition $\eta^2 \leq 4(e^D - 1 - D)$ ensures that $\left(1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right)\right) \geq 0$.

The upper bound converges only if $1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right) < 1$, so we solve for θ^* such that

$$1 - \eta\theta^* + (\theta^*)^2\left(e^D - 1 - D\right) \leq 1 - \frac{\eta\theta^*}{2}.$$

We obtain that

$$\theta^* \leq \frac{\eta}{2(e^D - 1 - D)}.$$

For such θ^*, we have

$$\mathbb{E}\left[e^{\theta^* Z(X_{k+1})}\right] \leq \left(1 - \frac{\eta\theta^*}{2}\right)^{k+1}\mathbb{E}\left[e^{\theta^* Z(X_0)}\right] + \theta^*(\eta + D)e^{\theta^*\kappa}\sum_{i=0}^{k+1}\left(1 - \frac{\eta\theta^*}{2}\right)^i$$

Then, taking the limit of the recursion as $k \to \infty$ we obtain the result. $\square$

6.5.3 Proof of Lemma 6.6

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

$$\begin{aligned} \max \quad & \langle \boldsymbol{q}(k), \boldsymbol{x} \rangle \\ s.t. \quad & \boldsymbol{x} \in ConvexHull\left(\mathcal{S}^{(m)}\right) \end{aligned} \tag{6.21}$$

and sets $\boldsymbol{s}(k)$ as one of these optimal solutions. To make the optimization problem linear, we use $ConvexHull\left(\mathcal{S}^{(m)}\right)$ as the feasible region instead of $\mathcal{S}^{(m)}$. However, this does not change the problem because the objective function is linear and, therefore, an optimal solution of Equation 6.21 is at an extreme point, i.e. at a point in $\mathcal{S}^{(m)}$.

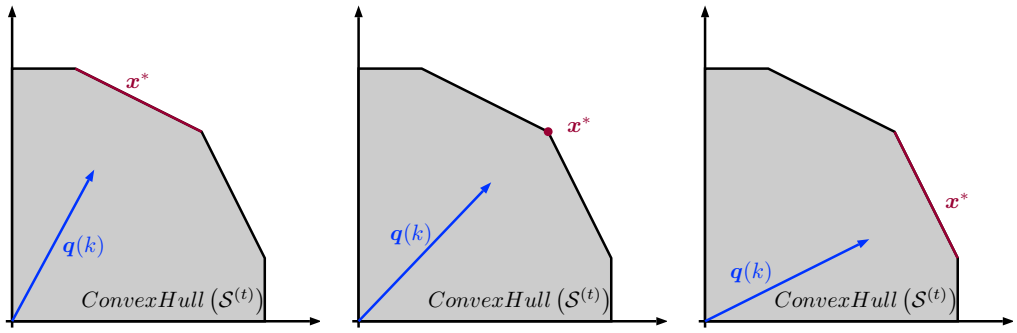
The gradient of the objective function is $\boldsymbol{q}(k)$. Then, depending on its direction, the optimal solution(s) $\boldsymbol{x}^*$ will belong to a different facet or vertex of $ConvexHull\left(\mathcal{S}^{(m)}\right)$. In Figure 6.1 we present pictorial examples where we show the optimal solution(s) when the vector of queue lengths goes in three different directions.

Recall that $\boldsymbol{q}_{\|}(k)$ goes in the same direction as $\boldsymbol{c}^{(\ell)}$. Also, if ϵ is small we expect that $\boldsymbol{q}(k) \approx \boldsymbol{q}_{\|}(k)$ by SSC. Then, if ϵ is small we expect that any optimal solution $\boldsymbol{x}^*$ to the linear program presented in Equation 6.21 satisfies $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}^* \rangle = b^{(m,\ell)}$ with high probability.

Now we present the technical details of the proof.

Proof of Lemma 6.6. First observe that if $\theta = 0$ the proof holds trivially. Now assume $\theta \neq 0$.

We start with a definition. Let $m \in \mathcal{M}$ and suppose that the channel state is $\overline{M} = m$.



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

(b) Example 2: Unique solution.

(c) Example 3: Another example of multiple solutions.

Figure 6.1: Example of optimal solutions depending on the queue lengths vector.

Then, let $\varphi^{(m)} \in \left(0, \frac{\pi}{2}\right]$ be an angle such that $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}} \rangle = b^{(m,\ell)}$ if $\frac{\|\overline{\boldsymbol{q}}_{\|}\|}{\|\overline{\boldsymbol{q}}\|} \geq \cos\left(\varphi^{(m)}\right)$. Let $\varphi_{\boldsymbol{q}}$ be the angle between $\overline{\boldsymbol{q}}_{\|}$ and $\overline{\boldsymbol{q}}$ and define $\varphi_{\min} \stackrel{\triangle}{=} \min_{m\in\mathcal{M}} \varphi^{(m)}$. Therefore, since $\overline{\boldsymbol{s}}$ is scheduled using MaxWeight algorithm, if channel state is m we have

$$b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}} \rangle \quad \text{implies} \quad \varphi^{(m)} < \varphi_{\boldsymbol{q}}. \tag{6.22}$$

In this proof we use the notation $\mathbb{E}_m\left[\,\cdot\,\right] = \mathbb{E}\left[\,\cdot\,|\overline{M} = m\right]$. By definition of conditional expectation we have

$$\begin{aligned}
&\mathbb{E}\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(\overline{B} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}^{(\epsilon)}\rangle\right)} - 1\right)\right] \\
&= \sum_{m\in\mathcal{M}} \psi_m \mathbb{E}_m\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}^{(\epsilon)}\rangle\right)} - 1\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}_m\left[\left(e^{\theta\epsilon\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle} - 1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}^{(\epsilon)}\rangle\right)} - 1\right)\right] \\
&\stackrel{(a)}{=} \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_{\|}\|} - 1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right)} - 1\right)\mathbb{1}_{\left\{b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right\}}\right] \\
&\stackrel{(b)}{\leq} \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_{\|}\|} - 1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right)} - 1\right)\mathbb{1}_{\left\{\varphi_{\boldsymbol{q}} > \varphi^{(m)}\right\}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_\perp\|\cot(\varphi_{\boldsymbol{q}})}-1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right)}-1\right)\mathbb{1}_{\left\{\varphi_{\boldsymbol{q}}>\varphi^{(m)}\right\}}\right]\\
&\overset{(c)}{\leq} \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_\perp\|\cot(\varphi^{(m)})}-1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right)}-1\right)\mathbb{1}_{\left\{\varphi_{\boldsymbol{q}}>\varphi^{(m)}\right\}}\right]\\
&\overset{(d)}{=} \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_\perp\|\cot(\varphi^{(m)})}-1\right)\left(e^{\theta\epsilon\left(b^{(m,\ell)}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right)}-1\right)\right]\\
&\overset{(e)}{\leq} \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_\perp\|\cot(\varphi^{(m)})}-1\right)^{j}\right]^{\frac{1}{j}}\mathbb{E}_m\left[\left(e^{\theta\epsilon\left(b^{(m,\ell)}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right)}-1\right)^{\frac{j}{j-1}}\right]^{\frac{j-1}{j}},
\end{aligned}
$$

where $j \in \mathbb{Z}_+$ satisfies $j > 1$. Here (a) holds by definition of indicator function and because $\overline{\boldsymbol{q}}_\| = \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle \boldsymbol{c}^{(\ell)}$ by definition of projection; (b) holds by Equation 6.22; (c) and (d) hold because $\cot(\varphi)$ is decreasing for $\varphi \in \left(0, \frac{\pi}{2}\right]$; (d) holds by Equation 6.22 and by definition of indicator function; and (e) holds by Hölder's inequality.

Using an argument similar to the one at the end of Lemma 3.9, it can be proved that

$$
0 \leq \mathbb{E}_m\left[\left(e^{\theta\epsilon\|\overline{\boldsymbol{q}}_\perp\|\cot(\varphi^{(m)})}-1\right)^{j}\right]^{\frac{1}{j}}
$$

converges to a constant as $\epsilon \downarrow 0$. On the other hand,

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

where $\overline{B}_{\max} = \max_{m\in\mathcal{M}} b^{(m,\ell)}$. In [34] , the authors prove that $\mathbb{P}\left[b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] = K\epsilon$ for a finite constant K, and their proof also holds here. Therefore,

$$
\mathbb{E}\left[\left(e^{\theta\epsilon\left(b^{(m,\ell)}-\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\right)}-1\right)^{\frac{j}{j-1}}\right] \text{ is } O\left(\epsilon^{1+\frac{j}{j-1}}\right)
$$

This completes the proof. □

6.5.4 Proof of Claim 6.8

Proof of Claim 6.8. We have

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

where (a) holds because $\overline{u}_i \leq \overline{s}_i \leq S_{\max}$ and $\boldsymbol{c}^{(\ell)} \geq 0$; and (b) holds by Lemma 6.7, because $\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle - \overline{B}\right] \leq 0$.

Therefore,

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

□

6.5.5 Proof of Claim 6.9

Proof of Claim 6.9. For the first expression, from Lemma 3.1 we have

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

where the last equality holds by Equation 6.7. Also,

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

where (a) holds because $\overline{a}_i \leq A_{\max}$ with probability 1 for all $i \in [n]$, $\boldsymbol{c}^{(\ell)} \geq 0$, $\overline{B}$ is bounded by a constant that we denote $B_{\max}$ and because all quantities are nonnegative; and (b) holds by Equation 6.7. Then,

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

Therefore,

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

This proves the first equation of the claim.

For the second expression, using Lemma 3.1 we obtain

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

But

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

where (a) holds because $\overline{s}_i \leq S_{\max}$ with probability 1 for all $i \in [n]$, $\boldsymbol{c}^{(\ell)} \geq 0$, $\overline{B} \leq B_{\max}$ and all quantities are nonnegative (see Equation 6.14 to see why $\mathbb{E}\left[\overline{B} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] \geq 0$); and (b) holds by Lemma 6.7 and because $\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] \geq 0$ since $\overline{\boldsymbol{u}} \geq 0$ and $\boldsymbol{c}^{(\ell)} \geq 0$.

Then,

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

Therefore,

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

This completes the proof. □

6.6 Conclusion and future work

In this chapter we presented the last SPN that we analyze with the transform method in this thesis. As mentioned before, the transform method is a simple and flexible approach to compute the distribution of the scaled heavy-traffic queue lengths. Future work is to generalize this method to other queueing systems, including SPNs that do not satisfy the CRP condition.

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 [34, 14, 15]. 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 (see chapter 7), because of the following reason. The queue lengths process is a multi-dimensional DTMC (or a CTMC 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 $\mathbb{E}\left[\epsilon \sum_i \overline{q}_i\right]$. If P is the transition matrix, then the stationary distribution is a unique solution of the equation, $\pi = \pi P$. Clearly, solving for the stationary distribution in general is hard. However, we know that it is unique and is characterized by this equation. If we take two-sided Laplace transform of the equation $\pi = \pi P$ 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 , even for the single server queue it is challenging to obtain a solution for this equation in all traffic. 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 ϵ goes to zero. Tightness follows directly from the bound on $E[\epsilon \sum_i \overline{q}_i]$ that one obtains 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 in [14, 15, 22] from such an equation.

CHAPTER 7

HEAVY-TRAFFIC ANALYSIS WITH NO COMPLETE RESOURCE POOLING

Based on:

D. Hurtado-Lange and S. T. Maguluri, "Heavy-traffic analysis of queueing systems with no complete resource pooling," *arXiv preprint arXiv:1904.10096*, 2019

D. Hurtado-Lange, S. M. Varma, and S. T. Maguluri, "Logarithmic heavy traffic error bounds in generalized switch and load balancing systems," *arXiv preprint arXiv:2003.07821*, 2020

7.1 Introduction

In this chapter we study a generalized switch without assuming that the CRP condition is satisfied, and so SSC may occur to a multidimensional subspace. Also, we assume that the arrival process to each queue is a sequence of independent and identically distributed (i.i.d.) random variables, but we do not require that these sequences are independent of each other. The main contributions of this chapter are:

(i) In Theorem 7.5 we characterize the heavy-traffic scaled mean of certain linear combinations of the queue lengths in steady state under the MaxWeight algorithm. Moreover, we obtain lower and upper bounds that are valid in all regimes (not necessarily heavy traffic), but are tight in the heavy-traffic regime. This result is immediately applicable in several systems, as we showcase in section 7.5, and it includes both, the CRP and the non-CRP cases. Little is known about SPNs that do not satisfy CRP, since the most common approach in the literature is the use of diffusion limits, and

solving a multidimensional RBM is an open question. In this chapter we contribute to understanding the heavy-traffic behavior of non-CRP systems by providing the mean of some linear combinations of the queue lengths.

(ii) In Corollary 7.8 we compute the heavy-traffic limit of the total queue length in an input-queued switch with correlated arrivals. As mentioned above, the input-queued switch has had considerable attention in the literature. However, it has only been studied under independent arrival processes [14, 15]. The input-queued switch is a model for an ideal data center network, and independent arrivals is an unrealistic assumption in this setting. In fact,data centers experience hot-spots, and, hence, the arrivals to different queues are highly correlated [100, 101].

(iii) We illustrate how Theorem 7.5 can be immediately applied to a variety of systems. Specifically, we show how to apply it to parallel-server systems (Corollary 7.13, and Corollary 7.15), the so-called $\mathcal{N}$-system (Corollary 7.16) and ad hoc wireless networks (Corollary 7.11).

(iv) In Section 7.5.2 we show that, if SSC is full-dimensional, then the heavy-traffic limit of the mean queue lengths does not depend on the correlation among arrival processes. In other words, if the systems experience full-dimensional SSC, the expected linear combination of queue lengths behaves as if the queues were independent. This result is rather surprising, and it was not known.

(v) In Theorem 7.24 we show that using the drift method with polynomial test functions, it is impossible to obtain the moments of all linear combinations of the queue lengths. We prove this result by presenting an alternate way of thinking of the drift method. Traditionally, the key step in using this approach, is to design the correct test function to obtain all the moments. However, it is not clear a priori if there are test functions that give all these moments. 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 is to obtain more equations using the constraints in the system, in order to solve for all the unknowns and obtain the complete joint distribution of the queue lengths when the CRP condition is not satisfied.

(vi) In Theorem 7.27 we obtain lower and upper bounds on the steady-state mean of an arbitrary linear combination of queue lengths. We do this by formulating a Linear Program (LP) using the under-determined system of equations from Theorem 7.24. We present numerical results in the case of Bernoulli arrivals, for different values of the traffic intensity. For simplicity of exposition, we do this only in the special case of an input-queued switch, and the same approach can be used for the generalized switch. We discuss this generalization in section 4.5

7.2 Related work

In section 6.2 we discussed the related work on MaxWeight algorithm.

The study of SPNs that do not satisfy the CRP condition is relatively new. The most relevant literature for this chapter is [34, 14, 15]. In [34], the authors develop the drift method and they apply it to the generalized switch under the CRP condition. In [14, 15] an input-queued switch that does not satisfy the CRP condition is studied using the drift method too.

The results in this chapter are clearly a generalization of the work of [34, 14, 15], but due to the generality of the model, several challenges arise.

We use the drift method to analyze the heavy-traffic behavior of the mean queue lengths in a generalized switch. The drift method is a two-step procedure to compute bounds on linear combinations of the queue lengths that are tight in heavy traffic. The first step is to prove SSC, which we do in Proposition 7.4, and the second step is to set to zero the drift of $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|\mathcal{H}}\|^2$,i.e., of the squared norm of the projection of the vector of queue lengths on the subspace where SSC occurs. While these steps are standard for the drift method

as developed in [34, 14, 15, 22], different challenges arise in each case depending on the system that one is studying. In this case we are working with the generalized switch, which is a very general model. Hence, we overcome difficulties that were not part of the work listed above. We summarize these below:

(a) Since the effective capacity region is the average of several individual capacity regions (see the definition of the capacity region in Equation 6.1), the vector of potential service does not necessarily belong to the effective capacity region. Then, it is not obvious how to deal with the terms that involve the service vector.

(b) In the case of an input-queued switch as studied in [14], the projected service vector $\overline{\boldsymbol{s}}_{\|\mathcal{H}}$ is constant due to the structure of the system. In the case of the generalized switch, this is not the case, and this leads to significant challenges. In particular, the computation of the term $\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}}_{\|\mathcal{H}} \rangle\right]$ is not trivial. We used the properties of the system and MaxWeight algorithm to bound this term.

(c) The final closed-form expression that we obtain for the steady-state expectation of the queue lengths is novel, and is a contribution by itself. To compute this expression (the term $\mathcal{T}_2$ in the proof of Theorem 7.5), we use the least squares problem to obtain an expression that is valid for any generalized switch. In [34, 14, 15] the underlying symmetry of the specific systems that were studied is explicitly used in the computations, and therefore, it is not clear how to generalize.

Challenge (a) is addressed in Lemmas 7.1 and 7.2, that we prove in the next section. These lemmas form an important part of the entire proof and are used repeatedly. Challenge (b) is addressed in 7.21. Finally, overcoming challenge (c) using the least square problem gave us the closed form expression for the right-hand side in Theorem 7.5.

7.3 Useful lemmas

In this section we present preliminary results that form the base of the analysis of the generalized switch. We use these results repeatedly in this chapter and, hence, we present them below.

Recall that $\overline{B}_\ell \stackrel{\triangle}{=} b^{(\overline{M},\ell)}$ and that, for each $m \in \mathcal{M}$, $b^{(m,\ell)}$ is the maximum $\boldsymbol{c}^{(\ell)}$-weighted service rate in $\mathcal{S}^{(m)}$. Similarly, $b^{(\ell)}$ can be interpreted as the maximum $\boldsymbol{c}^{(\ell)}$-weighted service rate in $\mathcal{C}$, and $\boldsymbol{c}^{(\ell)}$ and $b^{(\ell)}$ define a facet of $\mathcal{C}$. Hence, since the values of $\overline{B}_\ell$ occur according to the probability mass function of the channel state, and the capacity region $\mathcal{C}$ can be interpreted as the 'expected capacity region' according to Equation 6.1, we should expect $\mathbb{E}\left[\overline{B}_\ell\right] = b^{(\ell)}$. Additionally, $\boldsymbol{c}^{(\ell)}$ and $b^{(m,\ell)}$ define a half-space that passes through the boundary of ConvexHull$(\mathcal{S}^{(m)})$ and, hence, there must exist a vector $\boldsymbol{\nu}^{(m)}$ such that $b^{(m,\ell)} = \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)} \rangle$. We formalize these results in Lemma 7.1.

Lemma 7.1. *Let $\ell \in P$ and $m \in \mathcal{M}$. Then, there exists $\boldsymbol{\nu}^{(m)} \in \mathcal{S}^{(m)}$ such that $b^{(m,\ell)} = \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)} \rangle$. This implies that $b^{(\ell)} = \mathbb{E}\left[\overline{B}_\ell\right]$ for all $\ell \in P$.*

The proof of Lemma 7.1 follows immediately from the definition of the capacity region $\mathcal{C}$ in Equation 6.1, and of the parameters $b^{(m,\ell)}$ in Equation 6.4. We present the details below.

Proof of Lemma 7.1. First, recall $\boldsymbol{\nu} \in \mathcal{C}$ and $\mathcal{C}$ is a weighted sum of ConvexHull$(\mathcal{S}^{(m)})$ for $m \in \mathcal{M}$. Then, since each $\mathcal{S}^{(m)}$ is finite, for each $m \in \mathcal{M}$ there exists $\boldsymbol{\nu}^{(m)} \in \mathcal{S}^{(m)}$ such that

$$\boldsymbol{\nu} = \sum_{m \in \mathcal{M}} \psi_m \boldsymbol{\nu}^{(m)}$$

Also, by definition of the ℓ^{th} hyperplane, for each $\ell \in P$ we have

$$b^{(\ell)} = \max_{\boldsymbol{x} \in \mathcal{C}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle$$

$$
\begin{aligned}
&= \sum_{m\in\mathcal{M}} \psi_m \max\left\{\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle : \boldsymbol{x} \in ConvexHull\left(\mathcal{S}^{(m)}\right)\right\} \\
&\overset{(a)}{=} \sum_{m\in\mathcal{M}} \psi_m \max_{\boldsymbol{x}\in\mathcal{S}^{(m)}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle \\
&\overset{(b)}{=} \sum_{m\in\mathcal{M}} \psi_m b^{(m,\ell)}
\end{aligned}
$$

where (a) holds because the objective function of the maximization is linear and, therefore, the optimal solution is an extreme point of ConvexHull$\left(\mathcal{S}^{(m)}\right)$, which must be an element of $\mathcal{S}^{(m)}$ by definition of convex hull; and (b) holds by definition of $b^{(m,\ell)}$. This proves that $b^{(\ell)} = \mathbb{E}\left[\overline{B}_\ell\right]$.

Observe that the last equality also implies that $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle = b^{(m,\ell)}$ for all $m \in \mathcal{M}$, for the following reason. First, by definition of $b^{(m,\ell)}$ we know $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle \leq b^{(m,\ell)}$. Also, if there exists $m^* \in \mathcal{M}$ with $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m^*)}\rangle < b^{(m^*,\ell)}$, then

$$
\sum_{m\in\mathcal{M}} \psi_m \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle < \sum_{m\in\mathcal{M}} \psi_m b^{(m,\ell)}.
$$

But

$$
\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle = \sum_{m\in\mathcal{M}} \psi_m \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle \quad \text{and} \quad b^{(\ell)} = \sum_{m\in\mathcal{M}} \psi_m b^{(m,\ell)}.
$$

Therefore, we got a contradiction because $\boldsymbol{\nu} \in \bigcap_{\ell\in P} \mathcal{F}^{(\ell)}$ and, hence, $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle = b^{(\ell)}$. $\square$

As ϵ gets closer to zero, we know that $\boldsymbol{\lambda}^{(\epsilon)}$ gets closer to $\boldsymbol{\nu}$, and SSC implies that the vector of queue lengths can be approximated by its projection on $\mathcal{K}$. In other words, as $\epsilon \downarrow 0$ the vector of queue lengths can be well approximated by a conic combination of the vectors $\boldsymbol{c}^{(\ell)}$ with $\ell \in P$. Therefore, since the scheduling problem is solved using MaxWeight algorithm, and given that the channel state is m, one should expect that $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle = b^{(m,\ell)}$ with high probability. In the next lemma we formalize this intuition.

Lemma 7.2. *For each $m \in \mathcal{M}$ and $\ell \in P$, define $\pi^{(m,\ell)} \triangleq \mathbb{P}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle = b^{(m,\ell)} \;\middle|\; \overline{M} = m\right]$.*

Then, $1 - \pi^{(m,\ell)}$ *is* $O(\epsilon)$.

The proof of Lemma 7.2 is a generalization of [34, Claim 1], and we present below for completeness.

Proof of Lemma 7.2. For ease of exposition, we omit the dependence on ϵ of the variables, and we use the notation $\mathbb{E}_m[\cdot] \stackrel{\triangle}{=} \mathbb{E}\left[\cdot|\overline{M} = m\right]$. Define

$$\gamma^{(m)} \stackrel{\triangle}{=} \min\left\{b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle : \ \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle < b^{(m,\ell)}, \text{ for } \ell \in P, \, \boldsymbol{x} \in \mathcal{S}^{(m)}\right\}.$$

Observe that, for each $m \in \mathcal{M}$ we have $\gamma^{(m)} > 0$ because each $\mathcal{S}^{(m)}$ is a finite set and, therefore, $b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle$ cannot be arbitrarily close to zero.

From stability, for each $\ell \in P$ we have

$$\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] \geq \mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle\right].$$

Then, using Lemma 7.1, we obtain that for each $m \in \mathcal{M}$

$$\mathbb{E}_m\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] \geq \mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle\right] = (1-\epsilon)\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle = (1-\epsilon)b^{(m,\ell)}$$

On the other hand, by definition of $\pi^{(m,\ell)}$ we have

$$\mathbb{E}_m\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] = \pi^{(m,\ell)} b^{(m,\ell)} + \left(1 - \pi^{(m,\ell)}\right) \mathbb{E}_m\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \middle| \, b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right]$$

Putting these two results together we obtain

$$(1-\epsilon)b^{(m,\ell)} \leq b^{(m,\ell)}\pi^{(m,\ell)} + \left(1 - \pi^{(m,\ell)}\right) \mathbb{E}_m\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \ \middle| \ b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] \tag{7.1}$$

Also, by definition of $\gamma^{(m)}$ we have

$$\mathbb{E}_m \left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}} \rangle \;\middle|\; b^{(m,\ell)} \neq \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}} \rangle \right] \leq b^{(m,\ell)} - \gamma^{(m)}.$$

Using the last result in Equation 7.1 and rearranging terms we obtain

$$1 - \pi^{(m,\ell)} \leq \frac{\epsilon b^{(m,\ell)}}{\gamma^{(m)}}.$$

Since the sets P and $\mathcal{S}^{(m)}$ for each $m \in \mathcal{M}$ are finite, there exists $b_{\max} = \max_{m \in \mathcal{M},\, \ell \in P} \{b^{(m,\ell)}\}$ and $b_{\max} < \infty$. Therefore, we have $1 - \pi^{(m,\ell)}$ is $O(\epsilon)$.

□

7.4 Heavy-traffic analysis of the generalized switch.

In this section we perform heavy-traffic analysis of the generalized switch. Before presenting the details we specify the heavy-traffic parametrization, which is similar to the parametrization we used in Chapter 6. We fix a vector $\boldsymbol{\nu}$ in the boundary of $\mathcal{C}$ and we consider a set of generalized switches operating under MaxWeight as described above, parametrized by $\epsilon \in (0, 1)$. The heavy-traffic limit is the limit as $\epsilon \downarrow 0$ and, as ϵ gets small, the vector of mean arrival rates approaches $\boldsymbol{\nu}$. Formally, we parametrize the queueing system in the following way. We let $\boldsymbol{q}^{(\epsilon)}(k)$, $\boldsymbol{a}^{(\epsilon)}(k)$, $\boldsymbol{s}^{(\epsilon)}(k)$ and $\boldsymbol{u}^{(\epsilon)}(k)$ be the vectors of queue lengths, arrivals, potential service and unused service, respectively, in time slot k, in the system parametrized by ϵ. The parametrization is such that the vector of mean arrival rate is $\boldsymbol{\lambda}^{(\epsilon)} \stackrel{\triangle}{=} \mathbb{E}\left[\boldsymbol{a}^{(\epsilon)}(1)\right] = (1 - \epsilon)\boldsymbol{\nu}$. Therefore, $\boldsymbol{\lambda}^{(\epsilon)}$ belongs to the interior of $\mathcal{C}$ for each $\epsilon \in (0, 1)$ and, as $\epsilon \downarrow 0$, the arrival rate vector $\boldsymbol{\lambda}^{(\epsilon)}$ approaches the boundary of the capacity region at $\boldsymbol{\nu}$.

Heavy-traffic analysis of the generalized switch has been performed in the past, using the diffusion limits approach [8], and the Drift method [34]. However, in both cases, the analysis is under the assumption that SSC occurs into a one-dimensional subspace (CRP

condition), i.e., when the vector $\boldsymbol{\nu}$ is in the interior of a facet of the capacity region $\mathcal{C}$. In this chapter, we focus on cases where the vector $\boldsymbol{\nu}$ may live at the intersection of facets. Define $P \stackrel{\triangle}{=} \left\{\ell \in [L] : \boldsymbol{\nu} \in \mathcal{F}^{(\ell)}\right\}$, that is, P is the set of indices of all the facets that intersect at $\boldsymbol{\nu}$. Observe that, if P has only one element, we are under the CRP condition, and our results in this case agree with the results proved by [34]. In this chapter, we focus on the case where P is allowed to have more than one element.

For each $\epsilon \in (0,1)$, let $\overline{\boldsymbol{q}}^{(\epsilon)}$ be a steady-state random vector such that the Markov chain $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ converges in distribution to $\overline{\boldsymbol{q}}^{(\epsilon)}$ as $k \uparrow \infty$. Since MaxWeight is throughput optimal, the Markov chain $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent for each $\epsilon \in (0,1)$, so $\overline{\boldsymbol{q}}^{(\epsilon)}$ is well defined. Let $\overline{\boldsymbol{a}}^{(\epsilon)}$ be a steady-state vector which is equal in distribution to $\boldsymbol{a}^{(\epsilon)}(1)$. Then, $\mathbb{E}\left[\overline{\boldsymbol{a}}^{(\epsilon)}\right] = \boldsymbol{\lambda}^{(\epsilon)}$ and for each $i \in [n]$ we have $\overline{a}_i^{(\epsilon)} \leq A_{\max}$ with probability 1. Let $\Sigma_a^{(\epsilon)}$ be the covariance matrix of the vector $\overline{\boldsymbol{a}}^{(\epsilon)}$. Let $\overline{M}$ and $\overline{B}_\ell$ be steady-state random variables that are equal in distribution to $M(1)$ and $B_\ell(1)$ for each $\ell \in [L]$, respectively. Let $\overline{\boldsymbol{s}}^{(\epsilon)} \stackrel{\triangle}{=} \boldsymbol{s}(\overline{\boldsymbol{q}}^{(\epsilon)}, \overline{M})$ be the vector of potential service in steady-state, and $\overline{\boldsymbol{u}}^{(\epsilon)} \stackrel{\triangle}{=} \boldsymbol{u}(\overline{\boldsymbol{q}}^{(\epsilon)}, \overline{M}, \overline{\boldsymbol{a}}^{(\epsilon)})$ be the vector of unused service. Define $\left(\overline{\boldsymbol{q}}^{(\epsilon)}\right)^+ \stackrel{\triangle}{=} \overline{\boldsymbol{q}}^{(\epsilon)} + \overline{\boldsymbol{a}}^{(\epsilon)} - \overline{\boldsymbol{s}}^{(\epsilon)} + \overline{\boldsymbol{u}}^{(\epsilon)}$ as the vector of queue lengths one time slot after $\overline{\boldsymbol{q}}^{(\epsilon)}$ is observed, given that the vectors of arrivals and potential service are $\overline{\boldsymbol{a}}^{(\epsilon)}$ and $\overline{\boldsymbol{s}}^{(\epsilon)}$, respectively.

In Section 7.4.2 we prove that the state space collapses into the cone $\mathcal{K}$ described below. In other words, we show that the vector of queue lengths can be approximated by a vector in $\mathcal{K}$ in heavy traffic. Let $\mathcal{K}$ be the cone generated by $\left\{\boldsymbol{c}^{(\ell)} : \ell \in P\right\}$ and $\mathcal{H}$ be the subspace generated by the same set of vectors. Formally,

$$\mathcal{K} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^n : \ \boldsymbol{x} = \sum_{\ell \in P} \xi_\ell \boldsymbol{c}^{(\ell)} \ , \ \xi_\ell \geq 0 \ \forall \ell \in P \right\}. \tag{7.2}$$

A pictorial example of the capacity region $\mathcal{C}$ and the cone $\mathcal{K}$ when $n = 3$ is presented in Figure 7.1. Let $\tilde{P} \subset P$ be a set of indices such that the set $\left\{\boldsymbol{c}^{(\ell)} : \ell \in \tilde{P}\right\}$ is linearly independent, and let $C = \left[\boldsymbol{c}^{(\ell)}\right]_{\ell \in \tilde{P}}$ be a matrix where the columns are a linearly independent

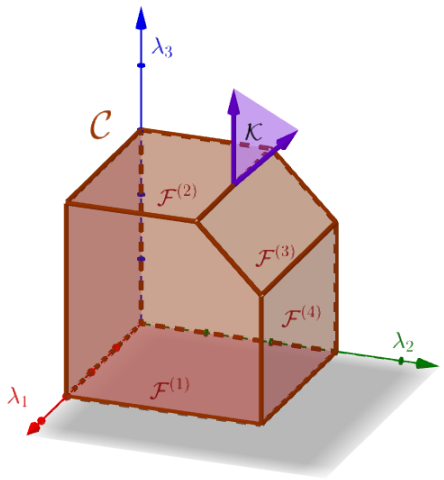


Figure 7.1: Example of capacity region $\mathcal{C}$ and cone $\mathcal{K}$.

subset of the vectors that generate the cone $\mathcal{K}$. Observe that the column space of the matrix C is exactly the subspace $\mathcal{H}$.

In subsection 7.4.1 we present a Universal Lower Bound (ULB), that is independent of the scheduling policy; in subsection 7.4.2 we present the SSC result formally; and in subsection 7.4.3 we present the main result of this chapter (Theorem 7.5), where we compute asymptotically tight bounds on linear combinations of the queue lengths.

7.4.1 Universal lower bound

In this section we compute a ULB for certain linear combinations of the vector of queue lengths. The bound is universal in the sense that it remains valid for all scheduling policies.

Proposition 7.3. *Consider a generalized switch parametrized by $\epsilon \in (0,1)$, as described in at the beginning of this section. Let $\boldsymbol{z} \in \mathcal{K}$ with $\boldsymbol{z} \neq \mathbf{0}$ and $\boldsymbol{r} \in \mathbb{R}_+^{|P|}$ be such that $\boldsymbol{z} = \sum_{\ell=1}^{L} r_\ell \boldsymbol{c}^{(\ell)}$. Then, for each $\epsilon \in (0,1)$ we have*

$$\mathbb{E}\left[\langle \boldsymbol{z}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle\right] \geq \frac{1}{2\epsilon \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle} \left(\boldsymbol{z}^T \Sigma_a^{(\epsilon)} \boldsymbol{z} + \boldsymbol{r}^T \Sigma_B \boldsymbol{r} \right) - f(\epsilon),$$

where $f(\epsilon) = \frac{b_{\max}\langle \mathbf{1}, \boldsymbol{r} \rangle}{2} - \frac{\epsilon \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle}{2}$ is $o\left(\frac{1}{\epsilon}\right)$ (i.e., $\lim_{\epsilon \downarrow 0} \epsilon f(\epsilon) = 0$) and $b_{\max} = \max_{m \in \mathcal{M}, \ell \in P} b^{(m,\ell)}$.

The proposition is proved by coupling the queue length vector of the generalized switch

with a single server queue $\left\{\Phi^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ constructed as follows. We let $\alpha^{(\epsilon)}(k) \triangleq \langle \boldsymbol{z}, \boldsymbol{a}^{(\epsilon)}(k)\rangle$ be the number of arrivals in time slot k and $\beta(k)$ be the potential service, where $\mathbb{P}\left[\beta(k) = \sum_{\ell \in P} r_\ell b^{(m,\ell)}\right] = \psi_m$ for each $m \in \mathcal{M}$. Then, it is easy to see that $\Phi^{(\epsilon)}(k)$ is stochastically smaller than $\langle \boldsymbol{z}, \boldsymbol{q}^{(\epsilon)}(k)\rangle$ (by definition of $b^{(m,\ell)}$ in Equation 6.4). Therefore, a lower bound to the expected value of $\Phi^{(\epsilon)}(k)$ in steady state is also a lower bound to $\mathbb{E}\left[\langle \boldsymbol{z}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle\right]$. The last step in the proof is to compute such lower bound, which we do by setting to zero the drift of $V_{ULB}(\Phi) = \Phi^2$. Note that it is essential that the weights r_ℓ are nonnegative to obtain a lower bound in the proof. This is the reason why $\boldsymbol{z} \in \mathcal{H}$ is not enough and we require $\boldsymbol{z} \in \mathcal{K}$. The rest of the proof is presented below.

Proof of Proposition 7.3. Let $\chi^{(\epsilon)}(k)$ the unused service in time slot k. We assume that in each time slot, arrivals occur before service. Then, for each $k \in \mathbb{Z}_+$ we have

$$\Phi^{(\epsilon)}(k+1) = \Phi^{(\epsilon)}(k) + \alpha^{(\epsilon)}(k) - \beta(k) + \chi^{(\epsilon)}(k). \tag{7.3}$$

Before computing the lower bound we need verify that the DTMC $\left\{\Phi^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent for each $\epsilon \in (0,1)$. To do that we show that $\mathbb{E}\left[\beta(k) - \alpha^{(\epsilon)}(k)\right] > 0$ for all $\epsilon \in (0,1)$. By definition of $\alpha^{(\epsilon)}(k)$ we have

$$\mathbb{E}\left[\alpha^{(\epsilon)}(k)\right] = \mathbb{E}\left[\langle \boldsymbol{z}, \boldsymbol{a}^{(\epsilon)}(k)\rangle\right] = (1-\epsilon)\langle \boldsymbol{z}, \boldsymbol{\nu}\rangle \tag{7.4}$$

where the last equality holds because $\mathbb{E}\left[\overline{\boldsymbol{a}}^{(\epsilon)}\right] = (1-\epsilon)\boldsymbol{\nu}$. By definition of $\beta(k)$ we have

$$\begin{aligned}
\mathbb{E}\left[\beta(k)\right] &= \sum_{\ell \in P} r^{(\ell)} \sum_{m \in \mathcal{M}} \psi_m b^{(m,\ell)} \\
&\overset{(a)}{=} \sum_{\ell \in P} r_\ell b^{(\ell)} \\
&\overset{(b)}{=} \sum_{\ell \in P} r_\ell \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle \\
&\overset{(c)}{=} \langle \boldsymbol{z}, \boldsymbol{\nu}\rangle,
\end{aligned} \tag{7.5}$$

where (a) holds because $b^{(\ell)} = \mathbb{E}\left[B_\ell(1)\right] \sum_{m \in \mathcal{M}} \psi_m b^{(m,\ell)}$ (as we proved Lemma 7.1); (b) holds because $\boldsymbol{\nu} \in \mathcal{F}^{(\ell)}$ for all $\ell \in P$; and (c) holds by definition of $\boldsymbol{z}$. Then, from Equation 7.4 and Equation 7.5 we obtain

$$\mathbb{E}\left[\beta(k) - \alpha^{(\epsilon)}(k)\right] = \epsilon \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle,$$

which is a positive number. Let $\overline{\Phi}^{(\epsilon)}$ be a steady-state vector which is limit in distribution of $\left\{\Phi^{(\epsilon)}(k) : k \geq 1\right\}$, and $\left(\overline{\Phi}^{(\epsilon)}\right)^+ \stackrel{\triangle}{=} \overline{\Phi}^{(\epsilon)} + \overline{\alpha}^{(\epsilon)} - \overline{\beta}^{(\epsilon)} + \overline{\chi}^{(\epsilon)}$, where $\overline{\alpha}^{(\epsilon)}$ and $\overline{\beta}^{(\epsilon)}$ are steady-state random variables with the distribution of $\alpha^{(\epsilon)}(1)$ and $\beta^{(\epsilon)}(1)$, respectively and $\overline{\chi}^{(\epsilon)}$ represents the unused service.

Now we show the result. We omit the dependence on ϵ in the rest of this proof, for ease of exposition. It can be easily proved that $\mathbb{E}\left[\overline{\Phi}^2\right] < \infty$ (e.g., we can use Lemma 2.7), and we omit the proof for brevity. Then, we set to zero the drift of $V(\Phi) = \Phi^2$ in steady state, and we obtain

$$\begin{aligned}
0 =& \mathbb{E}\left[\left(\overline{\Phi}^+\right)^2 - \overline{\Phi}^2\right] \\
=& \mathbb{E}\left[\left(\overline{\Phi}^+ - \overline{\chi}\right)^2 + \overline{\chi}^2 + 2\left(\overline{\Phi}^+ - \overline{\chi}\right)\overline{\chi} - \overline{\Phi}^2\right] \\
\stackrel{(a)}{=}& \mathbb{E}\left[\left(\overline{\Phi} + \overline{\alpha} - \overline{\beta}\right)^2 - \overline{\chi}^2 - \overline{\Phi}^2\right] \\
\stackrel{(b)}{=}& \mathbb{E}\left[\overline{\alpha}^2\right] + \mathbb{E}\left[\overline{\beta}^2\right] - 2\mathbb{E}\left[\overline{\alpha}\right]\mathbb{E}\left[\overline{\beta}\right] - 2\mathbb{E}\left[\overline{\Phi}\right]\mathbb{E}\left[\overline{\alpha} - \overline{\beta}\right] - \mathbb{E}\left[\overline{\chi}\right]
\end{aligned} \tag{7.6}$$

where (a) holds after expanding the product, and because $\overline{\Phi}^+\overline{\chi} = 0$ by definition of unused service; and (b) holds after expanding the product and using independence of the arrival, service and queue length processes.

We compute the terms in Equation 7.6 one by one. We already established that $\mathbb{E}\left[\overline{\alpha}\right] = (1-\epsilon)\langle \boldsymbol{z}, \boldsymbol{\nu} \rangle$ and $\mathbb{E}\left[\overline{\beta}\right] = \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle$. Now we compute the quadratic terms. By definition of

$\overline{\alpha}$ we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\alpha}^2\right] =& \mathbb{E}\left[\left(\sum_{i=1}^{n} z_i \overline{a}_i\right)^2\right] \\
=& \sum_{i=1}^{n}\sum_{i'=1}^{n} z_i z_{i'} \mathbb{E}\left[\overline{a}_i \overline{a}_{i'}\right] \\
\overset{(a)}{=}& \sum_{i=1}^{n}\sum_{i'=1}^{n} z_i z_{i'} \text{Cov}\left[\overline{a}_i, \overline{a}_{i'}\right] + \sum_{i=1}^{n}\sum_{i'=1}^{n} z_i z_{i'} \mathbb{E}\left[\overline{a}_i\right] \mathbb{E}\left[\overline{a}_{i'}\right] \\
\overset{(b)}{=}& \boldsymbol{z}^T \Sigma_a^{(\epsilon)} \boldsymbol{z} + (1-\epsilon)^2 \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle^2
\end{aligned}$$

where (a) holds by definition of covariance; and (b) holds by definition of covariance matrix and because $\mathbb{E}\left[\overline{\boldsymbol{a}}\right] = \boldsymbol{\lambda}^{(\epsilon)} = (1-\epsilon)\boldsymbol{\nu}$. For the service process, by definition of covariance matrix we obtain

$$\mathbb{E}\left[\overline{\beta}^2\right] = \boldsymbol{r}^T \Sigma_B \boldsymbol{r} + \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle^2,$$

where the last equality holds because $\boldsymbol{\nu} \in \mathcal{F}^{(\ell)}$ for all $\ell \in P$ and by definition of $\boldsymbol{z}$. For the last term we compute an upper bound. By definition of unused service, we have $\overline{\chi} \leq \overline{\beta}$ with probability 1. Then,

$$\begin{aligned}
\mathbb{E}\left[\overline{\chi}^2\right] \leq& \mathbb{E}\left[\overline{\beta}\overline{\chi}\right] \\
\overset{(a)}{\leq}& b_{\max}\left(\sum_{\ell \in P} r^{(\ell)}\right) \mathbb{E}\left[\overline{\chi}\right] \\
\overset{(b)}{=}& b_{\max}\left(\sum_{\ell \in P} r^{(\ell)}\right) \mathbb{E}\left[\overline{\beta} - \overline{\alpha}\right] \\
\overset{(c)}{=}& \epsilon \langle \boldsymbol{z}, \boldsymbol{\nu} \rangle b_{\max}\left(\sum_{\ell \in P} r^{(\ell)}\right),
\end{aligned}$$

where (a) holds by definition of $\overline{\beta}$ and $b_{\max}$; (b) holds because $\mathbb{E}\left[\overline{\chi}\right] = \mathbb{E}\left[\overline{\beta} - \overline{\alpha}\right]$, which can be easily proved by setting to zero the drift of $V_l(\Phi) = \Phi$; and (c) holds by Equation 7.4

and Equation 7.5.

Putting everything together in Equation 7.6 and rearranging terms we obtain the result.

□

7.4.2 State space collapse.

We prove SSC into the cone $\mathcal{K}$ defined in Equation 7.2 in heavy traffic. We start introducing the notation. For each $\epsilon \in (0,1)$, let $\boldsymbol{q}^{(\epsilon)}_{\|\mathcal{K}}(k)$ be the projection of $\boldsymbol{q}^{(\epsilon)}(k)$ on $\mathcal{K}$ and $\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{K}}(k) \stackrel{\triangle}{=} \boldsymbol{q}^{(\epsilon)}(k) - \boldsymbol{q}^{(\epsilon)}_{\|\mathcal{K}}(k)$. Similarly, define $\boldsymbol{q}^{(\epsilon)}_{\|\mathcal{H}}(k)$ as the projection of $\boldsymbol{q}^{(\epsilon)}(k)$ on $\mathcal{H}$ and $\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{H}}(k) \stackrel{\triangle}{=} \boldsymbol{q}^{(\epsilon)}(k) - \boldsymbol{q}^{(\epsilon)}_{\|\mathcal{H}}(k)$. We know the Markov chain $\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\}$ is positive recurrent for each $\epsilon \in (0,1)$, so by definition of projection we also have that $\{\boldsymbol{q}^{(\epsilon)}_{\|\mathcal{K}}(k) : k \in \mathbb{Z}_+\}$, $\{\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{K}}(k) : k \in \mathbb{Z}_+\}$, $\{\boldsymbol{q}^{(\epsilon)}_{\|\mathcal{H}}(k) : k \in \mathbb{Z}_+\}$ and $\{\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{H}}(k) : k \in \mathbb{Z}_+\}$ are positive recurrent for each $\epsilon \in (0,1)$. Then, we define $\overline{\boldsymbol{q}}^{(\epsilon)}_{\|\mathcal{K}}$, $\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{K}}$, $\overline{\boldsymbol{q}}^{(\epsilon)}_{\|\mathcal{H}}$ and $\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{H}}$ as steady-state vectors which are limit in distribution of each them, respectively. In the next proposition we state SSC formally.

Proposition 7.4. *Given a vector $\boldsymbol{\nu}$ in the boundary of $\mathcal{C}$ and $\epsilon \in (0,1)$, consider a generalized switch operating under MaxWeight as described in Section 6.3, parametrized by ϵ as described at the beginning of Section 7.4, and let P be defined as in there as well. Let $\delta > 0$ be such that $\delta \leq b^{(\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle$ for all $\ell \in [L] \setminus P$ if $[L] \setminus P \neq \emptyset$, and $\delta = 1$ if $[L] \setminus P = \emptyset$. If $\epsilon < \frac{\delta}{2\|\boldsymbol{\nu}\|}$, then for each $j \in \mathbb{Z}_+$ with $j \geq 1$, we have*

$$\mathbb{E}\left[\|\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{H}}\|^j\right] \leq \mathbb{E}\left[\|\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{K}}\|^j\right] \leq J_j,$$

where, defining $\Lambda = \max\{A_{\max}, S_{\max}\}$,

$$J_j \stackrel{\triangle}{=} \left(\frac{8n\Lambda^2}{\delta}\right)^j + (8\sqrt{n}\Lambda)^j \left(\frac{8\sqrt{n}\Lambda + \delta}{\delta}\right)^j j!$$

To prove Proposition 7.4, we adopt the technique introduced by [34] so our proof is similar to theirs. We present as sketch of the proof at the end of this subsection. The challenges

in obtaining our result arise in the second step of the drift method, which corresponds to Theorem 7.5.

SSC is a consequence of Proposition 7.4 for the following reason. As $\epsilon \downarrow 0$, $\left\|\overline{\boldsymbol{q}}^{(\epsilon)}\right\|$ goes to infinity (this can be easily concluded from Theorem 7.5). Therefore, Proposition 7.4 implies that as ϵ gets small, we can approximate $\overline{\boldsymbol{q}}^{(\epsilon)} \approx \overline{\boldsymbol{q}}^{(\epsilon)}_{\parallel\mathcal{K}}$ because all the moments of $\|\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{K}}\|$ are bounded.

Observe that the cone $\mathcal{K}$ is determined by the facets that intersect at $\boldsymbol{\nu}$. Moreover, the dimension of the cone is $n - d_{\boldsymbol{\nu}}$, where $d_{\boldsymbol{\nu}}$ is the dimension of the face of $\mathcal{C}$ where $\boldsymbol{\nu}$ is. For example, if $\boldsymbol{\nu}$ is in the relative interior of a facet then $d_{\boldsymbol{\nu}} = n - 1$, and this implies that $\mathcal{K}$ is one-dimensional. This is the CRP case, which was studied in [34] and [8]. Similarly, if $\boldsymbol{\nu}$ is a vertex of $\mathcal{C}$ then $d_{\boldsymbol{\nu}} = 0$ and, hence, $\mathcal{K}$ is n-dimensional. In the last case, we say that SSC is full dimensional. We study the full-dimensional case in subsection 7.5.2.

Now we present a proof sketch of Proposition 7.4. We use Lemma 2.7. Then, the main idea is to show that the conditions (C1) and (C2) are satisfied for $Z(\boldsymbol{q}) = \|\boldsymbol{q}_{\perp}\|$.

Proof sketch of Proposition 7.4. For ease of exposition, we omit the dependence on ϵ of the random variables in this proof. First observe that $\mathcal{K} \subset \mathcal{H}$ by definition. Therefore, for all $j \in \mathbb{Z}_+$ with $j \geq 1$, we have $\left\|\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{H}}\right\|^j \leq \left\|\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{K}}\right\|^j$ with probability 1. This proves the first inequality.

To prove the second inequality, we introduce the following notation. Let

$$V(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}\|^2, \quad V_{\parallel}(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_{\parallel\mathcal{K}}\|^2, \quad V_{\perp}(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_{\perp\mathcal{K}}\|^2 \quad \text{and} \quad W_{\perp}(\boldsymbol{q}) \stackrel{\triangle}{=} \|\boldsymbol{q}_{\perp\mathcal{K}}\|.$$

We use Lemma 2.7 with Lyapunov function $W_{\perp}(\boldsymbol{q})$. We first prove that condition (C2) is satisfied. By definition of drift, we have

$$\begin{aligned}\left|\Delta W_{\perp}(\boldsymbol{q})\right| &= \left|W_{\perp}(\boldsymbol{q}(k+1)) - W_{\perp}(\boldsymbol{q}(k))\right| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\ &= \left|\left\|\boldsymbol{q}_{\perp\mathcal{K}}(k+1)\right\| - \left\|\boldsymbol{q}_{\perp\mathcal{K}}(k)\right\|\right| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}}\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \left\| \boldsymbol{q}_{\perp\mathcal{K}}(k+1) - \boldsymbol{q}_{\perp\mathcal{K}}(k) \right\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\overset{(b)}{=} \left\| \boldsymbol{q}(k+1) - \boldsymbol{q}(k) - \left(\boldsymbol{q}_{\|\mathcal{K}}(k+1) - \boldsymbol{q}_{\|\mathcal{K}}(k) \right) \right\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\overset{(c)}{\leq} \left(\left\| \boldsymbol{q}(k+1) - \boldsymbol{q}(k) \right\| + \left\| \boldsymbol{q}_{\|\mathcal{K}}(k+1) - \boldsymbol{q}_{\|\mathcal{K}}(k) \right\| \right) \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\overset{(d)}{\leq} 2 \left\| \boldsymbol{q}(k+1) - \boldsymbol{q}(k) \right\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\overset{(e)}{=} 2 \left\| \boldsymbol{q} + \boldsymbol{a}(k) - \boldsymbol{s}(k) + \boldsymbol{u}(k) - \boldsymbol{q} \right\| \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&= \left(2 \sqrt{\sum_{i=1}^{n} \left| a_i(k) - s_i(k) + u_i(k) \right|^2} \right) \mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \\
&\overset{(f)}{\leq} 2\sqrt{n} \max\{A_{\max}, S_{\max}\} \quad \text{with probability 1,}
\end{aligned} \tag{7.7}
$$

where (a) holds by triangle inequality; (b) holds by definition of $\boldsymbol{q}_{\perp\mathcal{K}}$; (c) holds by triangle inequality; (d) holds because projection on the cone $\mathcal{K}$ is nonexpansive; (e) holds by the dynamics of the queues presented in Equation 1.2; and (f) holds because $a_i(k) \leq A_{\max}$ with probability 1 and $s_i(k) \leq S_{\max}$ for all $i \in [n]$ and all $k \in \mathbb{Z}_+$. Therefore, if we let $D = 2\sqrt{n} \max\{A_{\max}, S_{\max}\}$ we have that condition (C2) is satisfied.

Now we prove condition (C1). We start with an observation that was first used in [34, Lemma 7, part 1]. Note that $W_\perp(\boldsymbol{q}) = \sqrt{\|\boldsymbol{q}_{\perp\mathcal{K}}\|^2}$ and $f(x) = \sqrt{x}$ is a concave function. Then, using the definition of concavity and reorganizing terms we have

$$
\Delta W_\perp(\boldsymbol{q}) \leq \frac{1}{2\|\boldsymbol{q}_{\perp\mathcal{K}}\|} \left(\Delta V(\boldsymbol{q}) - \Delta V_\|(\boldsymbol{q}) \right). \tag{7.8}
$$

We bound the conditional expectation of the terms in the brackets separately. We start with $\mathbb{E}_{\boldsymbol{q}}\left[\Delta V(\boldsymbol{q})\right]$. We obtain

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[\Delta V(\boldsymbol{q})\right] \\
=&\mathbb{E}_{\boldsymbol{q}}\left[\left\| \boldsymbol{q}(k+1) \right\|^2 - \left\| \boldsymbol{q}(k) \right\|^2 \right] \\
=&\mathbb{E}_{\boldsymbol{q}}\left[\left\| \boldsymbol{q}(k+1) - \boldsymbol{u}(k) + \boldsymbol{u}(k) \right\|^2 - \left\| \boldsymbol{q}(k) \right\|^2 \right]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}(k+1)-\boldsymbol{u}(k)\right\|^2+\left\|\boldsymbol{u}(k)\right\|^2+2\langle\boldsymbol{q}(k+1)-\boldsymbol{u}(k),\,\boldsymbol{u}(k)\rangle-\left\|\boldsymbol{q}(k)\right\|^2\right]\\
\overset{(a)}{=}&\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}(k)+\boldsymbol{a}(k)-\boldsymbol{s}(k)\right\|^2-\left\|\boldsymbol{u}(k)\right\|^2-\left\|\boldsymbol{q}(k)\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}(k)\right\|^2+\left\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\right\|^2+2\langle\boldsymbol{q}(k),\boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle-\left\|\boldsymbol{u}(k)\right\|^2-\left\|\boldsymbol{q}(k)\right\|^2\right]\\
\overset{(b)}{\leq}&\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\right\|^2\right]+2\mathbb{E}_{\boldsymbol{q}}\left[\langle\boldsymbol{q}(k),\,\boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle\right] \qquad (7.9)
\end{aligned}
$$

where (a) holds by the dynamics of the queues presented in Equation 1.2, by definition of inner product and by Equation 1.3; and (b) holds because $\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{u}(k)\right\|^2\right]\geq 0$ by definition of norm.

We bound the terms in Equation 7.9 separately. First, observe

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\right\|^2\right]=&\mathbb{E}_{\boldsymbol{q}}\left[\sum_{i=1}^{n}\big(a_i(k)-s_i(k)\big)^2\right]\\
=&\mathbb{E}_{\boldsymbol{q}}\left[\sum_{i=1}^{n}\big(a_i^2(k)+s_i^2(k)-2a_i(k)s_i(k)\big)\right]\\
\leq&n(A_{\max}^2+S_{\max}^2), \qquad (7.10)
\end{aligned}
$$

where the inequality holds because $2a_i(k)s_i(k)\geq 0$ with probability 1 for all $i\in[n]$ and all $k\in\mathbb{Z}_+$; and because $0\leq a_i(k)\leq A_{\max}$ with probability 1 and $0\leq s_i(k)\leq S_{\max}$ for all $i\in[n]$ and all $k\in\mathbb{Z}_+$. Let $\zeta\triangleq n(A_{\max}^2+S_{\max}^2)$.

On the other hand,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}\left[\langle\boldsymbol{q}(k),\,\boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle\right]=&\langle\boldsymbol{q},\boldsymbol{\lambda}^{(\epsilon)}\rangle-\mathbb{E}_{\boldsymbol{q}}\left[\langle\boldsymbol{q}(k),\boldsymbol{s}(k)\rangle\right]\\
\overset{(a)}{=}&\langle\boldsymbol{q},(1-\epsilon)\boldsymbol{\nu}\rangle-\max_{\boldsymbol{x}\in\mathcal{C}}\langle\boldsymbol{q},\boldsymbol{x}\rangle\\
=&-\epsilon\langle\boldsymbol{q},\boldsymbol{\nu}\rangle+\min_{\boldsymbol{x}\in\mathcal{C}}\langle\boldsymbol{q},\boldsymbol{\nu}-\boldsymbol{x}\rangle\\
\leq&-\epsilon\langle\boldsymbol{q},\boldsymbol{\nu}\rangle+\langle\boldsymbol{q},\boldsymbol{\nu}-\boldsymbol{x}^*\rangle, \qquad (7.11)
\end{aligned}
$$

for any $\boldsymbol{x}^*\in\mathcal{C}$. Here, equality (a) holds by definition of $\boldsymbol{\lambda}^{(\epsilon)}$ and by Lemma 6.1.

We pick $\boldsymbol{x}^* = \boldsymbol{\nu} + \frac{\delta}{2\|\boldsymbol{q}_{\perp\mathcal{K}}\|}\boldsymbol{q}_{\perp\mathcal{K}}$. Before proceeding with the proof, we show that such $\boldsymbol{x}^* \in \mathcal{C}$. To do that, we show that $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}^* \rangle \leq b^{(\ell)}$ for all $\ell \in [L]$. We have two cases. If $\ell \in P$, then

$$\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}^* \rangle = \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle + \frac{\delta}{\|\boldsymbol{q}_{\perp\mathcal{K}}\|}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}_{\perp\mathcal{K}} \rangle \overset{(a)}{=} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle \overset{(b)}{=} b^{(\ell)}$$

where (a) holds because $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}_{\perp\mathcal{K}} \rangle = 0$ for all $\ell \in P$, by the orthogonality principle; and (b) holds because $\boldsymbol{\nu} \in \bigcap_{\ell \in P} \mathcal{F}^{(\ell)}$.

If $[L] \setminus P \neq \emptyset$ and $\ell \in [L] \setminus P$ we have $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle < b^{(\ell)}$. Then, for each $\ell \notin P$ there exists $\delta^{(\ell)} > 0$ such that $\boldsymbol{\nu} + \frac{\delta^{(\ell)}}{2\|\boldsymbol{q}_{\perp\mathcal{K}}\|}\boldsymbol{q}_{\perp\mathcal{K}} \in \mathcal{C}$. Then, since there are finitely many hyperplanes defining $\mathcal{C}$, we can pick $\delta = \min_{\ell \in [L]\setminus P} \left\{\delta^{\ell}\right\}$.

Then, from Equation 7.11 we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}(k),\, \boldsymbol{a}(k) - \boldsymbol{s}(k) \rangle\right] &\leq -\epsilon\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle + \langle \boldsymbol{q}, \boldsymbol{\nu} - \left(\boldsymbol{\nu} + \frac{\delta}{\left\|\boldsymbol{q}_{\perp\mathcal{K}}\right\|}\boldsymbol{q}_{\perp\mathcal{K}}\right) \rangle \\
&= -\epsilon\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle + \frac{\delta}{\left\|\boldsymbol{q}_{\perp\mathcal{K}}\right\|}\langle \boldsymbol{q}, \boldsymbol{q}_{\perp\mathcal{K}} \rangle \\
&= -\epsilon\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle + \delta\|\boldsymbol{q}_{\perp\mathcal{K}}\|, \qquad (7.12)
\end{aligned}$$

where the last equality holds because $\boldsymbol{q} = \boldsymbol{q}_{\|\mathcal{K}} + \boldsymbol{q}_{\perp\mathcal{K}}$ and $\langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{q}_{\perp\mathcal{K}} \rangle = 0$. Then, using Equation 7.10 and Equation 7.12 in Equation 7.9 we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V(\boldsymbol{q})\right] \leq \zeta - 2\epsilon\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle - 2\delta\left\|\boldsymbol{q}_{\perp\mathcal{K}}\right\|. \qquad (7.13)$$

To bound the second term in Equation 7.8 we use properties of projection. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{\|}(\boldsymbol{q})\right] \\
=&\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}_{\|\mathcal{K}}(k+1)\right\|^2 - \left\|\boldsymbol{q}_{\|\mathcal{K}}(k)\right\|^2\right] \\
=&\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k+1) + \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}_{\|\mathcal{K}}(k+1) - \boldsymbol{q}_{\|\mathcal{K}}(k) \rangle\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}_{\|\mathcal{K}}(k+1)-\boldsymbol{q}_{\|\mathcal{K}}(k)\right\|^2\right] + 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}_{\|\mathcal{K}}(k+1)-\boldsymbol{q}_{\|\mathcal{K}}(k)\rangle\right] \\
&\overset{(a)}{\geq} 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}_{\|\mathcal{K}}(k+1)-\boldsymbol{q}_{\|\mathcal{K}}(k)\rangle\right] \\
&\overset{(b)}{=} 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}(k+1)-\boldsymbol{q}(k)\rangle\right] - 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}_{\perp\mathcal{K}}(k+1)-\boldsymbol{q}_{\perp\mathcal{K}}(k)\rangle\right] \\
&= 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{a}(k)-\boldsymbol{s}(k)+\boldsymbol{u}(k)\rangle\right] - 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k), \boldsymbol{q}_{\perp\mathcal{K}}(k+1)\rangle\right] \\
&\qquad + 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{q}_{\perp\mathcal{K}}(k)\rangle\right] \\
&\overset{(c)}{\geq} 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle\right] \\
&= 2\langle \boldsymbol{q}_{\|\mathcal{K}}, (1-\epsilon)\boldsymbol{\nu}\rangle - 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}(k),\, \boldsymbol{s}(k)\rangle\right] \\
&= -2\epsilon\langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{\nu}\rangle + 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{\nu}-\boldsymbol{s}(k)\rangle\right] \qquad (7.14)
\end{aligned}
$$

where (a) holds because $\mathbb{E}_{\boldsymbol{q}}\left[\left\|\boldsymbol{q}_{\|\mathcal{K}}(k+1)-\boldsymbol{q}_{\|\mathcal{K}}(k)\right\|^2\right] \geq 0$; (b) holds because $\boldsymbol{q}_{\|\mathcal{K}}(k) = \boldsymbol{q}(k)-\boldsymbol{q}_{\perp\mathcal{K}}(k)$ for all $k \in \mathbb{Z}_+$; (c) holds because, since $\boldsymbol{q}_{\|\mathcal{K}}(k) \geq 0$ and $\boldsymbol{u}(k) \geq 0$ by definition, we have $\langle \boldsymbol{q}_{\|\mathcal{K}}(k), \boldsymbol{u}(k)\rangle \geq 0$, because $\langle \boldsymbol{q}_{\|\mathcal{K}}(k), \boldsymbol{q}_{\perp\mathcal{K}}(k)\rangle = 0$ since they are orthogonal by definition, and $\langle \boldsymbol{q}_{\|\mathcal{K}}(k), \boldsymbol{q}_{\perp\mathcal{K}}(k+1)\rangle \leq 0$ because $\boldsymbol{q}_{\|\mathcal{K}}(k)$ belongs to the cone $\mathcal{K}$ and $\boldsymbol{q}_{\perp\mathcal{K}}(k+1)$ belongs to the polar cone of $\mathcal{K}$, defined as $\mathcal{K}^\circ \overset{\triangle}{=} \{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq 0 \ \ \forall \boldsymbol{x} \in \mathcal{K}\}$.

Since $\boldsymbol{q}_{\|\mathcal{K}}(k)$ is the projection of $\boldsymbol{q}(k)$ on $\mathcal{K}$, there exist coefficients $\xi_\ell \geq 0$ with $\ell \in P$ such that

$$
\boldsymbol{q}_{\|\mathcal{K}}(k) = \sum_{\ell \in P} \xi_\ell \boldsymbol{c}^{(\ell)}.
$$

Then,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}_{\|\mathcal{K}},\, \boldsymbol{\nu}-\boldsymbol{s}(k)\rangle\right] &= \sum_{\ell \in P} \xi_\ell \mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{s}(k)\rangle\right] \\
&\overset{(a)}{=} \sum_{\ell \in P} \xi_\ell \left(b^\ell - \langle \boldsymbol{c}^{(\ell)}, \arg\max_{\boldsymbol{x} \in \mathcal{C}} \langle \boldsymbol{q}, \boldsymbol{x}\rangle\rangle\right) \\
&\overset{(b)}{=} \sum_{\ell \in P} \xi_\ell \left(b^{(\ell)} - \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{x}}\rangle\right)
\end{aligned}
$$

$$\overset{(c)}{\geq} \sum_{\ell \in P} \xi_\ell (b^{(\ell)} - b^{(\ell)}) = 0$$

for some $\overline{\boldsymbol{x}} \in \mathcal{C}$. Here, (a) holds because $\boldsymbol{\nu} \in \bigcap_{\ell \in P} \mathcal{F}^{(\ell)}$ and by Lemma 6.1; (b) holds for some $\overline{\boldsymbol{x}} \in \mathcal{C}$ because $\mathcal{C}$ is a closed and bounded set, so the maximum is attained at some point in $\mathcal{C}$; and (c) holds because, since $\overline{\boldsymbol{x}} \in \mathcal{C}$, then $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{x}} \rangle \leq b^{(\ell)}$ for all $\ell \in [L]$.

Therefore, from Equation 7.14 we obtain

$$\mathbb{E}_{\boldsymbol{q}} \left[\Delta V_{\|}(\boldsymbol{q}) \right] \geq -2\epsilon \langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{\nu} \rangle. \tag{7.15}$$

Then, using Equation 7.13 and Equation 7.15 in Equation 7.8 we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{q}} \left[\Delta W_{\perp}(\boldsymbol{q}) \right] \leq & \frac{1}{2\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} \left(\zeta - 2\epsilon \langle \boldsymbol{q}, \boldsymbol{\nu} \rangle - 2\delta \left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\| + 2\epsilon \langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{\nu} \rangle \right) \\
= & \frac{\zeta}{2\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} - \delta + \frac{\epsilon}{\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} \left(\langle \boldsymbol{q}_{\|\mathcal{K}}, \boldsymbol{\nu} \rangle - \langle \boldsymbol{q}, \boldsymbol{\nu} \rangle \right) \\
\overset{(a)}{=} & \frac{\zeta}{2\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} - \delta + \frac{\epsilon}{\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} \left(- \langle \boldsymbol{q}_{\perp\mathcal{K}}, \boldsymbol{\nu} \rangle \right) \\
\overset{(b)}{\leq} & \frac{\zeta}{2\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} - \delta + \epsilon \|\boldsymbol{\nu}\|
\end{aligned}$$

where (a) holds because $\boldsymbol{q} = \boldsymbol{q}_{\|\mathcal{K}} + \boldsymbol{q}_{\perp\mathcal{K}}$; and (b) holds by Cauchy-Schwarz inequality. Therefore, if $\epsilon \leq \frac{\delta}{2\|\boldsymbol{\nu}\|}$ we have

$$\mathbb{E}_{\boldsymbol{q}} \left[\Delta W_{\perp}(\boldsymbol{q}) \right] \leq \frac{\zeta}{2\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\|} - \frac{\delta}{2}.$$

Further, if $\left\| \boldsymbol{q}_{\perp\mathcal{K}} \right\| \geq \frac{2\zeta}{\delta}$ we have $\mathbb{E}_{\boldsymbol{q}} \left[\Delta W_{\perp}(\boldsymbol{q}) \right] \leq -\frac{\delta}{4}$. The last inequality verifies condition (C1) with $\eta = \frac{\delta}{4}$ and $\kappa = \frac{2\zeta}{\delta}$. This completes the proof. □

7.4.3 Asymptotically tight bounds.

In subsection 7.4.2 we showed SSC into the cone $\mathcal{K}$, which implies SSC into the subspace $\mathcal{H}$. In this section we present the main result of this chapter (Theorem 7.5), where we

provide asymptotically tight bounds to the expected value of certain linear combinations of the queue lengths in steady state. After the statement of the theorem we present some remarks and applications, and we delay the proof to section 7.6.

Theorem 7.5. *Given a vector $\boldsymbol{\nu}$ in the boundary of $\mathcal{C}$, let P be defined at the beginning of Section 7.4. Consider a set of generalized switches operating under MaxWeight, indexed by the heavy-traffic parameter $\epsilon \in (0,1)$ as described at the beginning of Section 7.4. Then, for any vector $\boldsymbol{w} \in \cap_{\ell \in P} \mathcal{F}^{(\ell)}$ we have*

$$\left| \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle \right] - \frac{1}{2\epsilon} \mathbf{1}^T \left(H \circ \Sigma_a^{(\epsilon)} \right) \mathbf{1} - \frac{1}{2\epsilon} \mathbf{1}^T \left((C^T C)^{-1} \circ \Sigma_B \right) \mathbf{1} \right| \leq \zeta \log\left(\frac{1}{\epsilon} \right), \quad (7.16)$$

where $H \stackrel{\triangle}{=} C(C^T C)^{-1} C^T$ is the projection matrix into $\mathcal{H}$ and ζ is a constant independent of ϵ and $\boldsymbol{w}$. This implies that, if $\lim_{\epsilon \downarrow 0} \Sigma_a^{(\epsilon)} = \Sigma_a$ component-wise, then,

$$\lim_{\epsilon \downarrow 0} \epsilon\, \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle \right] = \frac{1}{2} \left(\mathbf{1}^T \left(H \circ \Sigma_a \right) \mathbf{1} + \mathbf{1}^T \left((C^T C)^{-1} \circ \Sigma_B \right) \mathbf{1} \right). \quad (7.17)$$

First observe that Equation 7.16 gives bounds that are valid for all regimes, not necessarily heavy traffic. Additionally, it shows that the queue lengths grow to infinity as the traffic intensity grows (i.e., as $\epsilon \downarrow 0$).

In Equation 7.17, observe that the right-hand side has two terms: one corresponding to randomness in the arrival process, and the other one to randomness in the service process. The first term is a linear combination of the covariance matrix of the arrival process, and the weights of the linear combination are determined by the projection matrix on the subspace $\mathcal{H}$, which is where SSC occurs. The second term is a linear combination of the elements of a covariance matrix which is related to the channel state. Since the potential service rate vector is selected using MaxWeight algorithm (see Equation 6.3), it is not actually random once queue lengths and channel state are observed. However, the channel state is a random variable that defines the feasible set where MaxWeight is solved. Hence, the second term in Equation 7.17, which includes a covariance matrix related to channel state, represents

the randomness on the service process.

A third observation is that, in order to project on the subspace $\mathcal{H}$ generated by the cone $\mathcal{K}$, we had to drop the vectors $\boldsymbol{c}^{(\ell)}$ with $\ell \in P$ that are linearly dependent (recall that the columns of the matrix C are a linearly independent subset of the vectors that generate $\mathcal{K}$). Clearly, the cone generated by the columns of C is not equal to $\mathcal{K}$. However, projecting on the subspace $\mathcal{H}$ is sufficient, and we do not need to worry about these linearly dependent vectors that we dropped.

In the next remark we write Equation 7.17 in different ways to facilitate interpretation of the result.

Remark 7.6. *Equation 7.17 can be also written as*

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle\right] \\
=&\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n} \langle \boldsymbol{e}^{(i)}, \boldsymbol{e}_{\|\mathcal{H}}^{(i')} \rangle (\Sigma_a)_{i,j} + \sum_{\ell_1 \in \tilde{P}} \sum_{\ell_2 \in \tilde{P}} (C^T C)^{-1}_{\ell_1,\ell_2} (\Sigma_B)_{\ell_1,\ell_2}\right) \qquad (7.18)\\
=&\frac{1}{2}\left(Trace\left(H\Sigma_a^T\right) + Trace\left((C^T C)^{-1}\Sigma_B^T\right)\right), \qquad (7.19)
\end{aligned}$$

where the subscript $\|\mathcal{H}$ denotes projection on the subspace $\mathcal{H}$, $(\Sigma_a)_{i,j}$ is the element (i,j) of the covariance matrix Σ_a for each $i,j \in [n]$, and $(\Sigma_B)_{\ell_1,\ell_2}$ is the element (ℓ_1,ℓ_2) of Σ_B for each $\ell_1, \ell_2 \in \tilde{P}$.

In some cases, the projection of a vector on $\mathcal{H}$ is known in closed form, and it is simpler to work with than the projection matrix. For example, in the case of a completely saturated input-queued switch, one can directly compute the projections as in [14], but writing down the projection matrix is more involved.

We present the proof of Remark 7.6 below.

Proof of Remark 7.6. If we expand the products on the right-hand side of Equation 7.17

we obtain

$$\begin{aligned}
&\frac{1}{2}\left(\mathbf{1}^T\left(H\circ\Sigma_a\right)\mathbf{1}+\mathbf{1}^T\left((C^TC)^{-1}\circ\Sigma_B\right)\mathbf{1}\right)\\
&\overset{(a)}{=}\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n}h_{i,i'}(\Sigma_a)_{i,i'}+\sum_{\ell_1\in\tilde{P}}\sum_{\ell_2\in\tilde{P}}(C^TC)^{-1}_{\ell_1,\ell_2}(\Sigma_B)_{\ell_1,\ell_2}\right)\\
&\overset{(b)}{=}\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(\boldsymbol{e}^{(i)}\right)^T H\boldsymbol{e}^{(i')}(\Sigma_a)_{i,i'}+\sum_{\ell_1\in\tilde{P}}\sum_{\ell_2\in\tilde{P}}(C^TC)^{-1}_{\ell_1,\ell_2}(\Sigma_B)_{\ell_1,\ell_2}\right)\\
&\overset{(c)}{=}\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n}\langle\boldsymbol{e}^{(i)},\boldsymbol{e}^{(i')}_{\|\mathcal{H}}\rangle(\Sigma_a)_{i,i'}+\sum_{\ell_1\in\tilde{P}}\sum_{\ell_2\in\tilde{P}}(C^TC)^{-1}_{\ell_1,\ell_2}(\Sigma_B)_{\ell_1,\ell_2}\right),
\end{aligned}$$

where (a) holds by definition of Hadamard's product; (b) holds by definition of the canonical vectors $\boldsymbol{e}^{(i)}$ and by definition of matrix product; and (c) holds by definition of inner product and because $H\boldsymbol{e}^{(i')}$ is the projection of $\boldsymbol{e}^{(i')}$ on the subspace $\mathcal{H}$.

The proof of Equation 7.19 holds by properties of Hadamard's product and trace, and we omit it. □

Observe that the bounds presented in Proposition 7.3 and Theorem 7.5 may be for different linear combinations of the vector of queue lengths. In Proposition 7.3 the vector of weights is $\boldsymbol{z}\in\mathcal{K}$ and in Theorem 7.5 it is $\boldsymbol{w}\in\cap_{\ell\in P}\mathcal{F}^{(\ell)}$. In the next remark we give sufficient conditions under which these bounds correspond to the same linear combination of the queue lengths.

Remark 7.7. *Let A be a matrix with columns $\mathbf{c}^{(\ell)}$ for $\ell\in P$ and $\mathbf{b}_P$ be a vector with elements $b^{(\ell)}$ for $\ell\in P$. Observe that the column space of A is equal to the column space of C, but the columns of A may not be linearly independent. In fact, if the columns of A are linearly independent, then $A=C$. Then, Proposition 7.3 and Theorem 7.5 give bounds to the same linear combination of the queue lengths if $\mathcal{A}\overset{\triangle}{=}\left\{\boldsymbol{x}\in\mathbb{R}^{|P|}:\boldsymbol{x}^TA^TA\geq 0\ ,\ \boldsymbol{x}^T\mathbf{b}_P<0\right\}$ is empty.*

Proof of Remark 7.7. We can obtain bounds to the same linear combination of the queue

lengths if there exists a vector $\boldsymbol{y} \in \mathcal{K} \cap \left(\cap_{\ell \in P} \mathcal{F}^{(\ell)}\right)$. In other words, if the set $\mathcal{Y} \triangleq \left\{\boldsymbol{y} \in \mathbb{R}_{+}^{|P|} : AA^T\boldsymbol{y} = \boldsymbol{b}_P\right\}$ is nonempty. By Farkas' lemma [102, Theorem 4.6], proving that $\mathcal{Y} \neq \emptyset$ is equivalent to proving that $\mathcal{A} = \emptyset$. □

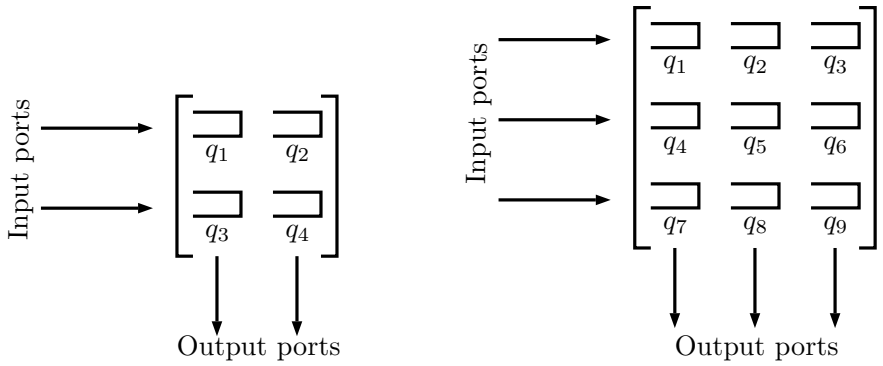
7.5 Applications of Theorem 7.5

The generalized switch is a model that subsumes several SPNs, such as ad hoc wireless networks, the input-queued switch, down-link base stations and the parallel-server system. In this section we elaborate on a few applications to give examples of the use of Theorem 7.5, and it is by no means an exhaustive list. We start with an input-queued switch in subsection 7.5.1, and then, in subsection 7.5.2, we present examples where full-dimensional SSC is observed.

7.5.1 Input-queued switch.

The drift method has been used to perform heavy-traffic analysis of the input-queued switch operating under MaxWeight in both, completely and incompletely saturated cases [14, 15], respectively. In both scenarios, the analysis is performed under the assumption that the arrivals to different queues are independent. However, this is an unrealistic assumption in data center networks. Indeed, it has been shown that the traffic exhibits hot-spots, i.e., there are subsets of queues that simultaneously perceive a surge on traffic [100, 101]. This implies that the arrival processes are highly correlated. In this section we focus on the completely saturated input-queued switch, and we obtain the heavy-traffic limit of the scaled total queue length when the arrivals are correlated, as a corollary of Theorem 7.5. Corollary 7.8 generalizes the main result proved by [14] and it is of special interest by itself, given the nature of the arrival processes to data center networks observed in reality. We start specifying the model.

Consider a system with N^2 queues operating in discrete time. There are N input ports, N output ports, and there is a different queue for each input/output pair. Each of these pairs



(a) 2×2 switch. (b) 3×3 switch.

Figure 7.2: Diagram of the queue length vector for the input-queued switch.

has its own arrival process and all the arriving packets have the same size, which is equal to one time slot. The service process must satisfy the following feasibility constraints. In each time slot, at most one packet can be transmitted from each input port, and each output port can process at most one packet. We can think of this system as a matrix of input/output pairs, where rows represent inputs and columns represent outputs. Then, the constraint described above can be also stated as follows. In each time slot, at most one queue can be active (i.e., processing jobs) in each row and each column.

This model corresponds to a generalized switch with $n = N^2$ queues, where the channel state is constant over time. As mentioned above, the input-queued switch has a natural matrix-shape interpretation. In [14, 15] represent the vectors of queue lengths, arrivals and services by $N \times N$ matrices, but they are treated as vectors. Specifically, dot products and norms are computed as if these matrices were column vectors. In this paper, however, we will write them as column vectors to be consistent with the notation we introduced in section 6.3. We enumerate the elements of the vectors row by row. For each $i \in [n]$ we have that $q_i(k)$ is the number of packets in line in input port $\lceil \frac{i}{N} \rceil$, waiting for service from output $i \bmod N$ if i is not a multiple of N, and output N otherwise. Similarly for the vectors of arrivals, potential service and unused service. In Figure 7.2 we show how to build the vectors in the case of $N = 2$ (Figure 7.2a) and $N = 3$ (Figure 7.2b).

For ease of exposition, we introduce the following notation. For each $i \in [N^2]$ let

$$row(i) \stackrel{\triangle}{=} \left\{ \left(\left\lceil \frac{i}{N} \right\rceil - 1 \right) N + j : j \in [N] \right\} \setminus \{i\}$$

$$col(i) \stackrel{\triangle}{=} \left\{ j \in \left[N^2\right] : i \mod N = j \mod N \right\} \setminus \{i\}$$

$$other(i) \stackrel{\triangle}{=} \left[N^2\right] \setminus \left(row(i) \cup col(i) \cup \{i\} \right).$$

In words, the set $row(i)$ contains the index of all elements in the same row as i, except by i; $col(i)$ contains the index of the elements in the same column as i, except by i; and $other(i)$ contains all indexes that do not correspond to the same row or column as i, or i itself.

We explicitly know the feasibility constraints in the input-queued switch. Then, we can compute the set of feasible service rate vectors $\mathcal{S}$ and the capacity region $\mathcal{C}$. We obtain

$$\mathcal{S} = \left\{ \boldsymbol{x} \in \{0,1\}^{N^2} : \sum_{i=1}^{N} x_{N(j-1)+i} \leq 1 \ \forall j \in [N] \text{ and } \sum_{j=1}^{N} x_{N(j-1)+i} \leq 1 \ \forall i \in [N] \right\},$$

and

$$\begin{aligned} \mathcal{C} &= \text{ConvexHull}(\mathcal{S}) \\ &= \left\{ \boldsymbol{x} \in \mathbb{R}_+^{N^2} : \right. \\ &\qquad \left. \sum_{i=1}^{N} x_{N(j-1)+i} \leq 1 \ \forall j \in [N] \text{ and } \sum_{j=1}^{N} x_{N(j-1)+i} \leq 1 \ \forall i \in [N] \right\}. \end{aligned} \tag{7.20}$$

Then, the number of hyperplanes that define the capacity region is $L = 2N$, the right-hand side parameters are $b^{(\ell)} = 1$ for all $\ell \in [2N]$, and the left-hand side vectors $\boldsymbol{c}^{(\ell)}$ are defined

as follows.

$$\boldsymbol{c}^{(\ell)} = \begin{cases} \sum_{i=N(\ell-1)+1}^{N\ell} \boldsymbol{e}^{(i)} & \text{, if } \ell \in [N] \\ \sum_{i \in \{i' : i' \bmod N = \ell \bmod N\}} \boldsymbol{e}^{(i)} & \text{, if } \ell \in [2N] \setminus [N]. \end{cases} \tag{7.21}$$

Completely saturated switch means that the vector $\boldsymbol{\nu}$ that we approach in the heavy-traffic limit satisfies all the inequalities in Equation 7.20 at equality. Formally, $\boldsymbol{\nu}$ satisfies $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle = b^{(\ell)}$ for all $\ell \in [2N]$. Then, $P = [2N]$. If $\boldsymbol{\nu}$ does not satisfy all the inequalities at equality, it is said that the switch is incompletely saturated. We do not study the incompletely saturated case here.

Recall that the cone $\mathcal{K}$ where SSC occurs is the cone generated by the vectors $\boldsymbol{c}^{(\ell)}$ with $\ell \in P$. In this case, since $P = [2N]$ and since we explicitly know the vectors $\boldsymbol{c}^{(\ell)}$, it can be easily proved that the cone $\mathcal{K}$ can be described as

$$\mathcal{K} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^{N^2} : \; x_i = \frac{1}{N} \sum_{j \in row(i) \cup \{i\}} x_j + \frac{1}{N} \sum_{j \in col(i) \cup \{i\}} x_j - \frac{1}{N^2} \sum_{j=1}^{N^2} x_j \right\}. \tag{7.22}$$

The proof of this claim is just algebra, and we omit it for brevity. In this case, it can be also proved that the subspace $\mathcal{H}$ generated by the cone $\mathcal{K}$ satisfies $\mathcal{K} = \mathcal{H} \cap \mathbb{R}_+^{N^2}$.

Now we present the heavy-traffic limit of the scaled sum of the queue lengths in a completely saturated switch with correlated arrival processes, as a corollary of Theorem 7.5. This corollary by itself is a contribution because, to the best of our knowledge, the input-queued switch has been studied only under independent arrivals assumption. However, it is known that in data centers this is not satisfied and, in fact, hot-spots are frequently observed.

Corollary 7.8. *Let $\boldsymbol{\nu}$ be an N^2-dimensional vector that satisfies $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle = b^{(\ell)}$ for all $\ell \in [2N]$, for $\boldsymbol{c}^{(\ell)}$ as defined in Equation 7.21 and $b^{(\ell)} = 1$ for all $\ell \in [2N]$. Consider a set of $N \times N$ input-queued switches as described above, parametrized by $\epsilon \in (0, 1)$ as*

described in Theorem 7.5. For each $i \in [N^2]$, let $\sigma_{a_i}^2 = (\Sigma_a)_{i,i}$. Then,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i=1}^{N^2} \overline{q}_i^{(\epsilon)}\right] = \frac{1}{2N} \sum_{i=1}^{N^2} \left((2N-1)\sigma_{a_i}^2 + (N-1) \sum_{j \in row(i) \cup col(i)} (\Sigma_a)_{i,j} - \sum_{j \in other(i)} (\Sigma_a)_{i,j} \right).$$

Proof of Corollary 7.8. We use Remark 7.6. We first compute $\boldsymbol{e}_{\|\mathcal{H}}^{(i)}$ for each $i \in [N^2]$. For any vector $\boldsymbol{y} \in \mathbb{R}_+^{N^2}$ we have $\boldsymbol{y}_{\|\mathcal{H}}$ has elements

$$\boldsymbol{y}_{\|\mathcal{H}j} = \frac{1}{N} \sum_{j' \in row(j) \cup \{j\}} y_{j'} + \frac{1}{N} \sum_{j' \in col(j) \cup \{j\}} y_{j'} - \frac{1}{N^2} \sum_{j'=1}^{N^2} y_{j'} \qquad \forall j \in [N^2].$$

Then, for each $i \in [N^2]$ the vector $\boldsymbol{e}_{\|\mathcal{H}}^{(i)}$ has elements

$$\boldsymbol{e}_{\|\mathcal{H}j}^{(i)} = \begin{cases} \frac{2N-1}{N^2} & \text{, if } j = i \\ \frac{N-1}{N^2} & \text{, if } j \in row(i) \text{ or } j \in col(i) \\ -\frac{1}{N^2} & \text{, if } j \in other(i) \end{cases} \qquad \forall j \in [N^2].$$

Using this expression in Remark 7.6 we immediately obtain the result. □

A special case of Corollary 7.8 is when the arrival processes to different input ports are independent. In this case, we recover the result from [14, Theorem 1], where they explicitly set to zero the drift of $V_{\|\mathcal{H}}(\boldsymbol{q}) = \left\|\boldsymbol{q}_{\|\mathcal{H}}\right\|^2$ (similarly to our approach in the proof of 7.5). We present the result below for completeness.

Corollary 7.9. *Consider a set of $N \times N$ input-queued switches operating under MaxWeight, parametrized by $\epsilon \in (0,1)$ as described in Corollary 7.8. Further, assume that the arrival*

processes to different queues are independent. Then,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i=1}^{N^2} \overline{q}_i^{(\epsilon)}\right] = \left(1 - \frac{1}{2N}\right) \sum_{i=1}^{N^2} \sigma_{a_i}^2.$$

The proof of Corollary 7.9 is easy after considering Corollary 7.8, since $(\Sigma_a)_{i,j} = 0$ for all $i \neq j$ under the independent arrivals assumption.

7.5.2 Full-dimensional SSC.

As mentioned in subsection 7.4.2, if the point $\boldsymbol{\nu}$ is a vertex of the capacity region $\mathcal{C}$, the cone $\mathcal{K}$ is n-dimensional. In other words, $\mathcal{K}$ is full-dimensional. In this section we explore this situation, and we present examples of SPNs where this phenomenon is observed. In particular, we present the case of an ad hoc wireless network, a parallel-server system operating in discrete time, an $\mathcal{N}$-system.

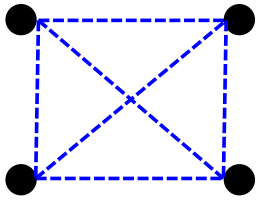
We first present the result in a general case.

Corollary 7.10. *Consider a set of generalized switches operating under MaxWeight, parametrized by $\epsilon \in (0,1)$ as described in Theorem 7.5. Let P, $\tilde{P}$ and $\boldsymbol{\nu}$ be as in Theorem 7.5 and suppose the cone $\mathcal{K}$ is n-dimensional. Let $\sigma_{a_i}^2 \stackrel{\triangle}{=} (\Sigma_a)_{i,i}$ for each $i \in [n]$ and $\boldsymbol{\sigma}_a$ be a vector with elements σ_{a_i}. Then,*

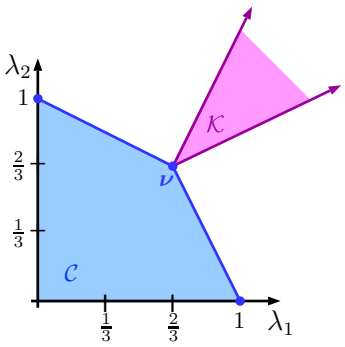
$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle\right] = \frac{1}{2}\left(\|\boldsymbol{\sigma}_a\|^2 + \mathbf{1}^T\left((C^T C)^{-1} \circ \Sigma_B\right)\mathbf{1}\right).$$

Observe that Corollary 7.10 gives a rather surprising result. The right-hand side of the limit does not depend on the correlation among arrivals to different queues. In other words, in the heavy-traffic limit, these linear combinations of the queue lengths behave as if the arrival processes were independent when SSC is full-dimensional.

The proof of Corollary 7.10 follows immediately from Theorem 7.5 because, if the cone $\mathcal{K}$ is full-dimensional, then the subspace $\mathcal{H} = \mathbb{R}^n$ and, therefore, the projection matrix on



(a) Example of an interference graph for an ad hoc wireless network with four links.



(b) Capacity region and cone for SPN in subsubsection 7.5.2.

Figure 7.3: Diagram of ad hoc wireless networks.

$\mathcal{H}$ satisfies $H = \mathbb{I}$. In the rest of this section, we present examples of SPNs that experience full-dimensional SSC.

Ad hoc wireless network.

An ad hoc wireless network is composed by a set of nodes with no infrastructure for central coordination, and packets are transmitted between nodes (a transmitter and a receiver) if there is a link. The links interfere with each other and, therefore, not all of them can be active at the same time. These interference constraints are frequently represented with a graph, where the vertices represent links and an edge between two links represents interference. In Figure 7.3a we present an example of the interference graph of an ad hoc wireless network with four links, where all links interfere with each other. The packets to be transmitted arrive to each of the links and wait in line until they can be processed. This model has been studied in a long line of literature, including but not limited to [103, 104, 34, 105], but in most of the cases the focus is on studying stability or optimality of the scheduling policy. Here we provide the heavy-traffic limit of linear combinations of the queue lengths under MaxWeight algorithm. A particular case of our result are the results obtained in [34].

An ad hoc wireless network can be modeled as a generalized switch with fixed channel state. Then, Theorem 7.5 can be immediately applied. In this section we provide an exam-

ple of an ad hoc wireless network that experiences full-dimensional SSC. We focus on a network with two links to illustrate the geometry of the capacity region and the cone where SSC occurs, but similar work can be done for larger networks.

Let $\sigma_1^2 \triangleq \text{Var}\left[\overline{a}_1^{(\epsilon)}\right]$, $\sigma_2^2 \triangleq \text{Var}\left[\overline{a}_2^{(\epsilon)}\right]$, and $\varphi \triangleq \text{Cov}\left[\overline{a}_1^{(\epsilon)}, \overline{a}_2^{(\epsilon)}\right]$, where these three parameters do not depend on ϵ. Suppose the set of feasible service rate vectors is $\mathcal{S} = \left\{(1,0),\, (0,1),\, \left(\frac{2}{3}, \frac{2}{3}\right)\right\}$. Then, the capacity region is

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}_+^2 : x_1 + 2x_2 \leq 2,\, 2x_1 + x_2 \leq 2\right\}.$$

Applying Theorem 7.5 we obtain the following corollary.

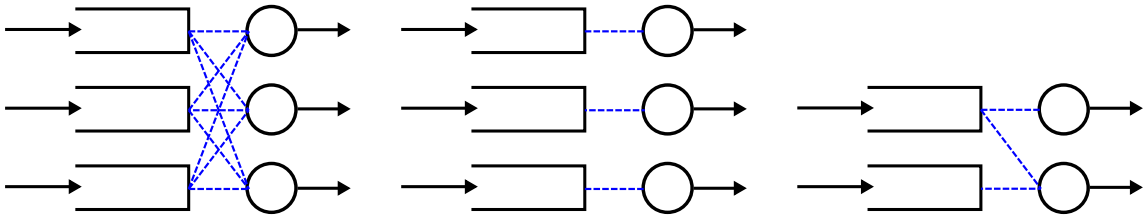
Corollary 7.11. *Consider an ad hoc wireless network as described above. Then,*

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1^{(\epsilon)} + \overline{q}_2^{(\epsilon)}\right] = \frac{3}{4}\left(\sigma_1^2 + \sigma_2^2\right)$$

In the proof of Corollary 7.11 we take the heavy-traffic limit as the vector of arrival rate approaches the point $\boldsymbol{\nu} = \frac{2}{3}(1,1)$ in the boundary of $\mathcal{C}$. The proof is simple, so we omit it. In Figure 7.3b we plot the capacity region, the point $\boldsymbol{\nu}$ and the cone $\mathcal{K}$ where SSC occurs. Observe that the cone $\mathcal{K}$ is two-dimensional and, therefore, this ad hoc wireless network experiences full-dimensional SSC.

Remark 7.12. *The input-queued switch can be modeled similarly to an ad hoc wireless network. However, the input-queued switch cannot experience full-dimensional SSC since all the vertices of its capacity region are on the coordinate axes. In other words, all the vertices in the capacity region of the input-queued switch require that the arrival rate to (at least) one queue is zero. This is equivalent to considering a queueing system where the zero-arrival rate queue does not exist, which already has a lower-dimensional state space.*

Figure 7.4: Diagrams of examples of parallel-server systems. The dotted lines represent the compatibility between job-types and servers.



(a) Fully flexible servers. (b) Dedicated servers. (c) $\mathcal{N}$-system.

Parallel-server system.

Consider a parallel-server system as follows. There are n types of jobs that arrive according to arrival processes as described in section 6.3. Each job type can be processed by a subset of servers, and these subsets are modeled by a compatibility graph. In Figure 7.4 we present three examples of parallel-server systems, where the dotted lines represent the compatibility of the job types with the servers. In Figure 7.4a, all jobs can be served by all servers (fully flexible system), in Figure 7.4b each job can be processed by only one server (dedicated system), and in Figure 7.4c, the jobs from the first queue can be processed by any server and the jobs from the second queue can only be processed by the second server ($\mathcal{N}$-system, to be studied in subsubsection 7.5.2). The parallel-server systems (also called process flexibility) have received plenty of attention in the literature [106, 107, 7, 3, 108, 13]. However, most of the prior work is under the CRP condition. In this section we show that the parallel-server system can be studied as an immediate application of Theorem 7.5, regardless of the CRP condition being satisfied.

To model a parallel-server system as a generalized switch, we assume that the service rate offered by each server in each time slot is a random variable that may depend on the service rate of other servers, but it is independent of the arrival and queueing processes. The joint distribution of the offered service rates is known, and we assume its state space is finite. Hence, the joint distribution of the offered service can be modeled as the channel state, and the compatibility graph determines the feasible service rate vectors in each time

slot. Since we need the set of feasible service rate vectors in each channel state to be finite, we only consider the maximal vectors and their projection on the coordinate axes. Once the offered service rates are observed, the scheduler follows MaxWeight algorithm to decide which job types will be served and at which rate. We obtain the following result.

Corollary 7.13. *Consider a set of parallel-server systems as described above, parametrized by ϵ as described in Theorem 7.5. Suppose the capacity region $\mathcal{C}$ has vertices that do not lie on the coordinate axes, and that $\boldsymbol{w}$ is one of them. Let Σ_B be as in Theorem 7.5 and $\boldsymbol{\sigma}_a$ be as in Corollary 7.10. Then,*

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle\right] = \frac{1}{2}\left(\|\boldsymbol{\sigma}_a\|^2 + \mathbf{1}^T\left((C^T C)^{-1} \circ \Sigma_B\right)\mathbf{1}\right).$$

The proof of Corollary 7.13 only requires modeling the parallel-server system as a generalized switch as we showed above, so we omit it.

Remark 7.14. *In Corollary 7.13 we considered a vector $\boldsymbol{w}$ in a vertex of the capacity region. However, Theorem 7.5 is immediately applicable for any $\boldsymbol{w}$ in the boundary of the capacity region. Here we focused on a special case to illustrate the full-dimensional SSC result.*

Before finishing this section we present one of the simplest parallel-server systems to illustrate the result in Corollary 7.13. Specifically, we work with a dedicated system, where every job type can be processed by exactly one server. A diagram with three job-types and three servers is presented in Figure 7.4b.

Consider an SPN with n servers, each with its own queue. Let $\{\hat{\boldsymbol{s}}(k) : k \in \mathbb{Z}_+\}$ be a sequence of i.i.d. random vectors, such that $\hat{s}_i(k)$ is the potential service in queue i in time slot k. Let $\boldsymbol{\mu} = \mathbb{E}\left[\hat{\boldsymbol{s}}(1)\right]$ and Σ_s be the covariance matrix of $\hat{\boldsymbol{s}}(1)$. Suppose the vector $\hat{\boldsymbol{s}}(1)$ has finite state space and that $\hat{s}_i(1) \leq S_{\max}$ with probability 1 for all $i \in [n]$. Suppose $\min_{i\in[n]} \mu_i > 0$.

The arrival process is defined as in section 6.3, and we model heavy traffic as described at the beginning of section 7.4. Specifically, let $\epsilon \in (0,1)$ be the heavy-traffic parameter. Then, for each $\epsilon \in (0,1)$ and each $i \in [n]$, let the arrival process to the system be $\left\{\boldsymbol{a}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$, which is a sequence of i.i.d. random vectors with mean $\boldsymbol{\lambda}^{(\epsilon)} = \mathbb{E}\left[\boldsymbol{a}^{(\epsilon)}(1)\right] = (1-\epsilon)\boldsymbol{\mu}$ and covariance matrix $\Sigma_a^{(\epsilon)}$.

Corollary 7.15. *Consider a set of dedicated parallel-server systems as described above, parametrized by $\epsilon \in (0,1)$ as described in Theorem 7.5. Suppose $\lim_{\epsilon \downarrow 0} \Sigma_a^{(\epsilon)} = \Sigma_a$ component-wise. Let $\sigma_{a_i}^2 = (\Sigma_a)_{i,i}$ and $\sigma_{s_i}^2 = (\Sigma_s)_{i,i}$ for each $i \in [n]$. Then,*

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i=1}^{n} \mu_i \overline{q}_i^{(\epsilon)}\right] = \frac{1}{2}\sum_{i=1}^{n}\left(\sigma_{a_i}^2 + \sigma_{s_i}^2\right).$$

From the discussion after Corollary 7.13, we expected that the correlation among the arrival processes would not be part of the right-hand side of the limit. However, observe that the correlation among the service processes does not appear in the answer either. Then, even though the arrival and potential service processes are correlated among queues, the linear combination of the mean queue lengths behaves as if the queues were independent. Moreover, Corollary 7.15 recovers Kingman's bound. We present the proof below.

Proof of Corollary 7.15. The capacity region of this queueing system is

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}_+^n : x_i \leq \mu_i \ , \ i \in [n]\right\}.$$

Then, we have $L = n$, and for each $i \in [n]$ we set $\boldsymbol{c}^{(i)} = \boldsymbol{e}^{(i)}$ and $b^{(i)} = \mu_i$. Therefore, the matrix C is the identity matrix, which implies that the projection matrix H is also the identity matrix.

Let $P = [n]$. Then, $\cap_{\ell \in P} \mathcal{F}^{(\ell)} = \{\boldsymbol{\mu}\}$, and the left-hand side of Equation 7.17 yields

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i=1}^{n} \mu_i \overline{q}_i^{(\epsilon)}\right]$$

Since the projection matrix satisfies $H = \mathbb{I}$, the first term on the right-hand side of Equation 7.17 yields

$$\frac{1}{2}\mathbf{1}^T \left(H \circ \Sigma_a\right) \mathbf{1} = \frac{1}{2}\mathbf{1}^T \left(\mathbb{I} \circ \Sigma_a\right) \mathbf{1} = \frac{1}{2}\sum_{i=1}^{n} \sigma_{a_i}^2.$$

To compute the second term of the right-hand side of Equation 7.17, we consider the following interpretation of the channel state. Let $\mathcal{M}$ be an enumeration of the elements of the state space of $\hat{\boldsymbol{s}}(1)$, and $\boldsymbol{s}^{(m)}$ be its m^{th} element for each $m \in \mathcal{M}$. For each $m \in \mathcal{M}$, let the set of feasible service rate vectors in channel state m be

$$\mathcal{S}^{(m)} = \left\{\boldsymbol{s}^{(m)}\right\} \cup \left\{\boldsymbol{s}^{(m)} - s_i^{(m)}\boldsymbol{e}^{(i)} : \; i \in [n]\right\},$$

i.e., the set $\mathcal{S}^{(m)}$ contains $\boldsymbol{s}^{(m)}$ and its projection on the coordinate axes. We assume that MaxWeight breaks ties by choosing maximal schedules. Then, if the channel state is m then the service rates vector is always $\boldsymbol{s}^{(m)}$. With this assumption we lose some generality because arrivals occur after deciding the optimal schedule. However, we are interested in heavy-traffic analysis so this slight loss of generality does not affect our result. Then, the probability mass function of the channel state $\boldsymbol{\psi}$ satisfies $\psi_m \stackrel{\triangle}{=} \mathbb{P}\left[\hat{\boldsymbol{s}}(1) = \boldsymbol{s}^{(m)}\right]$ for each $m \in \mathcal{M}$.

By definition of $b^{(m,\ell)}$ in Equation 6.4 and by definition of the sets $\mathcal{S}^{(m)}$ and the vectors $\boldsymbol{c}^{(\ell)}$ above, we obtain that for each $\ell \in [n]$ we have

$$b^{(m,\ell)} = \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{s}^{(m)} \rangle = \langle \boldsymbol{e}^{(\ell)}, \boldsymbol{s}^{(\ell)} \rangle = s_\ell^{(m)}.$$

Then, for each $\ell \in [n]$ the random variable $B_\ell(1)$ is such that $\mathbb{P}\left[B_\ell(1) = s_\ell^{(m)}\right] = \psi_m$ and $\mathbb{E}\left[B_\ell(1)\right] = \mu_\ell$. Therefore, the vectors $(B_1(1), \ldots, B_n(1))$ and $\hat{\boldsymbol{s}}(1)$ have the same distribution. Hence, $(\Sigma_B)_{i,j} = \text{Cov}\left[\hat{\boldsymbol{s}}_i, \hat{\boldsymbol{s}}_j\right]$, and the second term in the right-hand side of

Equation 7.17 becomes

$$\frac{1}{2}\mathbf{1}^T\left((C^TC)^{-1}\circ\Sigma_B\right)\mathbf{1} \overset{(a)}{=} \frac{1}{2}\mathbf{1}^T\left(\mathbb{I}\circ\Sigma_B\right)\mathbf{1} \overset{(b)}{=} \sum_{i=1}^{n}\sigma^2_{s_i},$$

where (a) holds because $C = \mathbb{I}$; and (b) holds by definition of Hadamard's product and because the diagonal of Σ_B contains the variance of $\hat{s}_i(1)$ for each $i \in [n]$. □

$\mathcal{N}$-system.

The $\mathcal{N}$-system model is a parallel-server system with two servers and two job types. One of the servers exclusively serves the jobs type 1, and the other server can process both. A diagram of the $\mathcal{N}$-system is presented in Figure 7.4c. According to [109], "the $\mathcal{N}$-system is one of the simplest parallel server system models that retains much of the complexity inherent in more general models". Consequently, it has received plenty of attention over the years and there is vast literature that only focuses on its performance under the CRP condition [106, 107, 7, 108]. Theorem 7.5 is immediately applicable to this system, and gives information about the mean queue lengths in both, the CRP and non-CRP cases. In this section we focus on the non-CRP case.

Let the arrival processes be as described in section 6.3 and suppose that each server processes jobs at rate 1. Then, the capacity region of this system is

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_1 \leq 1,\, x_2 \leq 1\right\}.$$

We consider the heavy-traffic parametrization $\boldsymbol{\lambda}^{(\epsilon)} = (1-\epsilon)\mathbf{1}$, for $\epsilon \in (0,1)$. Then, as $\epsilon \downarrow 0$, the arrival rate vector approaches a vertex of the capacity region and, hence, the $\mathcal{N}$-system experiences full-dimensional SSC. We now present the result formally.

Corollary 7.16. *Consider a set of $\mathcal{N}$-systems parametrized by $\epsilon \in (0,1)$, as described*

above. Let $\boldsymbol{\sigma}_a$ be as in 7.13. Then,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1^{(\epsilon)} + \overline{q}_2^{(\epsilon)}\right] = \frac{\sigma_{a_1}^2 + \sigma_{a_2}^2}{2}.$$

Corollary 7.16 can be easily proved as an immediate application of Corollary 7.10. Here we present an alternative proof, where we explicitly set to zero the drift of the function $V(\boldsymbol{q}) = \|\boldsymbol{q}\|^2$. Recall that, since SSC occurs into a two-dimensional subspace, in this case we have $\boldsymbol{q} = \boldsymbol{q}_{\|\mathcal{H}}$.

Alternative proof of Corollary 7.16. We set to zero the drift of $V(\boldsymbol{q}) = \|\boldsymbol{q}\|^2 = q_1^2 + q_2^2$. We obtain

$$\begin{aligned}
0 =& \mathbb{E}\left[(\overline{q}_1 + \overline{a}_1 - \overline{s}_1 + \overline{u}_1)^2 + (\overline{q}_2 + \overline{a}_2 - \overline{s}_2 + \overline{u}_2)^2 - \overline{q}_1^2 - \overline{q}_2^2\right] \\
=& \mathbb{E}\left[\overline{a}_1^2\right] + \mathbb{E}\left[\overline{s}_1^2\right] - 2\mathbb{E}\left[\overline{a}_1\right]\mathbb{E}\left[\overline{s}_1\right] + \mathbb{E}\left[\overline{a}_2^2\right] + \mathbb{E}\left[\overline{s}_2^2\right] - 2\mathbb{E}\left[\overline{a}_2\right]\mathbb{E}\left[\overline{s}_2\right] \\
& + 2\mathbb{E}\left[\overline{q}_1\left(\overline{a}_1 - \overline{s}_1\right)\right] + 2\mathbb{E}\left[\overline{q}_2(\overline{a}_2 - \overline{s}_2)\right] - \mathbb{E}\left[\overline{u}_1^2\right] - \mathbb{E}\left[\overline{u}_2^2\right] \\
=& \sigma_{a_1}^2 + \sigma_{a_2}^2 + \mathbb{E}\left[(1 - \epsilon - \overline{s}_1)^2\right] + \mathbb{E}\left[(1 - \epsilon - \overline{s}_2)^2\right] - 2\epsilon\mathbb{E}\left[\overline{q}_1\right] - 2\epsilon\mathbb{E}\left[\overline{q}_2\right] \qquad (7.23) \\
& + 2\mathbb{E}\left[\overline{q}_1(1 - \overline{s}_1)\right] + 2\mathbb{E}\left[\overline{q}_2(1 - \overline{s}_2)\right] - \mathbb{E}\left[\overline{u}_1^2\right] - \mathbb{E}\left[\overline{u}_2^2\right],
\end{aligned}$$

where the last equality holds by definition of variance and because $\mathbb{E}\left[\overline{a}_i\right] = 1 - \epsilon$. Now we compute each of the terms of Equation 7.23.

We start with the quadratic terms on the first line.

Claim 7.17. $\mathbb{E}\left[1 - \epsilon - \overline{s}_1)^2\right] + \mathbb{E}\left[(1 - \epsilon - \overline{s}_2)^2\right]$ *is* $O(\epsilon)$.

Proof. To prove it, define $\phi = \mathbb{P}\left[\overline{\boldsymbol{s}} = (1, 1)\right]$. Then, observe that for $i \in \{1, 2\}$, we have

$$\mathbb{E}\left[\overline{s}_i\right] = \phi + 0(1 - \phi) = \phi.$$

On the other hand, by stability we know $\mathbb{E}\left[\overline{s}_i\right] \geq \mathbb{E}\left[\overline{a}_i\right] = 1 - \epsilon$. Putting both together and

rearranging terms we obtain

$$1 - \phi \leq \epsilon,$$

which means that $1 - \phi$ is $O(\epsilon)$. Therefore, we have

$$\mathbb{E}\left[(1 - \epsilon - \overline{s}_i)^2\right] = \phi\epsilon^2 + (1 - \phi)(1 - \epsilon)^2 = O(\epsilon).$$

□

Now we show $\mathbb{E}\left[\overline{u}_1^2\right] + \mathbb{E}\left[\overline{u}_2^2\right]$ is $O(\epsilon)$.

Claim 7.18. $\mathbb{E}\left[\overline{u}_1^2\right] + \mathbb{E}\left[\overline{u}_2^2\right]$ *is* $O(\epsilon)$

Proof. By definition of unused service, and because potential service is bounded above by 2, we have

$$0 \leq \mathbb{E}\left[\overline{u}_i^2\right] \leq 2\mathbb{E}\left[\overline{u}_i\right] \qquad \forall i \in \{1, 2\}.$$

Then, it suffices to show that $\mathbb{E}\left[\overline{u}_i\right]$ is $O(\epsilon)$ for $i = 1, 2$. To do that, we set to zero the drift of the linear function $V_i(\boldsymbol{q}) = q_i$, for $i = 1, 2$. We obtain

$$\begin{aligned} 0 =& \mathbb{E}\left[(\overline{q}_i + \overline{a}_i - \overline{s}_i + \overline{u}_i) - \overline{q}_i\right] \\ \overset{(a)}{=}& (1 - \epsilon) - \mathbb{E}\left[\overline{s}_i\right] + \mathbb{E}\left[\overline{u}_i\right] \\ =& -\epsilon - \mathbb{E}\left[\overline{s}_i - 1\right] + \mathbb{E}\left[\overline{u}_i\right], \end{aligned}$$

where (a) holds because $\mathbb{E}\left[\overline{a}_i\right] = 1 - \epsilon$. From the proof of Claim 7.17, we know that $\mathbb{E}\left[1 - \overline{s}_i\right] = 1 - \phi$ is $O(\epsilon)$. Therefore, rearranging terms we obtain

$$\mathbb{E}\left[\overline{u}_i\right] = \epsilon + O(\epsilon).$$

This concludes the proof of the claim. □

To compute the remaining terms we use MaxWeight algorithm. We know

$$\mathbb{E}\left[\overline{q}_1(1-\overline{s}_1)+\overline{q}_2(1-\overline{s}_2)\right]\leq 0, \tag{7.24}$$

because $(1,1)$ is one of the two feasible schedules. It remains to show a lower bound that increases to 0 as $\epsilon \downarrow 0$. We prove such result in the following claim, and observe the proof is based on SSC.

Claim 7.19. $\mathbb{E}\left[\overline{q}_1(1-\overline{s}_1)+\overline{q}_2(1-\overline{s}_2)\right]$ *is* $\Theta(\sqrt{\epsilon})$.

Proof. We already know Equation 7.24. Now we show a lower bound that is $\Theta(\sqrt{\epsilon})$ using SSC to the cone $\mathcal{K}$.

Consider a vector $\boldsymbol{x}\in\mathbb{R}^2_+$. If $\boldsymbol{x}\notin\mathcal{K}$, then projecting $\boldsymbol{x}$ on the cone $\mathcal{K}$ is equivalent to projecting it on the line $\mathcal{L}=\left\{\boldsymbol{y}\in\mathbb{R}^2_+:\ y_1=y_2\right\}$. Then, if $\boldsymbol{x}\notin\mathcal{K}$ we have

$$\boldsymbol{x}_{\parallel\mathcal{K}}=\left(\frac{x_1+x_2}{2},\frac{x_1+x_2}{2}\right)\quad,\quad \boldsymbol{x}_{\perp\mathcal{K}}=\frac{x_2-x_1}{2}\left(-1,1\right). \tag{7.25}$$

Then, we have

$$\begin{aligned}
0\geq&\mathbb{E}\left[\overline{q}_1(1-\overline{s}_1)+\overline{q}_2(1-\overline{s}_2)\right]\\
\overset{(a)}{=}&\mathbb{E}\left[\left(\overline{q}_1(1-\overline{s}_1)+\overline{q}_2(1-\overline{s}_2)\right)\mathbb{1}_{\{\overline{\boldsymbol{s}}=(2,0)\}}\right]\\
=&\mu_1\mathbb{E}\left[(\overline{q}_1-\overline{q}_2)\mathbb{1}_{\{\overline{\boldsymbol{s}}=(0,\mu_1+\mu_2)\}}\right]\\
\overset{(b)}{=}&\mathbb{E}\left[(\overline{q}_1-\overline{q}_2)\mathbb{1}_{\{\overline{q}_2<\overline{q}_1\}}\right]\\
\overset{(c)}{=}&-\sqrt{2}\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp\|\ \mathbb{1}_{\{\overline{q}_2<\overline{q}_1\}}\right]\\
\overset{(d)}{\geq}&-\sqrt{2}\sqrt{\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp\|^2\right](1-\phi)}\\
\overset{(e)}{\geq}&-\sqrt{2}J_1\sqrt{1-\phi}
\end{aligned}$$

where (a) holds because the set of feasible service rates is $\mathcal{S} = \{(1,1),\,(2,0)\}$; (b) holds because, since we are using MaxWeight algorithm, the events $\{\overline{\boldsymbol{s}} = (2,0)\}$ and $\{\overline{q}_2 < \overline{q}_1\}$ are equivalent; (c) holds by definition of Euclidean norm and by Equation 7.25; (d) holds by Cauchy-Schwarz inequality; and (e) holds by SSC as established in Proposition 7.4.

From the proof of Claim 7.17, we know that $(1-\phi)$ is $O(\epsilon)$. Therefore, we have completed the proof. □

Putting all the claims together in Equation 7.23, we obtain

$$2\epsilon\mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] = \sigma_1^2 + \sigma_2^2 + f(\epsilon),$$

where $f(\epsilon)$ goes to zero as $\epsilon \downarrow 0$, and so we get

$$\lim_{\epsilon\downarrow 0} \epsilon\mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] = \frac{\sigma_1^2 + \sigma_2^2}{2}.$$

□

Remark 7.20. *Observe that we only use SSC in the last step of the proof. This implies that, if we did not have a SSC result, we can still obtain an upper bound. In fact, using Equation 7.24 along with Claim 7.17 and Claim 7.18 in Equation 7.23, we obtain*

$$2\epsilon\mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] \leq \sigma_1^2 + \sigma_2^2 + f(\epsilon),$$

where $f(\epsilon)$ goes to zero as $\epsilon \downarrow 0$, and so we get

$$\lim_{\epsilon\downarrow 0} \epsilon\mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] \leq \frac{\sigma_1^2 + \sigma_2^2}{2}.$$

This example shows that, in order to obtain bounds that are asymptotically tight in heavy traffic, we must use SSC.

7.6 Proof of Theorem 7.5.

In this section we present the proof of the main theorem of this chapter. We use the notation $\mathbb{E}_m[\,\cdot\,] = \mathbb{E}\left[\,\cdot\,\middle|\overline{M} = m\right]$, and we omit the dependence on ϵ of the variables for simplicity of exposition.

Proof of Theorem 7.5. First observe that $\langle \overline{\boldsymbol{q}}, \boldsymbol{w}\rangle = \langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle$. To show this statement, define $\boldsymbol{w}_\perp \stackrel{\triangle}{=} \boldsymbol{w} - \boldsymbol{\nu}$ for all $\boldsymbol{w} \in \cap_{\ell\in P}\mathcal{F}^{(\ell)}$, and observe that $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{w}_\perp\rangle = 0$ because both $\boldsymbol{\nu}, \boldsymbol{w} \in \mathcal{F}^{(\ell)}$ for all $\ell \in P$. Then,

$$\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle = \langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{w} - \boldsymbol{w}_\perp\rangle = \langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{w}\rangle \stackrel{(a)}{=} \langle \overline{\boldsymbol{q}}, \boldsymbol{w}\rangle,$$

where (a) holds because $\boldsymbol{w} \in \cap_{\ell\in P}\mathcal{F}^{(\ell)}$ and because $\overline{\boldsymbol{q}}_{\|\mathcal{H}} = \overline{\boldsymbol{q}} - \overline{\boldsymbol{q}}_{\perp\mathcal{H}}$. Hence, in the rest of the proof we focus on computing bounds for $\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle\right]$.

We set to zero the drift of $V_{\|\mathcal{H}}(\boldsymbol{q}) = \left\|\boldsymbol{q}_{\|\mathcal{H}}\right\|^2$, and bound the terms that arise one by one. Before setting the drift to zero we need to make sure that $\mathbb{E}\left[V_{\|\mathcal{H}}(\overline{\boldsymbol{q}}_{\|\mathcal{H}})\right]$ is finite. This result can be proved using the Foster-Lyapunov theorem with Lyapunov function $Z(\boldsymbol{q}) = \|\boldsymbol{q}\|^2$. This proves that $\mathbb{E}\left[\|\overline{\boldsymbol{q}}\|^2\right]$ is finite. Then, since projection is nonexpansive we have $\mathbb{E}\left[\left\|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\right\|^2\right]$ is also finite. The proof is simple, so we omit the details for ease of exposition. Now, setting to zero the drift of $V_{\|\mathcal{H}}(\boldsymbol{q})$ we obtain

$$\begin{aligned} 0 &= \mathbb{E}\left[\left\|\overline{\boldsymbol{q}}^+_{\|\mathcal{H}}\right\|^2 - \left\|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\right\|^2\right] \\ &\stackrel{(a)}{=} \mathbb{E}\left[\left\|\overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\right\|^2 + 2\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}},\, \overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\rangle - \left\|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\right\|^2 + 2\langle \overline{\boldsymbol{q}}^+_{\|\mathcal{H}},\, \overline{\boldsymbol{u}}_{\|\mathcal{H}}\rangle\right] \end{aligned} \tag{7.26}$$

where (a) holds by the dynamics of the queues presented in Equation 1.2, and reorganizing terms. Let

$$\mathcal{T}_1 \stackrel{\triangle}{=} 2\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}}_{\|\mathcal{H}} - \overline{\boldsymbol{a}}_{\|\mathcal{H}}\rangle\right], \quad \mathcal{T}_2 \stackrel{\triangle}{=} \mathbb{E}\left[\left\|\overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\right\|^2\right],$$

$$\mathcal{T}_3 \stackrel{\triangle}{=} \mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\right\|^2\right] \quad \text{and} \quad \mathcal{T}_4 \stackrel{\triangle}{=} 2\mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}},\, \overline{\boldsymbol{u}}_{\|\mathcal{H}}\rangle\right].$$

Then, reorganizing the terms in Equation 7.26 we obtain $\mathcal{T}_1 = \mathcal{T}_2 - \mathcal{T}_3 + \mathcal{T}_4$. We compute each term separately. We start with $\mathcal{T}_1$.

$$\begin{aligned}
\mathcal{T}_1 &\stackrel{(a)}{=} 2\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \overline{\boldsymbol{a}}\rangle\right] \\
&\stackrel{(b)}{=} 2\epsilon\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle\right] + \mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}\rangle\right] \\
&\stackrel{(c)}{=} 2\epsilon\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle\right] + O\left(\epsilon \log\left(\tfrac{1}{\epsilon}\right)\right),
\end{aligned} \tag{7.27}$$

where (a) holds by the orthogonality principle; (b) holds because $\mathbb{E}\left[\overline{\boldsymbol{a}}\right] = (1-\epsilon)\boldsymbol{\nu}$ and because $\overline{\boldsymbol{a}}$ is independent of the vector of queue lengths; and (c) holds by Claim 7.21 stated below.

Claim 7.21. *Consider a set of generalized switches as described in Theorem 7.5. Then, there exists $\epsilon_0' \in (0,1)$ and a finite constant $\zeta_1 > 0$ such that*

$$\left|\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}\rangle\right]\right| \leq \zeta_1 \epsilon \log\left(\tfrac{1}{\epsilon}\right) \qquad \forall \epsilon < \epsilon_0'.$$

We present the proof of Claim 7.21 in subsection 7.7.1. Now we compute $\mathcal{T}_2$. Expanding the product we obtain

$$\mathcal{T}_2 = \mathbb{E}\left[\left\|\overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\right\|^2\right] = \mathbb{E}\left[\left\|\overline{\boldsymbol{a}}_{\|\mathcal{H}}\right\|^2\right] + \mathbb{E}\left[\left\|\overline{\boldsymbol{s}}_{\|\mathcal{H}}\right\|^2\right] - 2\mathbb{E}\left[\langle \overline{\boldsymbol{a}}_{\|\mathcal{H}},\, \overline{\boldsymbol{s}}_{\|\mathcal{H}}\rangle\right]. \tag{7.28}$$

We compute each term in Equation 7.28 separately. For the first two terms, we solve the least squares problem and we use the projection matrix on the subspace $\mathcal{H}$, denoted as H. Let $h_{i,j}$ be its element (i,j) for each $i,j \in [n]$. For the first term we have

$$\mathbb{E}\left[\left\|\overline{\boldsymbol{a}}_{\|\mathcal{H}}\right\|^2\right] = \mathbb{E}\left[\left\|H\,\overline{\boldsymbol{a}}\right\|^2\right]$$

$$\stackrel{(a)}{=} \sum_{i=1}^{n}\sum_{i'=1}^{n} h_{i,i'}\text{Cov}\left[\overline{a}_i, \overline{a}_{i'}\right] + \sum_{i=1}^{n}\sum_{i'=1}^{n} h_{i,i'}\mathbb{E}\left[\overline{a}_i\right]\mathbb{E}\left[\overline{a}_{i'}\right]$$
$$\stackrel{(b)}{=} \mathbf{1}^T\left(H \circ \Sigma_a^{(\epsilon)}\right)\mathbf{1} + (1-\epsilon)^2\boldsymbol{\nu}^T H\boldsymbol{\nu}, \tag{7.29}$$

where (a) holds solving the least squares problem, by definition of norm, because H is a projection matrix (and therefore $H = H^T = H^2$), and by definition of covariance; and (b) holds by definition of the Hadamard's product and because $\mathbb{E}\left[\overline{a}_i\right] = \lambda_i^{(\epsilon)} = (1-\epsilon)\nu_i$ for each $i \in [n]$. For the second term in Equation 7.28 we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\boldsymbol{s}}_{\|\mathcal{H}}\right\|^2\right] &= \mathbb{E}\left[\|H\overline{\boldsymbol{s}}\|^2\right] \\
&\stackrel{(a)}{=} \mathbb{E}\left[\overline{\boldsymbol{s}}^T C(C^TC)^{-1}C^T\overline{\boldsymbol{s}}\right] \\
&\stackrel{(b)}{=} \sum_{\ell_1\in\tilde{P}}\sum_{\ell_2\in\tilde{P}}(C^TC)^{-1}_{\ell_1,\ell_2}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell_1)}, \overline{\boldsymbol{s}}\rangle\langle \boldsymbol{c}^{(\ell_2)}, \overline{\boldsymbol{s}}\rangle\right] \\
&\stackrel{(c)}{=} \sum_{\ell_1\in\tilde{P}}\sum_{\ell_2\in\tilde{P}}(C^TC)^{-1}_{\ell_1,\ell_2}\sum_{m\in\mathcal{M}}\psi_m\mathbb{E}_m\left[\langle \boldsymbol{c}^{(\ell_1)}, \overline{\boldsymbol{s}}\rangle\langle \boldsymbol{c}^{(\ell_2)}, \overline{\boldsymbol{s}}\rangle\right] \\
&\stackrel{(d)}{=} \mathbf{1}^T\left((C^TC)^{-1} \circ \Sigma_B\right)\mathbf{1} + \boldsymbol{\nu}^T H\boldsymbol{\nu} - O(\epsilon),
\end{aligned} \tag{7.30}$$

where $(C^TC)^{-1}_{\ell_1,\ell_2}$ is the element (ℓ_1, ℓ_2) of the matrix $(C^TC)^{-1}$ for each $\ell_1, \ell_2 \in \tilde{P}$. Here, (a) holds solving the least squares problems, and because $H = C(C^TC)^{-1}C^T$ by definition of projection matrix; (b) holds by definition of matrix multiplication, and because $C^T\overline{\boldsymbol{s}}$ is a vector with elements $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle$ for $\ell \in \tilde{P}$; (c) holds by law of total probability, conditioning on the channel state; and (d) holds using Lemma 7.2, the definition of covariance and reorganizing the terms. Now we compute the last term in Equation 7.28. We obtain

$$\begin{aligned}
-2\mathbb{E}\left[\langle \overline{\boldsymbol{a}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}}_{\|\mathcal{H}}\rangle\right] &\stackrel{(a)}{=} -2\mathbb{E}\left[\overline{\boldsymbol{a}}^T H\overline{\boldsymbol{s}}\right] \\
&\stackrel{(b)}{=} -2(1-\epsilon)\boldsymbol{\nu}^T\mathbb{E}\left[H\overline{\boldsymbol{s}}\right] \\
&\stackrel{(c)}{=} -2(1-\epsilon)\boldsymbol{\nu}^T H\boldsymbol{\nu} + O(\epsilon),
\end{aligned} \tag{7.31}$$

where (a) holds because for any vector $\boldsymbol{x}$, we have $\boldsymbol{x}_{\|\mathcal{H}} = H\boldsymbol{x}$ by the solution of the least squares problem, and because H is a projection matrix; (b) holds because $\overline{\boldsymbol{a}}$ is independent of $\overline{\boldsymbol{s}}$ and $\mathbb{E}\left[\overline{\boldsymbol{a}}\right] = \boldsymbol{\lambda}^{(\epsilon)} = (1-\epsilon)\boldsymbol{\nu}$; and (c) because $H = C\left(C^TC\right)^{-1}C^T$, because $C^T\overline{\boldsymbol{s}}$ has elements $\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle$ with $\ell \in \tilde{P}$, by Lemma 7.1 and Lemma 7.2, and because $\boldsymbol{\nu} \in \mathcal{F}^{(\ell)}$. Therefore, using Equation 7.29, Equation 7.30 and Equation 7.31 in Equation 7.28 we obtain

$$\left|\mathcal{T}_2 - \left(\mathbf{1}^T\left(H \circ \Sigma_a^{(\epsilon)}\right)\mathbf{1} + \mathbf{1}^T\left((C^TC)^{-1} \circ \Sigma_B\right)\mathbf{1} + \epsilon^2\boldsymbol{\nu}^TH\boldsymbol{\nu}\right)\right| \text{ is } O(\epsilon).$$

In other words, there exists a finite constant $\zeta_2 > 0$ such that

$$\left|\mathcal{T}_2 - \left(\mathbf{1}^T\left(H \circ \Sigma_a^{(\epsilon)}\right)\mathbf{1} + \mathbf{1}^T\left((C^TC)^{-1} \circ \Sigma_B\right)\mathbf{1} + \epsilon^2\boldsymbol{\nu}^TH\boldsymbol{\nu}\right)\right| \leq \zeta_2\epsilon. \tag{7.32}$$

Now we compute $\mathcal{T}_3$. We obtain

$$\begin{aligned} 0 \leq \mathcal{T}_3 &= \mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\right\|^2\right] \\ &\overset{(a)}{\leq} \sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle^2\right] \\ &\overset{(b)}{\leq} nS_{\max}C_{\max}\sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] \overset{(c)}{=} O(\epsilon), \end{aligned}$$

where $C_{\max} = \max\limits_{\ell\in P, i\in[n]}\left\{c_i^{(\ell)}\right\}$ and it is a finite constant. Here, (a) holds because the vectors $\boldsymbol{c}^{(\ell)}$ are not necessarily orthogonal for all $\ell \in P$; (b) holds because $\overline{\boldsymbol{u}} \leq \overline{\boldsymbol{s}} \leq S_{\max}\mathbf{1}$ with probability 1 by definition of the unused service; and (c) holds by Claim 7.22.

Claim 7.22. *Consider a set of generlized switches, as described in Theorem 7.5. Then,*

$$\sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] = \epsilon\sum_{\ell\in P}b^{(\ell)} - O(\epsilon). \tag{7.33}$$

We present the proof of Claim 7.22 in subsection 7.7.2. Therefore, $\mathcal{T}_3$ is $O(\epsilon)$, which

implies the existence of a finite constant $\zeta_3 > 0$ such that

$$\mathcal{T}_3 \leq \zeta_3 \epsilon. \tag{7.34}$$

Finally, we compute a bound for $\mathcal{T}_4$ as follows.

Claim 7.23. *Consider the system described in Theorem 7.5. Then, there exist $\epsilon_0'' \in (0,1)$ and a finite constant ζ_4 such that*

$$\mathcal{T}_4 \leq \zeta_4 \epsilon \log\left(\frac{1}{\epsilon}\right) \quad \forall \epsilon < \epsilon_0''. \tag{7.35}$$

We provide the proof of Claim 7.23 in subsection 7.7.3. Putting equations Equation 7.27, Equation 7.32, Equation 7.34 and Equation 7.35 together we obtain

$$\left| \mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle\right] - \frac{1}{2\epsilon} \mathbf{1}^T \left(H \circ \Sigma_a^{(\epsilon)}\right) \mathbf{1} - \frac{1}{2\epsilon} \mathbf{1}^T \left((C^T C)^{-1} \circ \Sigma_B\right) \mathbf{1} \right| \leq \zeta \log\left(\frac{1}{\epsilon}\right),$$

where $\zeta = \max\{\zeta_1, \zeta_2, \zeta_3, \zeta_4\}$. This completes the proof. □

Clearly, the above result and proof are much more general and more involved than the proof in the special case of an input-queued switch developed in [14, 15]. The bound in Theorem 7.5 is expressed in terms of a general projection of the second moments of arrival and service processes onto the space $\mathcal{H}$, and we obtain a tighter rate of convergence compared to [34, 14, 15].

The key idea in obtaining a logarithmic error bound is in picking the right exponent j in Hölder's inequality while bounding terms $\mathcal{T}_1$ and $\mathcal{T}_4$. We do this by minimizing the upper bound over j (for a fixed ϵ), which gives $j = \left\lfloor \log\left(\frac{1}{\epsilon}\right) \right\rfloor$. The idea of optimizing over the exponent in Hölder's inequality is motivated by [110].

We would like to point out a couple of conceptual differences from the proof in the case of input-queued switch. Firstly, in the proof of asymptotic upper bounds in an input-queued switch, the scheduling policy is not used. This means that for an input-queued switch, any

scheduling policy that exhibits SSC also has the same asymptotic upper bounds. In our proof here, we use the scheduling policy to upper bound the term $\mathcal{T}_1$ in Claim 7.21. Thus, we may not claim that any scheduling policy that exhibits SSC in Proposition 7.4 satisfies the bound in Theorem 7.5. Secondly, while SSC into the cone $\mathcal{K}$ was established in [14, 15] in the case of an input-queued switch, only the weaker result about collapse into the space $\mathcal{H}$ was used to obtain heavy-traffic queue length bounds. In contrast, we use the collapse into the cone $\mathcal{K}$ in the proof of Theorem 7.5 to lower bound the term $\mathcal{T}_1$. Both these differences are due to the fact that $\overline{\boldsymbol{s}}_{\|\mathcal{H}}$ is constant for all maximal schedules $\overline{\boldsymbol{s}} \in \mathcal{S}$ in the case of an input-queued switch, whereas in the case of the generalized switch this is not necessarily true.

7.7 Details of proof of Theorem 7.5

In this section we prove the claims used in the proof of Theorem 7.5.

7.7.1 Proof of Claim 7.21.

Proof of Claim 7.21. Conditioning on the channel state, we get

$$\begin{aligned}\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}\rangle\right] &= \sum_{m\in\mathcal{M}} \psi_m \mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle\right] \\ &\overset{(a)}{=} \sum_{m\in\mathcal{M}} \psi_m \mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right],\end{aligned}$$

where $\boldsymbol{\nu}^{(m)}$ is defined as in Lemma 7.1. Equality (a) holds because $\overline{\boldsymbol{q}}_{\|\mathcal{H}} = \sum_{\ell\in\tilde{P}} \tilde{\xi}_\ell \boldsymbol{c}^{(\ell)}$ for $\tilde{\xi}_\ell \in \mathbb{R}$ for all $\ell \in \tilde{P}$ (by definition of projection on the subspace $\mathcal{H}$) and if the channel state is m we have

$$\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle = b^{(m,\ell)}\right\}} = \left(b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}^{(m)}\rangle\right) \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle = b^{(m,\ell)}\right\}} = 0,$$

where the last equality holds by definition of $\boldsymbol{\nu}^{(m)}$.

It suffices to show that $\mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right]$ is $O\left(\epsilon \log\left(\frac{1}{\epsilon}\right)\right)$ because $\boldsymbol{\psi} = (\psi_m)_{m\in\mathcal{M}}$ is a probability mass function and, therefore, each ψ_m is bounded.

Observe that $\overline{\boldsymbol{q}} = \overline{\boldsymbol{q}}_{\|\mathcal{H}} + \overline{\boldsymbol{q}}_{\perp\mathcal{H}} = \overline{\boldsymbol{q}}_{\|\mathcal{K}} + \overline{\boldsymbol{q}}_{\perp\mathcal{K}}$, thus

$$\begin{aligned}
&\mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right] \\
&= \mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{K}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right] && (7.36)\\
&\quad + \mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\perp\mathcal{K}} - \overline{\boldsymbol{q}}_{\perp\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right]. && (7.37)
\end{aligned}$$

Now, we show that the terms in Equation 7.36 and Equation 7.37 are $O\left(\epsilon \log\left(\frac{1}{\epsilon}\right)\right)$. For Equation 7.36, we have $\mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{K}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right] \leq 0$ by the definition of projection on the cone $\mathcal{K}$ and by definition of $\boldsymbol{\nu}^{(m)}$ and $b^{(m,\ell)}$ in Lemma 7.1. Now, we have

$$\begin{aligned}
0 &\geq \mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\|\mathcal{K}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right] \\
&\overset{(a)}{\geq} -\mathbb{E}_m\left[\langle \overline{\boldsymbol{q}}_{\perp\mathcal{K}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right] \\
&\overset{(b)}{\geq} -\mathbb{E}\left[\|\overline{\boldsymbol{q}}_{\perp\mathcal{K}}\|^j\right]^{\frac{1}{j}} \mathbb{E}_m\left[\left\|\overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\right\|^{j'} \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right]^{\frac{1}{j'}} \\
&\overset{(c)}{\geq} -J_j^{\frac{1}{j}} \mathbb{E}_m\left[\left\|\overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\right\|^{j'} \mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right\}}\right]^{\frac{1}{j'}},
\end{aligned}$$

where $j, j' \in \mathbb{Z}_+$ satisfy $j, j' > 1$ and $\frac{1}{j} + \frac{1}{j'} = 1$. Here, (a) holds because $\overline{\boldsymbol{q}}_{\|\mathcal{K}} = \overline{\boldsymbol{q}} - \overline{\boldsymbol{q}}_{\perp\mathcal{K}}$, and because $\langle \overline{\boldsymbol{q}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \geq 0$ by the definition of MaxWeight in Equation 6.3 and since $\boldsymbol{\nu}^{(m)} \in \mathcal{S}^{(m)}$; (b) holds using Hölder's inequality; and (c) holds by SSC in Proposition 7.4. Now, by the definition of J_j, we have

$$J_j^{\frac{1}{j}} = \left(\left(\frac{8n\Lambda^2}{\delta}\right)^j + \left(8\sqrt{n}\Lambda\right)^j \left(\frac{8\sqrt{n}\Lambda + \delta}{\delta}\right)^j j!\right)^{\frac{1}{j}} \leq \zeta_1'(j!)^{\frac{1}{j}} \overset{(a)}{\leq} \zeta_1' e^{\frac{1}{j}-1} j^{1+\frac{1}{2j}},$$

where $\zeta_1' \triangleq \frac{8n\Lambda^2}{\delta} + 8\sqrt{n}\Lambda\left(\frac{8\sqrt{n}\Lambda+\delta}{\delta}\right)$. Here, (a) follows from Stirling's approximation for

the factorial.

Now we bound the remaining term $\mathbb{E}_m\left[\left\|\overline{\boldsymbol{s}}-\boldsymbol{\nu}^{(m)}\right\|^{j'}\mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\neq b^{(m,\ell)}\right\}}\right]^{\frac{1}{j'}}$ as follows.

$$\begin{aligned}
0 \leq & \mathbb{E}_m\left[\left\|\overline{\boldsymbol{s}}-\boldsymbol{\nu}^{(m)}\right\|^{j'}\mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\neq b^{(m,\ell)}\right\}}\right]^{\frac{1}{j'}} \\
\overset{(a)}{=} & \mathbb{E}_m\left[\left\|\overline{\boldsymbol{s}}-\boldsymbol{\nu}^{(m)}\right\|^{j'}\middle|\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\neq b^{(m,\ell)}\right]^{\frac{1}{j'}}\left(1-\pi^{(m,\ell)}\right)^{\frac{1}{j'}} \\
\overset{(b)}{\leq} & n\left(S_{\max}^{j'}+V_{\max}^{j'}\right)\left(1-\pi^{(m,\ell)}\right)^{\frac{1}{j'}} \overset{(c)}{=} \zeta_2'\epsilon^{\frac{1}{j'}},
\end{aligned} \tag{7.38}$$

where (a) holds by definition of $\pi^{(m,\ell)}$ in Lemma 7.2; (b) holds with $V_{\max}=\max_{m\in\mathcal{M},i\in[n]}\nu_i^{(m)}$; and (c) holds by Lemma 7.2 for $\zeta_2'\overset{\triangle}{=}n\left(S_{\max}^{j'}+V_{\max}^{j'}\right)\frac{b^{(m,\ell)}}{\gamma^{(m)}}$.

Putting everything together, we obtain

$$\begin{aligned}
0 \geq & \mathbb{E}_m\left[\langle\overline{\boldsymbol{q}}_{\|\mathcal{K}},\overline{\boldsymbol{s}}-\boldsymbol{\nu}^{(m)}\rangle\mathbb{1}_{\left\{\langle \boldsymbol{c}^{(\ell)},\overline{\boldsymbol{s}}\rangle\neq b^{(m,\ell)}\right\}}\right] \\
\geq & -\zeta_1'\zeta_2' e^{\frac{1}{j}-1}j^{1+\frac{1}{2j}}\epsilon^{\frac{1}{j'}} \\
\overset{(a)}{=} & -\zeta_1'\zeta_2' e^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}-1}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor^{1+\frac{1}{2\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon \\
\overset{(b)}{\geq} & -2\zeta_1'\zeta_2'\epsilon\log\left(\frac{1}{\epsilon}\right) \quad \forall\epsilon<\epsilon_0',
\end{aligned}$$

where (a) holds after choosing $j\overset{\triangle}{=}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor$; and (b) follows for ϵ_0' as defined below, and because by the definition of floor function we have

$$\begin{aligned}
& \lim_{\epsilon\downarrow 0} e^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}-1}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}} \\
& \leq \left(\lim_{\epsilon\downarrow 0} e^{\frac{1}{\log\left(\frac{1}{\epsilon}\right)-1}-1}\right)\left(\lim_{\epsilon\downarrow 0}\log\left(\frac{1}{\epsilon}\right)^{\frac{1}{2\log\left(\frac{1}{\epsilon}\right)-2}}\right)\left(\lim_{\epsilon\downarrow 0}\epsilon^{-\frac{1}{\log\left(\frac{1}{\epsilon}\right)}}\right) \\
& = \frac{1}{e}\times 1\times e = 1.
\end{aligned}$$

By definition of limit, there exists $\tilde{\epsilon}_0' > 0$ such that for all $\epsilon < \tilde{\epsilon}_0'$ we have

$$e^{\frac{1}{\left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor}-1}\left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor}} \leq 2.$$

The proof that Equation 7.37 is $O\left(\epsilon \log\left(\frac{1}{\epsilon}\right)\right)$ follows similarly by linearity of dot product, Hölder's inequality with $j = \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor$ and Equation 7.38. We omit the details for brevity.

□

7.7.2 Proof of Claim 7.22.

Proof of Claim 7.22. We set to zero the drift of $V_l(\boldsymbol{q}) = \sum_{\ell \in P}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{q}\rangle$. We obtain

$$0 = \mathbb{E}\left[\sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^{+}\rangle - \sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}\rangle\right] = \mathbb{E}\left[\sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}} - \overline{\boldsymbol{s}} + \overline{\boldsymbol{u}}\rangle\right],$$

where the last equality holds by definition of $\overline{\boldsymbol{q}}^{+}$ and by the dynamics of the queues presented in Equation 1.2. Rearranging terms we obtain

$$\begin{aligned}\sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}}\rangle\right] &= \sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] - \sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{a}}\rangle\right] \\ &= \sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] - \sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, (1-\epsilon)\boldsymbol{\nu}\rangle.\end{aligned}$$

But

$$\begin{aligned}&\sum_{\ell\in P}\mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle\right] \\ &= \mathbb{E}\left[\sum_{\ell\in P}\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle - b^{(\ell)}\right)\right] + \sum_{\ell\in P} b^{(\ell)} \\ &\overset{(a)}{=} \sum_{\ell\in P}\sum_{m\in\mathcal{M}}\psi_m \mathbb{E}_m\left[\left(\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle - b^{(m,\ell)}\right)\middle|\, \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\right]\left(1-\pi^{(m,\ell)}\right) + \sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle \\ &\overset{(b)}{=} \sum_{\ell\in P}\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu}\rangle - O(\epsilon),\end{aligned}$$

where (a) holds because $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle = b^{(\ell)}$ for all $\ell \in P.$; and (b) holds by Lemma 7.2. Then, since $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle = b^{(\ell)}$, we have

$$\sum_{\ell \in P} \mathbb{E}\left[\langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{u}} \rangle\right] = \epsilon \sum_{\ell \in P} b^{(\ell)} - O(\epsilon).$$

□

7.7.3 Proof of Claim 7.23

Proof of Claim 7.23. In this proof we use ideas and notation from [34, Equation (56)]. For each $\ell \in P$, let $\mathcal{L}_+^{(\ell)} \triangleq \left\{ i \in [n] : \ c_i^{(\ell)} > 0 \right\}$ and define

$$\widetilde{\boldsymbol{c}}^{(\ell)} = \left[c_i^{(\ell)} \right]_{i \in \mathcal{L}_+^{(\ell)}}, \widetilde{\overline{\boldsymbol{q}}}^{(\ell)} = \left[\overline{q}_i \right]_{i \in \mathcal{L}_+^{(\ell)}} \text{ and } \widetilde{\overline{\boldsymbol{u}}}^{(\ell)} = \left[\overline{u}_i \right]_{i \in \mathcal{L}_+^{(\ell)}}.$$

Then,

$$\begin{aligned} 0 \leq \left| \frac{\mathcal{T}_4}{2} \right| = \left| \mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\| \mathcal{H}}^+, \overline{\boldsymbol{u}}_{\| \mathcal{H}} \rangle \right] \right| &\overset{(a)}{=} \left| \mathbb{E}\left[-\langle \left(\widetilde{\overline{\boldsymbol{q}}}_{\perp \mathcal{H}}^{(\ell)} \right)^+, \widetilde{\overline{\boldsymbol{u}}}^{(\ell)} \rangle \right] \right| \\ &\overset{(b)}{\leq} \mathbb{E}\left[\left\| \left(\widetilde{\overline{\boldsymbol{q}}}_{\perp \mathcal{H}}^{(\ell)} \right)^+ \right\|^j \right]^{\frac{1}{j}} \mathbb{E}\left[\left\| \widetilde{\overline{\boldsymbol{u}}}^{(\ell)} \right\|^{j'} \right]^{\frac{1}{j'}}, \end{aligned}$$

where $j, j' \in \mathbb{Z}_+$ satisfy $j, j' > 1$ and $\frac{1}{j} + \frac{1}{j'} = 1$. Here, (a) follows using the definition of projection on the subspace to substitute $\overline{\boldsymbol{q}}_{\| \mathcal{H}}^+ = \sum_{\ell \in P} \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^+ \rangle \boldsymbol{c}^{(\ell)}$, then the key property Equation 1.3, and that $\left(\widetilde{\overline{\boldsymbol{q}}}^{(\ell)} \right)^+ = \left(\widetilde{\overline{\boldsymbol{q}}}_{\| \mathcal{H}}^{(\ell)} \right)^+ + \left(\widetilde{\overline{\boldsymbol{q}}}_{\perp \mathcal{H}}^{(\ell)} \right)^+$. Then, (b) holds by Hölder's inequality.

Now we bound each of the terms. For the first term we use SSC as presented in Proposition 7.4, and we obtain

$$\mathbb{E}\left[\left\| \left(\widetilde{\overline{\boldsymbol{q}}}_{\perp \mathcal{H}}^{(\ell)} \right)^+ \right\|^j \right]^{\frac{1}{j}} \leq \mathbb{E}\left[\left\| \overline{\boldsymbol{q}}_{\perp \mathcal{H}}^+ \right\|^j \right]^{\frac{1}{j}} \leq J_j^{\frac{1}{j}} \overset{(a)}{\leq} \zeta_1' e^{\frac{1}{j} - 1} j^{1 + \frac{1}{2j}},$$

where (a) holds by Stirling's approximation for the factorial and ζ_1' is defined as in the proof of Claim 7.21. For the second term we obtain

$$0 \leq \mathbb{E}\left[\left\|\widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\right\|^{j'}\right] \overset{(a)}{\leq} \sum_{\ell\in P}\sum_{i\in\mathcal{L}_+^{(\ell)}} \frac{\widetilde{c}_i^{(\ell)}}{\widetilde{c}_i^{(\ell)}}\mathbb{E}\left[\widetilde{u}_i^{j'}\right] \overset{(b)}{\leq} \frac{S_{\max}^{j'-1}}{\widetilde{c}_{\min}}\sum_{\ell\in P}\mathbb{E}\left[\langle\widetilde{\boldsymbol{c}}^{(\ell)}, \widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\rangle\right] \overset{(c)}{\leq} \zeta_3'\epsilon,$$

where $\widetilde{c}_{\min} = \min_{\ell\in P, i\in[n]}\{\widetilde{c}_i^{(\ell)}\}$. Here, (a) follows as all the terms in the summation are non negative; (b) holds by definition of dot product; and (c) follows from [98, Equation (43)] for a finite constant ζ_3', using a similar argument to the properties used to obtain Equation 7.34. Now, pick $j \overset{\triangle}{=} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor$ to get

$$\begin{aligned}
0 &\leq \left|\frac{\mathcal{T}_4}{2}\right| \\
&\leq \zeta_1'\zeta_3'\frac{S_{\max}^{1-\frac{1}{j'}}}{\widetilde{c}_{\min}^{\frac{1}{j'}}}|P|^{\frac{1}{j'}}e^{\frac{1}{j}-1}j^{1+\frac{1}{2j}}\epsilon^{\frac{1}{j'}} \\
&= \zeta_1'\zeta_3' S_{\max}^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\left(\frac{|P|}{\widetilde{c}_{\min}}\right)^{1-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}} e^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}-1}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor^{1+\frac{1}{2\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon \\
&\overset{(a)}{\leq} 2\zeta_1'\zeta_3'\frac{|P|}{\widetilde{c}_{\min}}\epsilon\log\left(\frac{1}{\epsilon}\right) \quad \forall\epsilon < \tilde{\epsilon}_0'',
\end{aligned}$$

where (a) follows as

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0} S_{\max}^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\left(\frac{|P|}{e\widetilde{c}_{\min}}\right)^{1-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}} \\
&\qquad \leq 1\times\frac{|P|}{e\widetilde{c}_{\min}}\times 1\times e = \frac{|P|}{\widetilde{c}_{\min}}.
\end{aligned}$$

Thus, there exists $\tilde{\epsilon}_0'' > 0$ such that for all $\epsilon < \tilde{\epsilon}_0''$ we have

$$S_{\max}^{\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\left(\frac{|P|}{e\widetilde{c}_{\min}}\right)^{1-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}}\epsilon^{-\frac{1}{\left\lfloor\log\left(\frac{1}{\epsilon}\right)\right\rfloor}} \leq \frac{2|P|}{\widetilde{c}_{\min}}.$$

This completes the proof. □

7.8 Individual queue lengths and higher moments in the input-queued switch

In this section we show that the drift method with polynomial test functions does not provide all the information that is necessary to compute the moments of all the linear combinations of the scaled queue lengths in systems that do not satisfy the CRP condition. We do this by presenting an alternate view of the drift method.

In the proof of Theorem 7.5 we use $V(\boldsymbol{q}) = \left\|\boldsymbol{q}_{\|\mathcal{H}}\right\|^2$ as test function to obtain bounds on certain linear combinations of the queue lengths in a generalized switch. This choice of test function was first proposed in [14], and the main reason to use it is that the term $\mathcal{T}_4$ consisting of the 'qu' terms (i.e., cross terms between the queue length and the unused service) converges to zero in the heavy-traffic limit. All of queueing theory in some sense is to get a handle on the unused service terms, and the drift method handles these terms by making sure that they 'cancel out' in heavy traffic, using SSC and our choice of the test function. In this section, instead of trying to cancel out the 'qu' terms, we consider them as unknowns and try to solve for them along with the mean queue lengths. We will see that this is impossible even if we use all possible quadratic test functions.

For simplicity of exposition, we present this result in the context of an input-queued switch, which is one of the simplest queueing systems that experience multidimensional SSC and it is a special case of the generalized switch, as shown in subsection 7.5.1. The organization of this section is as follows. In subsection 7.8.1 we present the main result, in subsection 7.8.2 we use this result to compute bounds on the first moment of linear combinations of the scaled queue lengths and in subsection 7.9.3 we discuss how to generalize this approach to other queueing systems that experience multidimensional SSC.

7.8.1 System of equations to compute linear combinations of the first moment of scaled queue lengths.

In this section we prove that the drift method with polynomial test functions is not sufficient to compute all the linear combinations of the first moment of the scaled queue lengths in queueing systems that do not satisfy the CRP condition. Specifically, we show that the use of polynomial test functions yields an under-determined system of equations.

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 these cross 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, $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|\mathcal{H}}\|^2$ has been successfully used [34, 14, 15, 22] to obtain the mean sum of the queue lengths, similarly to Theorem 7.5. Typically one uses polynomial test functions of degree $(m+1)$ to get bounds on the expected value of the m^{th} power of the queue lengths. Therefore, in order to obtain bounds on the mean queue lengths, one must use quadratic test functions. In order to get all the 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 the quadratic monomial test functions.

For ease of exposition, in this section we prove our result in the case of $N = 2$ and independent arrivals, i.e., in the case of a 2×2 input-queued switch with independent arrivals. We present generalizations to this result in section 7.9. Specifically, we present the case of a 2×2 input-queued switch with correlated arrivals in subsection 7.9.1, and the case of an $N \times N$ input-queued switch with independent arrivals in subsection 7.9.2. The latter result can be easily generalized to the case of correlated arrivals, but we do not present the result here for ease of exposition.

Theorem 7.24. *Consider a set of 2×2 input-queued switches with independent arrival processes, operating under MaxWeight, indexed by $\epsilon \in (0,1)$ as described in Corollary 7.9. Let $(\Sigma_a^{(\epsilon)})_{i,i} = \sigma_{a_i}^{(\epsilon)}$ and suppose $\lim_{\epsilon \downarrow 0} \sigma_{a_i}^{(\epsilon)} = \sigma_{a_i}$ for all $i \in [4]$. Then, the following system of equations is satisfied*

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1\right] \\
&= \frac{9\sigma_{a_1}^2 + \sigma_{a_2}^2 + \sigma_{a_3}^2 + \sigma_{a_4}^2}{16} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \left(\overline{u}_2 + \overline{u}_3\right)\right] - \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left(\overline{q}_2^+ + \overline{q}_3^+\right) \overline{u}_4\right]
\end{aligned} \tag{7.39}$$

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2\right] \\
&= \frac{\sigma_{a_1}^2 + 9\sigma_{a_2}^2 + \sigma_{a_3}^2 + \sigma_{a_4}^2}{16} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \left(\overline{u}_1 - \overline{u}_3 + \overline{u}_4\right)\right]
\end{aligned} \tag{7.40}$$

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_3\right] \\
&= \frac{\sigma_{a_1}^2 + \sigma_{a_2}^2 + 9\sigma_{a_3}^2 + \sigma_{a_4}^2}{16} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \left(\overline{u}_1 - \overline{u}_2 + \overline{u}_4\right)\right]
\end{aligned} \tag{7.41}$$

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] \\
&= \frac{3\sigma_{a_1}^2 + 3\sigma_{a_2}^2 - \sigma_{a_3}^2 - \sigma_{a_4}^2}{8} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \left(3\overline{u}_2 - \overline{u}_3\right)\right] \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \left(3\overline{u}_1 + \overline{u}_3\right)\right] + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \overline{u}_4\right]
\end{aligned} \tag{7.42}$$

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1 + \overline{q}_3\right] \\
&= \frac{3\sigma_{a_1}^2 - \sigma_{a_2}^2 + 3\sigma_{a_3}^2 - \sigma_{a_4}^2}{8} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \left(-\overline{u}_2 + 3\overline{u}_3\right)\right] \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \overline{u}_4\right] + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \left(3\overline{u}_1 + \overline{u}_2\right)\right]
\end{aligned} \tag{7.43}$$

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2 + \overline{q}_3\right] \\
&= \frac{\sigma_{a_1}^2 - 3\sigma_{a_2}^2 - 3\sigma_{a_3}^2 + \sigma_{a_4}^2}{8} + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \left(\overline{u}_1 + 3\overline{u}_3 + \overline{u}_4\right)\right] \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \left(\overline{u}_1 + 3\overline{u}_2 + \overline{u}_4\right)\right],
\end{aligned} \tag{7.44}$$

where we omitted the dependence on ϵ of the variables for ease of exposition.

The proof of Theorem 7.24 is presented in section 7.10. Observe that in Theorem 7.24 we have system of 6 equations and 11 variables, where the variables are

$$\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_1\right]\ ,\ \lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_2\right]\ ,\ \lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_3\right],$$
$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_1^{+}\overline{u}_2\right]\ ,\ \lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_1^{+}\overline{u}_3\right],$$
$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_2^{+}\overline{u}_1\right]\ ,\ \lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_2^{+}\overline{u}_3\right]\ ,\ \lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_2^{+}\overline{u}_4\right],$$
$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_3^{+}\overline{u}_1\right]\ ,\ \lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_3^{+}\overline{u}_2\right]\ ,\ \lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_3^{+}\overline{u}_4\right].$$

Therefore, it cannot be solved uniquely. However, a specific linear combination of the scaled queue lengths can be obtained, as shown in the next Corollary.

Corollary 7.25. *Consider a set of* 2×2 *input-queued switches as described in Theorem 7.24. Then,*

$$\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_2+\overline{q}_3\right]=\frac{3}{8}\left(\sigma_{a_1}^2+\sigma_{a_2}^2+\sigma_{a_3}^2+\sigma_{a_4}^2\right)$$

Proof of Corollary 7.25. Consider the following linear combination of the equations in Theorem 7.24:

$$\begin{aligned}&Equation\ 7.39+Equation\ 7.40+Equation\ 7.41\\&-\frac{1}{2}Equation\ 7.42-\frac{1}{2}Equation\ 7.43+\frac{1}{2}Equation\ 7.44.\end{aligned}$$

Then, reorganizing terms we obtain the result. □

Corollary 7.25 can be also obtained as a consequence of Corollary 7.9 in the following way.

Alternative proof of Corollary 7.25. From Corollary 7.9 for $N = 2$ we know

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1 + \overline{q}_2 + \overline{q}_3 + \overline{q}_4\right] = \frac{3}{4}\left(\sigma_{a_1}^2 + \sigma_{a_2}^2 + \sigma_{a_3}^2 + \sigma_{a_4}^2\right). \tag{7.45}$$

From SSC as proved in Proposition 7.4 and by definition of the cone $\mathcal{K}$ in Equation 7.22 we also know that for all $i \in [4]$ we have

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_{\|\mathcal{H}i}\right] = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_i\right],$$

where $\overline{q}_{\|\mathcal{H}i}$ is the i^{th} element of $\overline{\boldsymbol{q}}_{\|\mathcal{H}}$. Also, one interpretation of the cone $\mathcal{K}$ presented in [14] is that for each vector in $\mathcal{K}$, all schedules have the same weight in MaxWeight algorithm. This can be easily verified by definition of the cone $\mathcal{K}$ in Equation 7.22. Then,

$$\overline{q}_{\|\mathcal{H}1} + \overline{q}_{\|\mathcal{H}4} = \overline{q}_{\|\mathcal{H}2} + \overline{q}_{\|\mathcal{H}3}. \tag{7.46}$$

Putting everything together we obtain the result in Corollary 7.25. □

A special case where we can solve for each of the expected individual queue lengths is the symmetric case, i.e., when all the arrival processes have the same distribution. We present the result below.

Corollary 7.26. *Consider a set of* 2×2 *input-queued switches as described in Theorem 7.24, where all the arrival processes have the same distribution. Let* $\sigma_a^{(\epsilon)} \triangleq \sigma_{a_i}^{(\epsilon)}$ *for all* $i \in [4]$ *and suppose* $\lim_{\epsilon \downarrow 0} \sigma_a^{(\epsilon)} = \sigma_a$*. Then, for each* $i \in [4]$*, we have*

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_i\right] = \frac{3}{4}\sigma_a^2.$$

Proof of Corollary 7.26. In this case, since the arrivals are symmetric, all the queue lengths have the same expectation. Using this fact in Corollary 7.25 we obtain the result. □

In Theorem 7.24 we prove that setting to zero the drift of all monomials of degree 2

leads to a system of 6 equations in 11 variables. Therefore, the solution is not unique. However, in [14, 15] the authors obtain the limit of specific linear combinations of the scaled queue lengths. These linear combinations can be obtained because some of the variables cancel out, as shown in the first proof of Corollary 7.25. However, 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 additional equations.

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 geometrics, and can be obtained using reversibility arguments. Using the drift approach described above, we 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.

7.8.2 Bounds on linear combinations of the scaled queue lengths in heavy-traffic.

In subsection 7.8.1 we presented a linear system of equations that the vector of queue lengths must satisfy in heavy-traffic. In this section we use this system of equations to obtain bounds on linear combinations of the expected scaled queue lengths in heavy traffic. A similar approach was studied in [111, 112], 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 2×2 input-queued switch.

Theorem 7.27. *Consider the equations*

$$v_1 - \frac{w_1 - w_2 + w_5 + w_8}{2} = \frac{9\sigma_{a_1}^2 + \sigma_{a_2}^2 + \sigma_{a_3}^2 + \sigma_{a_4}^2}{16} \tag{7.47}$$

$$v_2 - \frac{w_3 + w_4 - w_5}{2} = \frac{\sigma_{a_1}^2 + 9\sigma_{a_2}^2 + \sigma_{a_3}^2 + \sigma_{a_4}^2}{16} \tag{7.48}$$

$$v_3 - \frac{w_6 + w_7 - w_8}{2} = \frac{\sigma_{a_1}^2 + \sigma_{a_2}^2 + 9\sigma_{a_3}^2 + \sigma_{a_4}^2}{16} \tag{7.49}$$

$$v_1 + v_2 - \frac{3w_+ w_2 - 3w_3 - w_4 - w_8}{2} = \frac{3\sigma_{a_1}^2 + 3\sigma_{a_2}^2 - \sigma_{a_3}^2 - \sigma_{a_4}^2}{8} \tag{7.50}$$

$$v_1 + v_3 + \frac{w_1 - 3w_2 - w_5 - 3w_6 - w_7}{2} = \frac{3\sigma_{a_1}^2 - \sigma_{a_2}^2 + 3\sigma_{a_3}^2 - \sigma_{a_4}^2}{8} \tag{7.51}$$

$$v_2 + v_3 - \frac{w_3 - 3w_4 - w_5 - w_6 - 3w_7 - w_8}{2} = \frac{\sigma_{a_1}^2 - 3\sigma_{a_2}^2 - 3\sigma_{a_3}^2 + \sigma_{a_4}^2}{8} \tag{7.52}$$

$$-v_1 + v_2 + v_3 \geq 0 \tag{7.53}$$

$$-w_1 + w_7 \geq 0 \tag{7.54}$$

$$-w_2 + w_4 \geq 0 \tag{7.55}$$

and define $\mathcal{P} \triangleq \left\{(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{R}_+^3 \times \mathbb{R}_+^8 : \textit{Equation 7.47-Equation 7.55 are satisfied}\right\}$. *For* $\boldsymbol{\alpha} \in \mathbb{R}^3$*, define*

$$\underline{f}(\boldsymbol{\alpha}) \triangleq \min\left\{\langle \boldsymbol{\alpha}, \boldsymbol{v}\rangle : \ \exists \boldsymbol{w} \textit{ such that } (\boldsymbol{v}, \boldsymbol{w}) \in \mathcal{P}\right\}$$
$$\textit{and} \quad \overline{f}(\boldsymbol{\alpha}) \triangleq \max\left\{\langle \boldsymbol{\alpha}, \boldsymbol{v}\rangle : \ \exists \boldsymbol{w} \textit{ such that } (\boldsymbol{v}, \boldsymbol{w}) \in \mathcal{P}\right\}.$$

Then,

$$\underline{f}(\boldsymbol{\alpha}) \leq \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\langle \boldsymbol{\alpha}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle\right] \leq \overline{f}(\boldsymbol{\alpha}), \tag{7.56}$$

where ϵ *and* $\overline{\boldsymbol{q}}^{(\epsilon)}$ *are defined as in Theorem 7.24. Furthermore, for any* $B \in \mathbb{R}_+$

$$\mathbb{P}\left[\lim_{\epsilon \downarrow 0} \epsilon \langle \boldsymbol{\alpha}, \overline{\boldsymbol{q}}^{(\epsilon)}\rangle \geq B\right] \leq \frac{\overline{f}(\boldsymbol{\alpha})}{B}. \tag{7.57}$$

Proof of Theorem 7.27. For ease of exposition we omit the dependence on ϵ of the variables. Let

$$v_1 = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1\right] \ , \ v_2 = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2\right] \ , \ v_3 = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_3\right],$$

$$w_1 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \overline{u}_2\right] \ , \ w_2 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \overline{u}_3\right],$$

$$w_3 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \overline{u}_1\right] \ , \ w_4 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \overline{u}_3\right] \ ,$$

$$w_5 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \overline{u}_4\right] \ , \ w_6 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \overline{u}_1\right] \ ,$$

$$w_7 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \overline{u}_2\right] \ , \ w_8 = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \overline{u}_4\right].$$

Then, the proof of Equation 7.56 follows from Theorem 7.24 because the set $\mathcal{P}$ represents the system of equations presented there together with non-negativity constraints for all the variables. In particular, Equation 7.53-Equation 7.55 represent non-negativity constraints associated to $\overline{q}_4$. These must be considered because, even though $\overline{q}_4$ does not appear in the system of equations explicitly, there are underlying constraints of the system related to $\overline{q}_4$ that affect its performance. Specifically, using Equation 7.46 and the definition of the variables above, we obtain that the inequalities

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_4\right] \geq 0 \ , \quad \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_4^+ \overline{u}_i\right] \geq 0 \ \forall i \in \{1, 2, 3\}$$

can be rewritten as Equation 7.53, Equation 7.54, Equation 7.55 and $w_3 + w_6 \geq 0$ but the last inequality is implied by $w_3 \geq 0$ and $w_6 \geq 0$, so we do not write it in the definition of $\mathcal{P}$.

Also, from Markov's inequality we know

$$\mathbb{P}\left[\lim_{\epsilon \downarrow 0} \epsilon \langle \boldsymbol{\alpha}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle \geq B\right] \leq \frac{\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\langle \boldsymbol{\alpha}, \overline{\boldsymbol{q}}^{(\epsilon)} \rangle\right]}{B} \leq \frac{\overline{f}(\boldsymbol{\alpha})}{B},$$

where the last inequality holds by Equation 7.56. □

Table 7.1: Numerical results for LP with objective function $\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2^{(\epsilon)} + \overline{q}_3^{(\epsilon)}\right]$.

ϵ	Solution to LP	Mean from simulation	Error
0.01	0.375	0.378	0.87%
0.05	0.374	0.351	6.69%
0.10	0.371	0.336	10.38%

Theorem 7.27 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.

In the rest of this section we present numerical results to compare the bounds that we obtain from the linear program presented in Theorem 7.27 with the mean values that we obtain from simulation. We test four different objective functions, viz. $\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_i\right]$ for $i \in \{1, 2, 3\}$ and $\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2 + \overline{q}_3\right]$. We use the last function because in this case the system of equations has a unique solution, as shown in Corollary 7.25.

For simplicity, we assume that the arrivals to each queue are Bernoulli processes with mean $\lambda_i^{(\epsilon)} = \frac{1-\epsilon}{2}$ for all $i \in [4]$. We take $\epsilon \in \{0.01, 0.05, 0.1\}$ to evaluate the performance under different traffic intensities.

To allow the system to reach steady-state, we ran the simulation for 10^9 time slots when $\epsilon \in \{0.05, 0.1\}$ and for 10^{10} time slots in the case of $\epsilon = 0.01$. The reason is that, for smaller ϵ, the system takes more time to reach steady state. In both cases we compute the mean value of the variables considering the last 2×10^6 time slots[1]. We present our results in Table 7.1 and Table 7.2. We ran three replicas of each experiment, and we obtained similar results. The results we present in Table 7.1 and Table 7.2 were computed as an average of the three replicas.

In Table 7.1 we present the right-hand side of the expression proved in Corollary 7.25, the mean value of $\epsilon \left(\overline{q}_2^{(\epsilon)} + \overline{q}_3^{(\epsilon)}\right)$ obtained from the simulation, and the percentage error of the solution of the system of equations with respect to the simulation.

Observe that, as ϵ decreases, the solution to the LP becomes a better approximation for

[1]The code is publicly available here: https://github.com/dhurtadolange/2x2-switch-simulation

Table 7.2: Numerical results for individual queue lengths.

ϵ	Min	Max	Average min and max	Simulation			
				Mean $\epsilon\overline{q}_1$	Mean $\epsilon\overline{q}_2$	Mean $\epsilon\overline{q}_3$	Mean $\epsilon\overline{q}_4$
0.01	0.062	0.312	0.187	0.187	0.187	0.192	0.191
0.05	0.062	0.312	0.187	0.174	0.176	0.175	0.176
0.10	0.062	0.309	0.186	0.168	0.168	0.168	0.169

the simulated result. In fact, when $\epsilon = 0.01$, the error is below 1%. Even in the case of $\epsilon = 0.1$, which is not considered heavy-traffic, the error is around 10%.

In Table 7.2 we compute a lower and an upper bound to the mean individual queue lengths, and we compare these results with the mean value of $\epsilon\overline{q}_1$, $\epsilon\overline{q}_2$, $\epsilon\overline{q}_3$ and $\epsilon\overline{q}_4$ obtained from simulation. The reason to present only one optimal value for all the queue lengths is that solving the linear program presented in Theorem 7.27 with objective function $\epsilon\mathbb{E}\left[\overline{q}_i^{(\epsilon)}\right]$ gives the same optimal value for all $i = 1, 2, 3$, because of the symmetric arrival pattern. We additionally present the average between the minimum and maximum value of the individual queue lengths.

Observe that for all the cases presented in Table 7.2, the mean obtained by simulation is between the lower and upper bound obtained solving the LP. The bounds are not necessarily tight, but the average of both gives a good approximation of the mean individual queue lengths. Additionally, the LP presented in Theorem 7.5 is simple and, hence, it can be solved in fractions of a second, as opposed to the simulation that may take hours.

7.9 Generalizations of Theorem 7.24

In this section we present generalizations of Theorem 7.24. We first present the result for a 2×2 switch with correlated arrivals, then for an $N \times N$ switch with independent arrivals, and at the end, we discuss the number of variables and equations that one would obtain in a generalized switch with n queue and SSC into a d-dimensional subspace, with $d > 1$.

7.9.1 System of equations for the 2×2 input-queued switch with correlated arrivals.

In this section we provide a generalization of Theorem 7.24 to the case of a switch with correlated arrivals. We omit the proof, since it is similar to the proof of Theorem 7.24.

Theorem 7.28. *Consider a set of* 2×2 *input-queued switches operating under MaxWeight, indexed by* $\epsilon \in (0,1)$ *as described in Corollary 7.8. Suppose* $\Sigma^{(\epsilon)}$ *is the covariance matrix of the arrival processes, and* $\lim_{\epsilon \downarrow 0} \Sigma^{(\epsilon)} = \Sigma$ *component-wise. Then, the following system of equations is satisfied*

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1\right] \\
&= \frac{9\Sigma_{1,1} + 6\Sigma_{1,2} + 6\Sigma_{1,3} - 6\Sigma_{1,4} + \Sigma_{2,2} + 2\Sigma_{2,3} - 2\Sigma_{2,4} + \Sigma_{3,3} - 2\Sigma_{3,4} + \Sigma_{4,4}}{16} \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \left(\overline{u}_2 + \overline{u}_3\right)\right] - \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left(\overline{q}_2^+ + \overline{q}_3^+\right) \overline{u}_4\right] \\
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_2\right] \\
&= \frac{\Sigma_{1,1} + 6\Sigma_{1,2} - 2\Sigma_{1,3} + 2\Sigma_{1,4} + 9\Sigma_{2,2} - 6\Sigma_{2,3} - 6\Sigma_{2,4} + \Sigma_{3,3} - 2\Sigma_{3,4} + \Sigma_{4,4}}{16} \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \left(\overline{u}_1 - \overline{u}_3 + \overline{u}_4\right)\right] \\
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_3\right] \\
&= \frac{\Sigma_{1,1} - 2\Sigma_{1,2} + 6\Sigma_{1,3} + 2\Sigma_{1,4} + \Sigma_{2,2} - 6\Sigma_{2,3} - 2\Sigma_{2,4} + 9\Sigma_{3,3} + 6\Sigma_{3,4} + \Sigma_{4,4}}{16} \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \left(\overline{u}_1 - \overline{u}_2 + \overline{u}_4\right)\right] \\
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_1 + \overline{q}_2\right] \\
&= \frac{3\Sigma_{1,1} + 18\Sigma_{1,2} - 6\Sigma_{1,3} + 6\Sigma_{1,4} + 3\Sigma_{2,2} - 2\Sigma_{2,3} + 2\Sigma_{2,4} - \Sigma_{3,3} + 2\Sigma_{3,4} - \Sigma_{4,4}}{8} \\
&\quad + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_1^+ \left(3\overline{u}_2 - \overline{u}_3\right)\right] + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_2^+ \left(3\overline{u}_1 + \overline{u}_3\right)\right] + \frac{1}{2} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{q}_3^+ \overline{u}_4\right]
\end{aligned}$$

$$\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_1+\overline{q}_3\right]$$
$$=\frac{3\Sigma_{1,1}-6\Sigma_{1,2}+18\Sigma_{1,3}+6\Sigma_{1,4}-\Sigma_{2,2}+6\Sigma_{2,3}+2\Sigma_{2,4}+3\Sigma_{3,3}+6\Sigma_{3,4}-\Sigma_{4,4}}{8}$$
$$+\frac{1}{2}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_1^+\left(-\overline{u}_2+3\overline{u}_3\right)\right]+\frac{1}{2}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_2^+\overline{u}_4\right]+\frac{1}{2}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_3^+\left(3\overline{u}_1+\overline{u}_2\right)\right]$$

$$\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_2+\overline{q}_3\right]$$
$$=\frac{\Sigma_{1,1}-2\Sigma_{1,2}+6\Sigma_{1,3}+2\Sigma_{1,4}-3\Sigma_{2,2}+18\Sigma_{2,3}+6\Sigma_{2,4}-3\Sigma_{3,3}-2\Sigma_{3,4}+\Sigma_{4,4}}{8}$$
$$+\frac{1}{2}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_2^+\left(\overline{u}_1+3\overline{u}_3+\overline{u}_4\right)\right]+\frac{1}{2}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\overline{q}_3^+\left(\overline{u}_1+3\overline{u}_2+\overline{u}_4\right)\right],$$

where we omitted the dependence on ϵ of the variables for ease of exposition.

7.9.2 System of equations for the $N \times N$ input-queued switch.

For ease of exposition, in this section we use the matrix-shape interpretation of the switch and we assume the arrivals to different input ports are independent of each other. With a slight abuse of notation, we adhere to the notation introduced in section 7.8 for the variables, and we use two subscripts, one for the input port and one for the output port. For example, $q_{i,j}(k)$ is the number of packets in line at input port i and output port j, for $i,j\in[N]$. Before presenting the theorem we introduce the following notation. For $i,j\in[N]$, define

$$[N]_i \triangleq [N]\setminus\{i\} \quad \text{and} \quad [N]_{i,j}\triangleq [N]\setminus\{i,j\}$$

Theorem 7.29. *Consider a set of input-queued switches operating under MaxWeight, indexed by $\epsilon\in(0,1)$ as described in Corollary 7.9. Further, for all $i,j\in[N]$ let $\sigma_{i,j}^{(\epsilon)}\triangleq Var\left[\overline{a}_{i,j}^{(\epsilon)}\right]$ and assume $\sigma_{i,j}^2=\lim_{\epsilon\downarrow 0}\left(\sigma_{i,j}^{(\epsilon)}\right)^2$. Then, the following system of equations is satisfied, where we omit the dependence on ϵ of the variables by ease of exposition.*

$$\langle\overline{\boldsymbol{q}}_{\|},\boldsymbol{p}\rangle=\sum_{i=1}^{n}\overline{q}_{\|i,i}\qquad\forall\boldsymbol{p}\in\mathcal{S}. \tag{7.58}$$

$$
\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_{1,j}\right] \\
=&\frac{1}{2N^3}\left((2N-1)^2\sigma_{1,j}^2+(N-1)^2\left(\sum_{i'\in[N]_1}\sigma_{i',j}^2+\sum_{j'\in[N]_j}\sigma_{1,j'}^2\right)+\sum_{i'\in[N]_1}\sum_{j'\in[N]_j}\sigma_{i',j'}^2\right) \\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(N-1)\left(\sum_{i\in[N]_1}\overline{q}_{1,j}^+\overline{u}_{i',j}+\sum_{j'\in[N]_j}\overline{q}_{1,j}^+\overline{u}_{1,j'}\right)-\sum_{i'\in[N]_1}\sum_{j'\in[N]_j}\overline{q}_{1,j}^+\overline{u}_{i',j'}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall j\in[N]
\end{aligned}
\tag{7.59}
$$

$$
\begin{aligned}
&\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\overline{q}_{i,1}\right] \\
=&\frac{1}{2N^3}\left((2N-1)^2\sigma_{i,1}^2+(N-1)^2\left(\sum_{i'\in[N]_i}\sigma_{i',1}^2+\sum_{j\in[N]_1}\sigma_{i,j'}^2\right)+\sum_{i\in[N]_i}\sum_{j'\in[N]_1}\sigma_{i',j'}^2\right) \\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(N-1)\left(\sum_{i'\in[N]_i}\overline{q}_{i,1}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{i,j'}\right)-\sum_{i'\in[N]_i}\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{i',j'}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall i\in[N]_1
\end{aligned}
\tag{7.60}
$$

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_{1,1}+\overline{q}_{i,1}\right]\\
=&\frac{(N-1)}{N^3}\left((2N-1)\left(\sigma_{1,1}^2+\sigma_{i,1}^2\right)+(N-1)\sum_{i\in[N]_{1,i}}\sigma_{i',1}^2-\sum_{j'\in[N]_1}\sigma_{1,j'}^2-\sum_{j'\in[N]_1}\sigma_{i,j'}^2\right)\\
&+\frac{1}{N^3}\sum_{i'\in[N]_{1,i}}\sum_{j\in[N]_1}\sigma_{i',j'}^2\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{1,1}^+\overline{u}_{i,1}+(N-1)\left(\sum_{i'\in[N]_{1,i}}\overline{q}_{1,1}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{1,1}^+\overline{u}_{i,j'}\right)\right]\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{i,1}^+\overline{u}_{1,1}+(N-1)\left(\sum_{i'\in[N]_{1,i}}\overline{q}_{i,1}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{1,j'}\right)\right]\\
&-\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\sum_{i'\in[N]_i}\sum_{j'\in[N]_1}\overline{q}_{1,1}^+\overline{u}_{i',j'}+\sum_{i'\in[N]_1}\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{i',j'}\right]
\end{aligned}
$$

$$\forall i\in[N]_1 \tag{7.61}$$

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_{1,j}+\overline{q}_{1,m}\right]\\
=&\frac{(N-1)}{N^3}\left((2N-1)\left(\sigma_{1,j}^2+\sigma_{1,m}^2\right)+(N-1)\sum_{j'\in[N]_{j,m}}\sigma_{1,j'}^2\right)\\
&-\frac{(N-1)}{N^3}\left(\sum_{i'\in[N]_1}\sigma_{i',j}^2-\sum_{i'\in[N]_m}\sigma_{i',m}^2\right)+\frac{1}{N^3}\sum_{i'\in[N]_1}\sum_{j'\in[N]_{j,m}}\sigma_{i',j'}^2\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{1,j}^+\overline{u}_{1,m}+(N-1)\left(\sum_{i'\in[N]_1}\overline{q}_{1,j}^+\overline{u}_{i',m}+\sum_{j'\in[N]_{j,m}}\overline{q}_{1,j}^+\overline{u}_{1,j'}\right)\right]\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{1,m}^+\overline{u}_{1,j}+(N-1)\left(\sum_{i'\in[N]_1}\overline{q}_{1,m}^+\overline{u}_{i',j}+\sum_{j'\in[N]_{j,m}}\overline{q}_{1,m}^+\overline{u}_{1,j'}\right)\right]\\
&-\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\sum_{i'\in[N]_1}\sum_{j'\in[N]_m}\overline{q}_{1,j}^+\overline{u}_{i',j'}+\sum_{i'\in[N]_1}\sum_{j'\in[N]_j}\overline{q}_{1,m}^+\overline{u}_{i',j'}\right]
\end{aligned}
$$

$$\forall(j,m)\in\mathcal{A}_1 \tag{7.62}$$

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_{i,1}+\overline{q}_{l,1}\right]\\
=&\frac{(N-1)}{N^3}\left((2N-1)\left(\sigma_{i,1}^2+\sigma_{l,1}^2\right)+(N-1)\sum_{i'\in[N]_{i,l}}\sigma_{i',1}^2\right)\\
&-\frac{(N-1)}{N^3}\left(\sum_{j'\in[N]_1}\sigma_{i,j'}^2-\sum_{j'\in[N]_1}\sigma_{l,j'}^2\right)+\frac{1}{N^3}\sum_{i'\in[N]_{i,l}}\sum_{j'\in[N]_1}\sigma_{i',j'}^2\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{i,1}^+\overline{u}_{l,1}+(N-1)\left(\sum_{i'\in[N]_{i,l}}\overline{q}_{i,1}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{l,j'}\right)\right] \qquad (7.63)\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{l,1}^+\overline{u}_{i,1}+(N-1)\left(\sum_{i'\in[N]_{i,l}}\overline{q}_{l,1}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{l,1}^+\overline{u}_{i,j'}\right)\right]\\
&-\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\sum_{i'\in[N]_l}\sum_{j'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{i',j'}+\sum_{i'\in[N]_i}\sum_{j'\in[N]_1}\overline{q}_{l,1}^+\overline{u}_{i',j'}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall(i,l)\in\mathcal{A}_2
\end{aligned}
$$

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon\mathbb{E}\left[\overline{q}_{1,j}+\overline{q}_{i,1}\right]\\
=&\frac{1}{N^3}\left(-(2N-1)\left(\sigma_{1,j}^2+\sigma_{i,1}^2\right)+(N-1)^2\sigma_{1,1}^2+\sum_{i'\in[N]_{1,i}}\sum_{j'\in[N]_{1,j}}\sigma_{i',j'}^2\right)\\
&-\frac{(N-1)}{N^3}\left(\sum_{i\in[N]_1}\sigma_{i',j}^2+\sum_{j'\in[N]_{1,j}}\sigma_{1,j'}^2+\sum_{i'\in[N]_{1,i}}\sigma_{i',1}^2+\sum_{j'\in[N]_{1,j}}\sigma_{1,j'}^2\right)\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{1,j}^+\overline{u}_{i,1}+(N-1)\left(\sum_{i'\in[N]_i}\overline{q}_{1,j}^+\overline{u}_{i',1}+\sum_{j'\in[N]_1}\overline{q}_{1,j}^+\overline{u}_{i,j'}\right)\right] \qquad (7.64)\\
&+\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[(2N-1)\overline{q}_{i,1}^+\overline{u}_{1,j}+(N-1)\left(\sum_{i'\in[N]_1}\overline{q}_{i,1}^+\overline{u}_{i',j}+\sum_{j'\in[N]_j}\overline{q}_{i,1}^+\overline{u}_{1,j'}\right)\right]\\
&-\frac{1}{N}\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\sum_{(i',j')\in[N]_i\times[N]_1\setminus\{(1,j)\}}\overline{q}_{1,j}^+\overline{u}_{i',j'}+\sum_{(i',j')\in[N]_1\times[N]_j\setminus\{(i,1)\}}\overline{q}_{i,1}^+\overline{u}_{i',j'}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall i,j\in[N]_1
\end{aligned}
$$

where $\mathcal{P}$ is the set of $N \times N$ permutation matrices and

$$\mathcal{A}_1 = \{(x,y) \in [N] \times [N] : \ y \geq x+1\}$$
$$\mathcal{A}_2 = \{(x,y) \in [N] \times [N] : \ y \geq x+1 \ , \ 2 \leq x \leq N-1\} .$$

Equation 7.58 is one interpretation of SSC, which says that all the schedules have the same weight in the cone $\mathcal{K}$. Observe that in Theorem 7.24 we did not have an equation of the form of Equation 7.58. However, we used this condition in the proof to obtain a system of equations (see Claim 7.31). In this case, we decided to write it as an equation to make explicit the use of SSC.

7.9.3 Generalization to other queueing systems and higher moments.

In this section we focused on an input-queued switch 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, which is what we discuss in this subsection. Specifically, we focus on a generalized switch with n queues, where SSC occurs into a d-dimensional subspace.

In [34], the authors show how to compute the moments of $\|\boldsymbol{q}_{\|\mathcal{H}}\|$ using the drift method in queueing systems that satisfy the CRP condition. In this case, setting to zero the drift of $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|\mathcal{H}}\|^{m+1}$ 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 $V(\boldsymbol{q})$, terms of the form $q^+_{\|\mathcal{H}i} u_{\|\mathcal{H}i}$ arise and, since $\boldsymbol{q}^+_{\|\mathcal{H}}$ and $\boldsymbol{u}_{\|\mathcal{H}}$ 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 $\boldsymbol{q}$ lives in a d-dimensional subspace, where $d > 1$. In this case, for each i, $q^+_{\|\mathcal{H}i} u_{\|\mathcal{H}i}$ cannot be approximated by $q_i^+ u_i$ because of the following reason. In heavy traffic we only have the approximation (with some abuse of notation) $q^+_{\|\mathcal{H}i} u_{\|\mathcal{H}i} \approx q_i^+(u_{k_1} + u_{k_2} + \cdots + u_{k_d})$, where $k_1, \ldots, k_d$ represent

the d dimensions that characterize SSC. In other words, cross terms arise exactly as the 'qu' terms in Theorem 7.24, Theorem 7.28 and Theorem 7.29 for the input-queued switch. 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 the monomials of degree $m+1$. Since SSC occurs into a d-dimensional subspace, we need to consider all the 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 $k \leq m$. 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 $k < m+1$. 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 $\boldsymbol{q}_{\|\mathcal{H}}$ and, on the other hand, we have the heavy-traffic limit of the expected value of the product between the elements of $\boldsymbol{q}_{\|\mathcal{H}}$ 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 $\binom{m+d-1}{d-1}$ 'q' variables. The 'qu' variables 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 $\boldsymbol{u}_{\|\mathcal{H}}$ 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 this section to compute the m^{th} mo-

ment 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 Theorem 7.24 arise in any queueing system with multidimensional SSC.

7.10 Proof of Theorem 7.24.

For ease of exposition, in this proof we use subscript $\|$ instead of $\|\ \mathcal{H}$, since we only use projection on the subspace $\mathcal{H}$ and not on the cone $\mathcal{K}$.

Proof of Theorem 7.24. We know that SSC occurs into a subspace of dimension $2N-1=3$. Therefore, 3 variables are necessary to compute the most general quadratic polynomial. In fact, we know $\overline{q}_{\|4}=\overline{q}_{\|2}+\overline{q}_{\|3}-\overline{q}_{\|1}$. Then, we only need to consider the variables $\overline{q}_{\|1}$, $\overline{q}_{\|2}$ and $\overline{q}_{\|3}$. The most general quadratic polynomial with these variables is

$$V(\boldsymbol{q}) =\alpha_1\overline{q}^2_{\|1}+\alpha_2\overline{q}^2_{\|2}+\alpha_3\overline{q}^2_{\|3}+\alpha_4\overline{q}_{\|1}\overline{q}_{\|2}+\alpha_5\overline{q}_{\|1}\overline{q}_{\|3}+\alpha_6\overline{q}_{\|2}\overline{q}_{\|3},$$

where $\alpha_i\in\mathbb{R}$ for all $i\in[6]$.

Setting to zero the drift of $V(\boldsymbol{q})$ is equivalent to setting to zero the drift of each monomial separately. Then, we set to zero the drift of the following 6 test functions:

$$\begin{aligned}&V_1(\boldsymbol{q})=\overline{q}^2_{\|1},\ V_2(\boldsymbol{q})=\overline{q}^2_{\|2},\ V_3(\boldsymbol{q})=\overline{q}^2_{\|3},\\ &V_4(\boldsymbol{q})=\overline{q}_{\|1}\overline{q}_{\|2},\ V_5(\boldsymbol{q})=\overline{q}_{\|1}\overline{q}_{\|3}\text{ and }V_6(\boldsymbol{q})=\overline{q}_{\|2}\overline{q}_{\|3}.\end{aligned}$$

Before setting to zero the drift of $V_i(\boldsymbol{q})$ for $i\in[6]$ observe that, by definition of the cone $\mathcal{K}$ in Equation 7.22 we have for any vector $\boldsymbol{y}\in\mathbb{R}^4$

$$y_{\|1}=\frac{y_1+y_2}{2}+\frac{y_1+y_3}{2}-\frac{y_1+y_2+y_3+y_4}{4}=\frac{3y_1+y_2+y_3-y_4}{4},\tag{7.65}$$

$$y_{\|2} = \frac{y_1 + y_2}{2} + \frac{y_2 + y_4}{2} - \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{y_1 + 3y_2 - y_3 + y_4}{4}, \tag{7.66}$$

$$y_{\|3} = \frac{y_3 + y_4}{2} + \frac{y_1 + y_3}{2} - \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{y_1 - y_2 + 3y_3 + y_4}{4}. \tag{7.67}$$

Then, since the switch is completely saturated, we have

$$\mathbb{E}\left[\overline{a}_{\|i}\right] = \frac{1-\epsilon}{2} + \frac{1-\epsilon}{2} - \frac{2(1-\epsilon)}{4} = \frac{1-\epsilon}{2} \qquad \forall i \in [4] \tag{7.68}$$

and since $\overline{\boldsymbol{s}}$ is a maximal schedule we have

$$\overline{s}_{\|i} = \frac{1}{2} + \frac{1}{2} - \frac{2}{4} = \frac{1}{2} \qquad \forall i \in [4]. \tag{7.69}$$

We first set to zero the drift of $V_1(\boldsymbol{q})$. We obtain

$$\begin{aligned}
0 &= \mathbb{E}\left[\left(\overline{q}^+_{\|1}\right)^2 - \overline{q}^2_{\|1}\right] \\
&= \mathbb{E}\left[\left(\overline{q}^+_{\|1} - \overline{u}_{\|1} + \overline{u}_{\|1}\right)^2 - \overline{q}^2_{\|1}\right] \\
&= \mathbb{E}\left[\left(\overline{q}^+_{\|1} - \overline{u}_{\|1}\right)^2 + \overline{u}^2_{\|1} + 2\left(\overline{q}^+_{\|1} - \overline{u}_{\|1}\right)\overline{u}_{\|1} - \overline{q}^2_{\|1}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left(\overline{q}_{\|1} + \overline{a}_{\|1} - \overline{s}_{\|1}\right)^2 - \overline{u}^2_{\|1} + 2\overline{q}^+_{\|1}\overline{u}_{\|1} - \overline{q}^2_{\|1}\right] \\
&= \mathbb{E}\left[\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)^2 + 2\overline{q}_{\|1}\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right) - \overline{u}^2_{\|1} + 2\overline{q}^+_{\|1}\overline{u}_{\|1}\right],
\end{aligned} \tag{7.70}$$

where (a) holds by Equation 1.2 and reorganizing the terms. We compute each term separately. For the first term we have

$$\begin{aligned}
\mathbb{E}\left[\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)^2\right] &\overset{(a)}{=} \mathbb{E}\left[\left(\overline{a}_{\|1} - \frac{1}{2}\right)^2\right] \\
&\overset{(b)}{=} \text{Var}\left[\overline{a}_{\|1}\right] + \left(\mathbb{E}\left[\overline{a}_{\|1}\right]\right)^2 + \frac{1}{4} - \mathbb{E}\left[\overline{a}_{\|1}\right] \\
&= \text{Var}\left[\overline{a}_{\|1}\right] + \left(\mathbb{E}\left[\overline{a}_{\|1}\right] - \frac{1}{2}\right)^2
\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{=} \text{Var}\left[\frac{3\overline{a}_1 + \overline{a}_2 + \overline{a}_3 - \overline{a}_4}{4}\right] + \frac{\epsilon^2}{4} \\
&\overset{(d)}{=} \frac{9\left(\sigma_{a_1}^{(\epsilon)}\right)^2 + \left(\sigma_{a_2}^{(\epsilon)}\right)^2 + \left(\sigma_{a_3}^{(\epsilon)}\right)^2 + \left(\sigma_{a_4}^{(\epsilon)}\right)^2}{16} + \frac{\epsilon^2}{4},
\end{aligned} \tag{7.71}$$

where (a) holds by Equation 7.69; (b) holds by definition of variance and reorganizing terms; (c) holds by definition of $\overline{a}_{\|1}$ as in Equation 7.65, and by Equation 7.68; and (d) holds because the arrival processes to different queues are independent. For the second term we obtain

$$\begin{aligned}
2\mathbb{E}\left[\overline{q}_{\|1}\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)\right] &\overset{(a)}{=} 2\mathbb{E}\left[\overline{q}_{\|1}\left(\overline{a}_{\|1} - \frac{1}{2}\right)\right] \\
&\overset{(b)}{=} 2\mathbb{E}\left[\overline{q}_{\|1}\right]\left(\mathbb{E}\left[\overline{a}_{\|1}\right] - \frac{1}{2}\right) \\
&\overset{(c)}{=} -\epsilon\mathbb{E}\left[\overline{q}_{\|1}\right],
\end{aligned} \tag{7.72}$$

where (a) holds by Equation 7.69; (b) holds because the arrival processes are independent of the queue lengths; and (c) holds by Equation 7.68. For the third term, observe

$$0 \leq \mathbb{E}\left[\overline{u}_{\|1}^2\right] \leq \mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right].$$

From the proof of Theorem 7.5, we know $\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]$ is $O(\epsilon)$ (see Equation 7.34). Therefore,

$$\mathbb{E}\left[\overline{u}_{\|1}^2\right] \text{ is } O(\epsilon). \tag{7.73}$$

Now we compute the last term. By definition of $\overline{\boldsymbol{q}}_{\|}$ and $\overline{\boldsymbol{q}}_{\perp}$ we have

$$2\mathbb{E}\left[\overline{q}_{\|1}^{+}\overline{u}_{\|1}\right] = 2\mathbb{E}\left[\overline{q}_1^{+}\overline{u}_{\|1}\right] - 2\mathbb{E}\left[\overline{q}_{\perp 1}^{+}\overline{u}_{\|1}\right].$$

Claim 7.30. *Consider the queueing system described in Theorem 7.24. Then,*

$$\mathbb{E}\left[\overline{q}^+_{\perp 1}\overline{u}_{\parallel 1}\right] \text{ is } O(\sqrt{\epsilon}).$$

The proof of Claim 7.30 is presented in subsection 7.11.1. Then,

$$\begin{aligned}
2\mathbb{E}\left[\overline{q}^+_{\parallel 1}\overline{u}_{\parallel 1}\right] =& 2\mathbb{E}\left[\overline{q}^+_1\overline{u}_{\parallel 1}\right] + O(\sqrt{\epsilon}) \\
\overset{(a)}{=}& \frac{1}{2}\mathbb{E}\left[\overline{q}^+_1\left(3\overline{u}_1 + \overline{u}_2 + \overline{u}_3 - \overline{u}_4\right)\right] + O(\sqrt{\epsilon}) \\
\overset{(b)}{=}& \frac{1}{2}\mathbb{E}\left[\overline{q}^+_1\left(\overline{u}_2 + \overline{u}_3 - \overline{u}_4\right)\right] + O(\sqrt{\epsilon})
\end{aligned}$$

where (a) holds by Equation 7.65; and (b) holds by Equation 1.3.

Claim 7.31. *Consider the queueing system described in Theorem 7.24. Then,*

$$\mathbb{E}\left[\overline{q}^+_1\overline{u}_4\right] = \mathbb{E}\left[\overline{q}^+_2\overline{u}_4\right] + \mathbb{E}\left[\overline{q}^+_3\overline{u}_4\right] + O(\sqrt{\epsilon}).$$

The proof of Claim 7.31 is presented in subsection 7.11.2. Therefore, we obtain

$$2\mathbb{E}\left[\overline{q}^+_{\parallel 1}\overline{u}_{\parallel 1}\right] = \frac{1}{2}\mathbb{E}\left[\overline{q}^+_1\left(\overline{u}_2 + \overline{u}_3\right)\right] - \frac{1}{2}\mathbb{E}\left[\overline{q}^+_2\overline{u}_4\right] - \frac{1}{2}\mathbb{E}\left[\overline{q}^+_3\overline{u}_4\right] + O(\sqrt{\epsilon}) \tag{7.74}$$

Using Equation 7.71, Equation 7.72, Equation 7.73 and Equation 7.74 in Equation 7.70, and reorganizing the terms we obtain

$$\begin{aligned}
&\epsilon\mathbb{E}\left[\overline{q}_{\parallel 1}\right] \\
=&\frac{9\left(\sigma^{(\epsilon)}_{a_1}\right)^2 + \left(\sigma^{(\epsilon)}_{a_2}\right)^2 + \left(\sigma^{(\epsilon)}_{a_3}\right)^2 + \left(\sigma^{(\epsilon)}_{a_4}\right)^2}{16} + \frac{1}{2}\mathbb{E}\left[\overline{q}^+_1\left(\overline{u}_2 + \overline{u}_3\right)\right] - \frac{1}{2}\mathbb{E}\left[\overline{q}^+_2\overline{u}_4\right] \\
&- \frac{1}{2}\mathbb{E}\left[\overline{q}^+_3\overline{u}_4\right] + \frac{\epsilon^2}{4} + O(\sqrt{\epsilon}).
\end{aligned}$$

Taking the limit as $\epsilon \downarrow 0$ on both sides we obtain Equation 7.39. The proof of Equa-

tion 7.40 and of Equation 7.41 hold similarly, after setting to zero the drift of $V_2(\boldsymbol{q})$ and $V_3(\boldsymbol{q})$ respectively. We omit the details for brevity.

To obtain Equation 7.42 we set to zero the drift of $V_4(\boldsymbol{q})$. After similar manipulation as above, we obtain

$$\begin{aligned} 0 &= \mathbb{E}\left[\overline{q}^+_{\|1}\overline{q}^+_{\|2} - \overline{q}_{\|1}\overline{q}_{\|2}\right] \\ &= \mathbb{E}\left[\overline{q}_{\|1}\left(\overline{a}_{\|2} - \overline{s}_{\|2}\right)\right] + \mathbb{E}\left[\overline{q}_{\|2}\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)\right] + \mathbb{E}\left[\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)\left(\overline{a}_{\|2} - \overline{s}_{\|2}\right)\right] \\ &\quad + \mathbb{E}\left[\overline{q}^+_{\|1}\overline{u}_{\|2}\right] + \mathbb{E}\left[\overline{q}^+_{\|2}\overline{u}_{\|1}\right] - \mathbb{E}\left[\overline{u}_{\|1}\overline{u}_{\|2}\right]. \end{aligned} \tag{7.75}$$

We compute term by term. For the first term we have

$$\mathbb{E}\left[\overline{q}_{\|1}\left(\overline{a}_{\|2} - \overline{s}_{\|2}\right)\right] = -\frac{\epsilon}{2}\mathbb{E}\left[\overline{q}_{\|1}\right], \tag{7.76}$$

where we used that $\overline{s}_{\|2} = \frac{1}{2}$ and independence of the arrivals and queue lengths processes. Similarly, for the second term we obtain

$$\mathbb{E}\left[\overline{q}_{\|2}\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)\right] = -\frac{\epsilon}{2}\mathbb{E}\left[\overline{q}_{\|2}\right]. \tag{7.77}$$

For the third term we have

$$\begin{aligned} &\mathbb{E}\left[\left(\overline{a}_{\|1} - \overline{s}_{\|1}\right)\left(\overline{a}_{\|2} - \overline{s}_{\|2}\right)\right] && (7.78) \\ &\overset{(a)}{=} \mathbb{E}\left[\left(\overline{a}_{\|1} - \frac{1}{2}\right)\left(\overline{a}_{\|2} - \frac{1}{2}\right)\right] \\ &\overset{(b)}{=} \mathrm{Cov}\left[\overline{a}_{\|1}, \overline{a}_{\|2}\right] + \mathbb{E}\left[\overline{a}_{\|1}\right]\mathbb{E}\left[\overline{a}_{\|2}\right] - \frac{1}{2}\mathbb{E}\left[\overline{a}_{\|1}\right] - \frac{1}{2}\mathbb{E}\left[\overline{a}_{\|2}\right] + \frac{1}{4} \\ &\overset{(c)}{=} \mathrm{Cov}\left[\frac{3\overline{a}_1 + \overline{a}_2 + \overline{a}_3 - \overline{a}_4}{4}, \frac{\overline{a}_1 + 3\overline{a}_2 - \overline{a}_3 + \overline{a}_4}{4}\right] + \frac{\epsilon^2}{4} \\ &\overset{(d)}{=} \frac{3\left(\sigma^{(\epsilon)}_{a1}\right)^2 + 3\left(\sigma^{(\epsilon)}_{a2}\right)^2 - \left(\sigma^{(\epsilon)}_{a3}\right)^2 - \left(\sigma^{(\epsilon)}_{a4}\right)^2}{16} + \frac{\epsilon^2}{4} && (7.79) \end{aligned}$$

where (a) holds by Equation 7.69; (b) holds by definition of covariance and reorganizing

terms; (c) holds by Equation 7.65, Equation 7.66 and Equation 7.68; and (d) holds because the arrival processes to different queues are independent. For the fourth term we have

$$\begin{aligned}
\mathbb{E}\left[\overline{q}_{\|1}^{+}\overline{u}_{\|2}\right] &\overset{(a)}{=} \mathbb{E}\left[\overline{q}_{1}^{+}\overline{u}_{\|2}\right] - \mathbb{E}\left[\overline{q}_{\perp 1}^{+}\overline{u}_{\|2}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\overline{q}_{1}^{+}\overline{u}_{\|2}\right] + O(\sqrt{\epsilon}) \\
&\overset{(c)}{=} \frac{1}{4}\mathbb{E}\left[\overline{q}_{1}^{+}\left(\overline{u}_1 + 3\overline{u}_2 - \overline{u}_3 + \overline{u}_4\right)\right] + O(\sqrt{\epsilon}) \\
&\overset{(d)}{=} \frac{1}{4}\mathbb{E}\left[\overline{q}_{1}^{+}\left(3\overline{u}_2 - \overline{u}_3 + \overline{u}_4\right)\right] + O(\sqrt{\epsilon}) \\
&\overset{(e)}{=} \frac{1}{4}\mathbb{E}\left[\overline{q}_{1}^{+}\left(3\overline{u}_2 - \overline{u}_3\right)\right] + \frac{1}{4}\mathbb{E}\left[\overline{q}_{2}^{+}\overline{u}_4\right] + \frac{1}{4}\mathbb{E}\left[\overline{q}_{3}^{+}\overline{u}_4\right] + O(\sqrt{\epsilon}),
\end{aligned} \tag{7.80}$$

where (a) holds by definition of $\overline{\boldsymbol{q}}_{\|}$ and $\overline{\boldsymbol{q}}_{\perp}$; (b) holds similarly to Claim 7.30; (c) holds by Equation 7.66; (d) holds by Equation 1.3; and (e) holds by Claim 7.31. Similarly, for the fifth term we have

$$\mathbb{E}\left[\overline{q}_{\|2}^{+}\overline{u}_{\|1}\right] = \frac{1}{4}\mathbb{E}\left[\overline{q}_{2}^{+}\left(3\overline{u}_1 + \overline{u}_3 - \overline{u}_4\right)\right] + O(\sqrt{\epsilon}). \tag{7.81}$$

For the sixth term we have

$$\begin{aligned}
0 \leq \mathbb{E}\left[\overline{u}_{\|1}\overline{u}_{\|2}\right] &\overset{(a)}{\leq} \sqrt{\mathbb{E}\left[\overline{u}_{\|1}^{2}\right]\mathbb{E}\left[\overline{u}_{\|2}^{2}\right]} \\
&\leq \sqrt{\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]} \\
&= \mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]
\end{aligned}$$

where (a) holds by the Cauchy-Schwarz inequality. Also, since $\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]$ is $O(\epsilon)$, we obtain

$$\mathbb{E}\left[\overline{u}_{\|1}\overline{u}_{\|2}\right] \text{ is } O(\epsilon). \tag{7.82}$$

Using Equation 7.76, Equation 7.77, Equation 7.79, Equation 7.80, Equation 7.81 and

Equation 7.82 in Equation 7.75, and reorganizing terms we obtain

$$\begin{aligned}
&\epsilon\mathbb{E}\left[\overline{q}_{\|1}\right] + \epsilon\mathbb{E}\left[\overline{q}_{\|2}\right] \\
=&\frac{3\left(\sigma_{a_1}^{(\epsilon)}\right)^2 + 3\left(\sigma_{a_2}^{(\epsilon)}\right)^2 - \left(\sigma_{a_3}^{(\epsilon)}\right)^2 - \left(\sigma_{a_4}^{(\epsilon)}\right)^2}{8} + \frac{\epsilon^2}{2} + O(\sqrt{\epsilon}) \\
&+ \frac{1}{2}\mathbb{E}\left[\overline{q}_1^+\left(3\overline{u}_2 - \overline{u}_3\right)\right] + \frac{1}{2}\mathbb{E}\left[\overline{q}_3^+\overline{u}_4\right] + \frac{1}{2}\mathbb{E}\left[\overline{q}_2^+\left(3\overline{u}_1 + \overline{u}_3\right)\right].
\end{aligned}$$

Taking the limit as $\epsilon \downarrow 0$ on both sides we obtain Equation 7.42. The proof of Equation 7.43 and Equation 7.44 hold similarly, after setting to zero the drift of $V_5(\boldsymbol{q})$ and $V_6(\boldsymbol{q})$, respectively. This completes the proof of Theorem 7.24. □

7.11 Details of the proof of Theorem 7.24

We prove the claims we made in the proof of Theorem 7.24.

7.11.1 Proof of Claim 7.30

Proof of Claim 7.30. Observe

$$\begin{aligned}
\mathbb{E}\left[\overline{q}_{\perp 1}^+\overline{u}_{\|1}\right] &\leq \mathbb{E}\left[\left|\overline{q}_{\perp 1}^+\right|\left|\overline{u}_{\|1}\right|\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{4}\left|\overline{q}_{\perp i}^+\right|\left|\overline{u}_{\|i}\right|\right] \\
&\overset{(a)}{\leq} \sqrt{\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp\|^2\right]\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_\|\right\|^2\right]} \\
&\overset{(b)}{\leq} \sqrt{J_2}\sqrt{\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_\|\right\|^2\right]},
\end{aligned}$$

where (a) holds by Cauchy-Schwarz inequality; and (b) holds by Proposition 7.4. Similarly,

$$\mathbb{E}\left[\overline{q}_{\perp 1}^+\overline{u}_{\|1}\right] \geq -\mathbb{E}\left[\left|\overline{q}_{\perp 1}^+\right|\left|\overline{u}_{\|1}\right|\right] \geq -\sqrt{J_2}\sqrt{\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_\|\right\|^2\right]}.$$

Then,

$$\left|\mathbb{E}\left[\overline{q}_{\perp 1}^{+}\overline{u}_{\|1}\right]\right| \leq \sqrt{J_2}\sqrt{\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]}$$

and $\mathbb{E}\left[\left\|\overline{\boldsymbol{u}}_{\|}\right\|^2\right]$ is $O(\epsilon)$. This proves the claim. □

7.11.2 Proof of Claim 7.31

Proof of Claim 7.31. We use Claim 7.30. We obtain

$$\begin{aligned}
\mathbb{E}\left[\overline{q}_1^{+}\overline{u}_4\right] =& \mathbb{E}\left[\overline{q}_{\|1}^{+}\overline{u}_4\right] + O(\sqrt{\epsilon}) \\
\overset{(a)}{=}& \mathbb{E}\left[\left(\overline{q}_{\|2}^{+} + \overline{q}_{\|3}^{+} - \overline{q}_{\|4}^{+}\right)\overline{u}_4\right] + O(\sqrt{\epsilon}) \\
\overset{(b)}{=}& \mathbb{E}\left[\left(\overline{q}_2^{+} + \overline{q}_3^{+} - \overline{q}_4^{+}\right)\overline{u}_4\right] + O(\sqrt{\epsilon}) \\
\overset{(c)}{=}& \mathbb{E}\left[\left(\overline{q}_2^{+} + \overline{q}_3^{+}\right)\overline{u}_4\right] + O(\sqrt{\epsilon})
\end{aligned}$$

where (a) holds by SSC; (b) holds by Claim 7.30; and (c) holds by Equation 1.3. □

7.12 Conclusion and future work

In this chapter we studied one of the most general single-hop SPNs with control in service: the generalized switch. This model subsumes several queueing systems, such as the input-queued switch, parallel-server systems, ad hoc wireless networks, etc. Our result is widely applicable, since we do not assume the CRP condition, neither independence of the arrival processes.

We showcase the generality of our result with three particular SPNs: the input-queued switch, parallel-server systems, and an ad hoc wireless network. Each of these results are interesting by themselves since they have been studied separately in the literature, and we can easily compute them as applications of Theorem 7.5.

Additionally, we prove that if the heavy-traffic limit is to a vertex of the capacity region,

then SSC does not result in a reduction on the dimension of the state space. In other words, in this case we observe full-dimensional SSC. Under this condition, regardless of the correlation among arrival processes, the mean of the linear combinations of the queue lengths that we obtain behave as if the queues were independent in heavy traffic.

Our result is widely applicable to several SPNs, but it only allows to compute certain linear combinations of the queue lengths. In the case of an input-queued switch, this linear combination turns out to be the total queue length, and in parallel-server systems, the weights of the linear combination are the mean service rates.

We also show that obtaining other linear combinations is a nontrivial problem, since using the drift method with polynomial test functions is equivalent to solving an underdetermined system of linear equations. The results we obtain in this paper can be also obtained by taking specific linear combinations of these equations, such that some unknowns cancel out. An immediate line of future work is to extend the method so that all the linear combinations can be computed. This would allow us to also obtain higher moments and, eventually, the joint distribution of the queue lengths.

REFERENCES

[1] Y. Einav, "Amazon found every 100ms of latency cost them 1% in sales," *Gigaspaces.–01.20*, 2019.

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

[3] R Williams, "On dynamic scheduling of a parallel server system with complete resource pooling," *Fields Institute Communications*, vol. 28, no. 49-71, pp. 5–1, 2000.

[4] J. Dai and W. Lin, "Asymptotic optimality of maximum pressure policies in stochastic processing networks," *The Annals of Applied Probability*, vol. 18, no. 6, pp. 2239–2299, 2008.

[5] D. Gamarnik and A. Zeevi, "Validity of heavy traffic steady-state approximations in Generalized Jackson Networks," *The Annals of Applied Probability*, pp. 56–90, 2006.

[6] J Harrison, "Brownian models of queueing networks with heterogeneous customer populations," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, Springer, 1988, pp. 147–186.

[7] J Harrison, "Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies," *Annals of Applied Probability*, pp. 822–848, 1998.

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

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

[10] W Kang and R Williams, "Diffusion approximation for an input-queued switch operating under a maximum weight matching policy," *Stochastic Systems*, vol. 2, no. 2, pp. 277–321, 2012.

[11] J. Dai, "Steady-state approximations: Achievement lecture," in *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, ACM, 2018, pp. 1–1.

[12] D. Shah, J. Tsitsiklis, and Y. Zhong, "Optimal scaling of average queue sizes in an input-queued switch: An open problem," *Queueing Systems*, vol. 68, no. 3-4, pp. 375–384, 2011.

[13] R. Williams, "Stochastic processing networks," *Annual Review of Statistics and Its Application*, vol. 3, pp. 323–345, 2016.

[14] S. T. Maguluri and R Srikant, "Heavy traffic queue length behavior in a switch under the MaxWeight algorithm," *Stochastic Systems*, vol. 6, no. 1, pp. 211–250, 2016.

[15] S. T. Maguluri, S. Burle, and R Srikant, "Optimal heavy-traffic queue length scaling in an incompletely saturated switch," *Queueing Systems*, vol. 88, no. 3-4, pp. 279–309, 2018.

[16] G Foschini and J. Salz, "A basic dynamic routing problem and diffusion," *IEEE Transactions on Communications*, vol. 26, no. 3, pp. 320–327, 1978.

[17] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, no. 1, 181–189, 1977.

[18] N Vvedenskaya, R Dobrushin, and F Karpelevich, "Queueing system with selection of the shortest of two queues: An asymptotic approach," *Problems of Information Transmission*, vol. 32, no. 1, pp. 15–27, 1996.

[19] M. Mitzenmacher, "Load balancing and density dependent jump Markov processes," in *FOCS*, IEEE, 1996, p. 213.

[20] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 12, no. 10, pp. 1094–1104, 2001.

[21] 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*, vol. 37, no. 12, pp. 1936–1948, 1992.

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

[23] J Kingman, "On queues in heavy traffic," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 383–392, 1962.

[24] A. Braverman, J. Dai, and M. Miyazawa, "Heavy traffic approximation for the stationary distribution of a Generalized Jackson Network: The BAR approach," *Stochastic Systems*, vol. 7, no. 1, pp. 143–196, 2017.

[25] I. Gurvich, "Diffusion models and steady-state approximations for exponentially ergodic Markovian queues," *The Annals of Applied Probability*, vol. 24, no. 6, pp. 2527–2559, 2014.

[26] A. Braverman, "Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime," *Mathematics of Operations Research*, 2020.

[27] A. Braverman, J. Dai, and J. Feng, "Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models," *Stochastic Systems*, vol. 6, no. 2, pp. 301–366, 2017.

[28] A. Braverman and J. Dai, "Stein's method for steady-state diffusion approximations of M/Ph/n+ M systems," *The Annals of Applied Probability*, vol. 27, no. 1, pp. 550–581, 2017.

[29] X. Liu and L. Ying, "A simple steady-state analysis of load balancing algorithms in the sub-Halfin-Whitt regime," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 2, pp. 15–17, 2019.

[30] A. Stolyar, "Tightness of stationary distributions of a flexible-server system in the halfin-whitt asymptotic regime," *Stochastic Systems*, vol. 5, no. 2, pp. 239–267, 2015.

[31] L. 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*, ser. SIGMETRICS '16, Antibes Juan-les-Pins, France: ACM, 2016, pp. 285–297, ISBN: 978-1-4503-4266-7.

[32] L. Ying, "Stein's method for mean field approximations in light and heavy traffic regimes," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 1, 12:1–12:27, Jun. 2017.

[33] C. Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, The Regents of the University of California, 1972.

[34] A. Eryilmaz and R. Srikant, "Asymptotically tight steady-state queue length bounds implied by drift conditions," *Queueing Systems*, vol. 72, no. 3-4, pp. 311–359, 2012.

[35] A. Skorokhod, "Stochastic equations for diffusion processes in a bounded region," *Theory of Probability & Its Applications*, vol. 6, no. 3, pp. 264–274, 1961.

[36] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press, 2014, ISBN: 9781107036055.

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

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

[39] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis, "Performance of multiclass markovian queueing networks via piecewise linear Lyapunov functions," *The Annals of Applied Probability*, vol. 11, no. 4, pp. 1384–1428, Nov. 2001.

[40] A. Gut, *Probability: A Graduate Course*. Springer Science & Business Media, 2012, vol. 75.

[41] S. Mou and S. T. Maguluri, "Heavy traffic queue length behaviour in a switch under markovian arrivals," *arXiv preprint arXiv:2006.06150*, 2020.

[42] P. Jhunjhunwala and S. T. Maguluri, "Heavy traffic steady state distribution of switch system under maxweight scheduling," Working paper.

[43] S. M. Varma and S. T. Maguluri, "A heavy traffic theory of two-sided queues," Working paper.

[44] D. Hurtado-Lange and S. T. Maguluri, "Transform methods for heavy-traffic analysis," *Stochastic Systems*, vol. 10, no. 4, pp. 275–309, 2020.

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

[46] J. Köllerström, "Heavy traffic theory for queues with several servers. i," *Journal of Applied Probability*, vol. 11, no. 3, pp. 544–552, 1974.

[47] J Kingman, "The single server queue in heavy traffic," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, vol. 57, 1961, pp. 902–904.

[48] J Lehoczky, "Real-time queueing theory," *Real-Time Systems Symposium*, p. 186, 1996.

[49] J. Lehoczky, "Using real-time queueing theory to control lateness in real-time systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 25, no. 1, pp. 158–168, 1997.

[50] J Kingman, "Some inequalities for the queue GI/G/1," *Biometrika*, pp. 315–324, 1962.

[51] K. Marshall, "Some inequalities in queuing," *Operations research*, vol. 16, no. 3, pp. 651–668, 1968.

[52] D. Lindley, "The theory of queues with a single server," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, vol. 48, 1952, pp. 277–289.

[53] R. Weber, "On the optimal assignment of customers to parallel servers," *Journal of Applied Probability*, vol. 15, no. 2, pp. 406–413, 1978.

[54] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE Transactions on Automatic Control*, vol. 25, no. 4, pp. 690–693, 1980.

[55] H. Chen and H. Ye, "Asymptotic optimality of balanced routing," *Operations Research*, vol. 60, no. 1, pp. 163–179, 2012.

[56] B. Li, X. Kong, and L. Wang, "Optimal load-balancing for high-density wireless networks with flow-level dynamics," *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 316–317, 2018.

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

[58] P. Eschenfeldt and D. Gamarnik, "Join the Shortest Queue with many servers. The heavy-traffic asymptotics," *Mathematics of Operations Research*, vol. 43, no. 3, pp. 867–886, 2018.

[59] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. Larus, and A. Greenberg, "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services," *Performance Evaluation*, vol. 68, no. 11, pp. 1056–1071, 2011.

[60] A Stolyar, "Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers," *Queueing Systems*, vol. 85, no. 1-2, pp. 31–65, 2017.

[61] L. Ying, R. Srikant, and X. Kang, "The power of slightly more than one sample in randomized load balancing," *Mathematics of Operations Research*, vol. 42, no. 3, pp. 692–722, 2017.

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

[63] S. T. Maguluri, R. Srikant, and L. Ying, "Heavy traffic optimal resource allocation algorithms for cloud computing clusters," *Performance Evaluation*, vol. 81, pp. 20–39, 2014.

[64] C.-H. Wang, S. T. Maguluri, and T. Javidi, "Heavy traffic queue length behavior in switches with reconfiguration delay," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, IEEE, 2017, pp. 1–9.

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

[66] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[67] D. Hurtado-Lange and S. T. Maguluri, "Throughput and delay optimality of power-of-d choices in inhomogeneous load balancing systems," *Operations Research Letters*, 2021.

[68] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized Join-the-Shortest-Queue (JSQ) schemes in large heterogeneous processor-sharing systems," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 2, pp. 116–126, 2015.

[69] A. Mukhopadhyay, A Karthik, and R. R. Mazumdar, "Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay," *Stochastic Systems*, vol. 6, no. 1, pp. 90–131, 2016.

[70] S. Foss and N. Chernova, "On the stability of a partially accessible multi-station queue with state-dependent routing," *Queueing Systems*, vol. 29, no. 1, pp. 55–73, 1998.

[71] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications*. Springer, 1979, vol. 143.

[72] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal, "Balanced allocations," *SIAM Journal on Computing*, vol. 29, no. 1, pp. 180–200, 1999.

[73] R. Menich and R. F. Serfozo, "Optimality of routing and servicing in dependent parallel processing systems," *Queueing Systems*, vol. 9, no. 4, pp. 403–418, 1991.

[74] D. Hurtado-Lange and S. T. Maguluri, "Load balancing system under join the shortest queue: Many-server-heavy-traffic asymptotics," *arXiv preprint arXiv:2004.04826v2*, 2020.

[75] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Operations research*, vol. 29, no. 3, pp. 567–588, 1981.

[76] R. Atar, "A diffusion regime with nondegenerate slowdown," *Operations Research*, vol. 60, no. 2, pp. 490–500, 2012.

[77] R. Badonnel and M. Burgess, "Dynamic pull-based load balancing for autonomic servers," in *NOMS 2008-2008 IEEE Network Operations and Management Symposium*, IEEE, 2008, pp. 751–754.

[78] D. Mukherjee, S. C. Borst, J. S. Van Leeuwaarden, and P. A. Whiting, "Universality of power-of-d load balancing in many-server systems," *Stochastic Systems*, vol. 8, no. 4, pp. 265–292, 2018.

[79] V. Gupta and N. Walton, "Load balancing in the Nondegenerate Slowdown Regime," *Operations Research*, vol. 67, no. 1, pp. 281–294, 2019.

[80] S. Foss and A. L. Stolyar, "Large-scale join-idle-queue system with general service times," *Journal of Applied Probability*, pp. 995–1007, 2017.

[81] S. Banerjee and D. Mukherjee, "Join-the-shortest queue diffusion limit in Halfin–Whitt regime: Tail asymptotics and scaling of extrema," *The Annals of Applied Probability*, vol. 29, no. 2, pp. 1262–1309, 2019.

[82] S. Banerjee and D. Mukherjee, "Join-the-shortest queue diffusion limit in Halfin–Whitt regime: Sensitivity on the heavy-traffic parameter," *The Annals of Applied Probability*, vol. 30, no. 1, pp. 80–144, 2020.

[83] D. Mukherjee, S. C. Borst, J. S. Van Leeuwaarden, P. A. Whiting, *et al.*, "Universality of load balancing schemes on the diffusion scale," *Journal of Applied Probability*, vol. 53, no. 4, pp. 1111–1124, 2016.

[84] X. Liu and L. Ying, in *On Universal Scaling of Distributed Queues under Load Balancing*, arXiv preprint arXiv:1912.11904, 2019.

[85] Z. Zhao, S. Banerjee, and D. Mukherjee, "Many-server asymptotics for join-the-shortest queue in the super-halfin-whitt scaling window," *arXiv preprint arXiv:2106.00121*, 2021.

[86] W. Weng and W. Wang, "Dispatching parallel jobs to achieve zero queuing delay," *arXiv preprint arXiv:2004.02081*, 2020.

[87] M. Bramson, "State space collapse with application to heavy-traffic limits for multiclass queueing networks," *Queueing Systems Theory and Applications*, pp. 89 – 148, 1998.

[88] J. Dai and T. Tezcan, "State space collapse in many-server diffusion limits of parallel server systems," *Mathematics of Operations Research*, vol. 36, no. 2, pp. 271–320, 2011.

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

[90] D. Shah and D. Wischik, "Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse," *The Annals of Applied Probability*, vol. 22, no. 1, pp. 70–127, 2012.

[91] M. Miyazawa, "Diffusion approximation for stationary analysis of queues and their networks: A review," *Journal of the Operations Research Society of Japan*, vol. 58, no. 1, pp. 104–148, 2015.

[92] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International statistical review*, vol. 70, no. 3, pp. 419–435, 2002.

[93] N. Ross, "Fundamentals of Stein's method," *Probability Surveys*, vol. 8, pp. 210–293, 2011.

[94] R. L. Wheeden, *Measure and integral: an introduction to real analysis*. CRC press, 2015, vol. 308.

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

[96] G. Gupta and N. Shroff, "Delay analysis for wireless networks with single hop traffic and general interference constraints," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 2, pp. 393–405, 2010.

[97] S. Meyn, "Stability and asymptotic optimality of generalized maxweight policies," *SIAM Journal on Control and Optimization*, vol. 47, no. 6, pp. 3259–3294, 2009.

[98] D. Hurtado-Lange and S. T. Maguluri, "Heavy-traffic analysis of queueing systems with no complete resource pooling," *arXiv preprint arXiv:1904.10096*, 2019.

[99] D. Hurtado-Lange, S. M. Varma, and S. T. Maguluri, "Logarithmic heavy traffic error bounds in generalized switch and load balancing systems," *arXiv preprint arXiv:2003.07821*, 2020.

[100] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 267–280, 2010.

[101] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pp. 202–208, 2009.

[102] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997, vol. 6.

[103] A. Dimakis and J. Walrand, "Sufficient conditions for stability of longest queue first scheduling," *Adv. Appl. Prob.*, pp. 505–521, 2006.

[104] B. H. Siva Theja Maguluri and R. Srikant, "The stability of longest-queue-first scheduling with variable packet sizes," in *Proc. Conf. on Decision and Control*, 2011.

[105] X. Kang, W. Wang, J. J. Jaramillo, and L. Ying, "On the performance of largest-deficit-first for scheduling real-time traffic in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 72–84, 2014.

[106] S. L. Bell and R. J. Williams, "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy," *Annals of Applied Probability*, vol. 11, no. 3, pp. 608–649, 2001.

[107] O. Garnett and A. Mandelbaum, "An introduction to skills-based routing and its operational complexities," *Teaching notes*, 2000.

[108] C. Shi, Y. Wei, and Y. Zhong, "Process flexibility for multiperiod production systems," *Operations Research*, 2019.

[109] S. Ghamami and A. R. Ward, "Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy," *Mathematics of Operations Research*, vol. 38, no. 4, pp. 761–824, 2013.

[110] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam, "Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[111] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 39, no. 8, pp. 1600–1611, 1994.

[112] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *The Annals of Applied Probability*, pp. 43–75, 1994.