Operations Research Letters 49 (2021) 616–622

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Throughput and delay optimality of power-of-*d* choices in inhomogeneous load balancing systems

Daniela Hurtado-Lange *, Siva Theja Maguluri

Department of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, GA 30332, United States of America

ARTICLE INFO

Article history:
Received 30 March 2020
Received in revised form 3 March 2021
Accepted 8 June 2021
Available online 17 June 2021

Keywords:
Power-of-*d* choices
Load balancing
Throughput optimality
Heavy-traffic optimality

ABSTRACT

It is well-known that the power-of-*d* choices routing algorithm maximizes throughput and is heavy-traffic optimal in load balancing systems with homogeneous servers. However, if the servers are heterogeneous, throughput optimality does not hold in general. We find necessary and sufficient conditions for throughput optimality of power-of-*d* choices when the servers are heterogeneous, and we prove that almost the same conditions are sufficient to show heavy-traffic optimality. Additionally, we generalize the sufficient condition for throughput optimality to a larger class of routing policies.

1. Introduction

Load balancing systems are multi-server Stochastic Processing Networks (SPNs) with a single stream of job arrivals. A single dispatcher routes arrivals to one of the queues immediately after they enter the system and, after being routed, the jobs wait in the corresponding line until the assigned server can process them. The policy to route the jobs used by the dispatcher is called a routing algorithm. An essential goal when designing routing algorithms is to balance the workload of the servers such that the delay is minimized, and the stability region of the SPN is maximal. When a routing algorithm achieves maximal stability region, it is said to be throughput optimal. For a formal definition of throughput optimality in the case of a load balancing system, we refer the reader to Definition 2.

The most basic algorithm is random routing, under which new arrivals are routed to a queue selected uniformly at random. Advantages of this routing algorithm are that the dispatcher does not require any information about the servers' speed or queue length. However, it has been proved that it is not delay optimal and, if the servers are heterogeneous, the stability region is not maximal [15].

A popular routing algorithm is Join the Shortest Queue (JSQ), under which the new arrivals are routed to the server with the least number of jobs in line (including the one being served, if any). It has been proved that JSQ is optimal among policies that do not know job durations, under several optimality criteria. For example, [18,19] prove that JSQ maximizes the number of customers that complete service by a given time t. In [19], Poisson arrivals and exponential job sizes are assumed, whereas [18] relaxes these assumptions. In [3], it is shown that JSQ minimizes the total time needed to finish processing all the jobs that arrive by a fixed time t. All these consider a continuous time model in a general setting, i.e., without taking any asymptotic regime. In [5] it is proved that JSQ minimizes the delay in the heavy-traffic regime, i.e., when the arrival rate approaches the maximum capacity of the system. This characteristic of a policy is known as heavy-traffic optimality. More recently, [4] showed that JSQ is both throughput and heavy-traffic optimal in the context of a load balancing system operating in discrete time. In this case, instead of proving that the delay is minimized, the authors prove that the total number of jobs in the system is minimized. Even though JSQ is optimal under multiple criteria, a drawback is that it requires knowledge of all the queue lengths at any point of time.

Comparing JSQ to random routing suggests a trade-off between the expected delay and the amount of information required by the routing algorithms. A policy that can be considered to be in between them is the power-of-*d* choices algorithm, where d is an integer between 1 and the total number of servers n. Under this algorithm, d servers are sampled uniformly at random and the new arrivals are routed to the server with the shortest queue among these. If $d = 1$, then power-of-*d* is the same as random routing, and if $d = n$, it is the same as JSQ. In the case of load balancing systems with identical servers, it has been proved that even if

* Corresponding author.
E-mail addresses: d.hurtado@gatech.edu (D. Hurtado-Lange), siva.theja@gatech.edu (S.T. Maguluri).

https://doi.org/10.1016/j.orl.2021.06.010
0167-6377/

$d = 2$, power-of-d choices is throughput and heavy-traffic optimal [9]. It has also been shown that power-of-d choices yields substantial improvement in the tail probabilities of the queue lengths in the mean-field regime (i.e., when the number of servers increases to infinity) [12,13]. Also, for small values of d, the amount of information required by the dispatcher to route new arrivals is significantly smaller than under JSQ.

A disadvantage of power-of-d choices is that throughput and delay optimality have been proved only when the servers are identical. If the service rates are different, there are known counterexamples for throughput optimality [15]. 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 [2,16]. 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.

The primary contribution of this paper 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.

In [6] 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 paper, we provide conditions that only depend on the service rates and the sampling scheme. Hence, our conditions are easier to check.

The second contribution of this paper is the computation of the joint distribution of the scaled 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.

Heavy-traffic means that we analyze the system when it is loaded to its maximum capacity. In the limit, many systems behave as if their dimension was smaller, phenomenon known as State Space Collapse (SSC). For the heterogeneous load balancing system operating under power-of-d choices we prove that, in the limit, the n-dimensional queueing system behaves as a one-dimensional system, i.e., a single server queue. Then, we use this result to find the joint distribution of queue lengths. We develop our analysis in discrete time (i.e., in a time slotted fashion), so we use the notion of SSC developed in [4]. Then, we find the joint distribution of the queue lengths using the Moment Generating Function (MGF) method introduced in [8]. 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 [9]. 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.

The third contribution of this paper 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 paper, 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.

The organization of the rest of this paper is as follows. In Section 2 we formally introduce a model for the load balancing system and we define the power-of-d choices algorithm; in Section 3 we prove necessary and sufficient conditions for throughput optimality of power-of-d choices; in Section 4 we perform heavy-traffic analysis; in Section 5 we present the generalization; and in Section 6 we present details of the proofs of the previous sections.

1.1. Notation

Before establishing the details of our model we introduce our notation. We use $\mathbb{R}$ and $\mathbb{Z}$ to denote the set of real and integer numbers, respectively. We add a subscript $+$ to indicate nonnegativity, and a number in the superscript to denote vector spaces. For any number $n \in \mathbb{Z}_+$, we use $[n] \triangleq \{i \in \mathbb{Z}_+ : 1 \le i \le n\}$ and for $d \in \mathbb{Z}_+$ with $n \ge d$ we use $\binom{n}{d}$ to denote the binomial coefficient. We use bold letters to denote vectors and the same letter but not bold and with a subscript i to denote its i^{th} element. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, the notation $x_{(i)}$ refers to the i^{th} smallest element of $\boldsymbol{x}$. Given two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we use $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ to denote dot product and $\|\boldsymbol{x}\|$ to the Euclidean norm. Given a set $\mathcal{C} \subset \mathbb{R}^n$, $\text{Int}(\mathcal{C})$ denotes its interior.

If X is a random variable, then $\mathbb{E}[X]$ is its expected value and $\text{Var}[X]$ its variance. For an event A, the notation $\mathbb{1}_{\{A\}}$ is the indicator function of A. Additionally, we use the notation $\mathbb{E}_{\boldsymbol{q}}[\cdot] \triangleq \mathbb{E}[\cdot \,|\, \boldsymbol{q}(k) = \boldsymbol{q}]$ for the conditional expectation on the vector of queue lengths in time slot k.

For any function $V : \mathbb{Z}_+^n \to \mathbb{R}_+$ let

$$\Delta V(\boldsymbol{q}) \triangleq \left[V\left(\boldsymbol{q}(k+1)\right) - V\left(\boldsymbol{q}(k)\right)\right] \mathbb{1}_{\{\boldsymbol{q}(k) = \boldsymbol{q}\}}.$$

Thus, $\Delta V(\boldsymbol{q})$ is a random variable that measures the amount of change in the value of V in one step, starting from $\boldsymbol{q}$. We refer to $\Delta V(\boldsymbol{q})$ as the drift of $V(\boldsymbol{q})$.

2. Model

We model the load balancing system in discrete time, i.e., in a time slotted fashion, and we use $k \in \mathbb{Z}_+$ to index time. Consider a system with n servers, each of them with an infinite buffer. Let $\boldsymbol{q}(k)$ be the vector of queue lengths at the beginning of time slot k, i.e., for each $i \in [n]$, $q_i(k)$ is the number of jobs in queue i when the k^{th} time slot starts including the job in service, if any. There is a single stream of arrivals to the system, and a dispatcher routes all the arrivals of each time slot to one of the queues, according to some routing policy. Let $\{a(k) : k \in \mathbb{Z}_+\}$ be a sequence of i.i.d. random variables such that $a(k)$ is the total number of arrivals in time slot k. The vector $\boldsymbol{a}(k)$ represents the number of jobs that arrive to each of the queues in time slot k after routing. Then, if the dispatcher routes the arrivals to queue i^*, we have $a_{i^*}(k) = a(k)$ and $a_i(k) = 0$ for all $i \neq i^*$. We assume that the routing time is negligible and that the routing decision depends only on the current vector of queue lengths $\boldsymbol{q}(k)$. Let $\boldsymbol{s}(k)$ be the potential service vector in time slot k, i.e., for each $i \in [n]$, $s_i(k)$ is the number of jobs that can be processed in queue i in time slot k if there are enough jobs in line. Let $\{\boldsymbol{s}(k) : k \in \mathbb{Z}_+\}$ be a sequence of i.i.d. random vectors, which is independent of the arrival and the queue length processes. The difference between the potential and actual service is called unused service, and we use $\boldsymbol{u}(k)$ to denote the vector of unused service in time slot k. Observe that $\boldsymbol{u}(k)$ is a function of $\boldsymbol{q}(k)$, $\boldsymbol{a}(k)$ and $\boldsymbol{s}(k)$.

We assume that the arrivals and routing occur before service in each time slot. Then, the following equation describes the dynamics of the queues. For each $i \in [n]$ and each $k \in \mathbb{Z}_+$,

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k). \tag{1}$$

From (1), observe that $\{\boldsymbol{q}(k) : k \in \mathbb{Z}_+\}$ is a Discrete Time Markov chain (DTMC). Also, for every $i \in [n]$,

$$q_i(k+1)u_i(k) = 0 \tag{2}$$

because the unused service in queue i is nonzero only if the potential service to that queue is larger than the number of jobs available to be served (queue length and arrivals). Therefore, if the unused service is nonzero, the queue is empty at the beginning of the next time slot.

We assume that the arrival and the potential service to each queue have finite second moment. Let $\lambda \triangleq \mathbb{E}[a(1)]$, $\boldsymbol{\mu} \triangleq \mathbb{E}[\boldsymbol{s}(1)]$ and $\mu_\Sigma \triangleq \sum_{i=1}^n \mu_i$. 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 $\sigma_a^2 \triangleq \mathrm{Var}[a(1)]$ be the variance of the arrival process and Σ_s the covariance matrix of $\boldsymbol{s}(1)$. For each $i \in [n]$, define $\sigma_{s_i}^2 \triangleq (\Sigma_s)_{i,i}$. It is well known that the capacity region of the load balancing system is

$$\mathcal{C} \triangleq \{x \in \mathbb{R}_+ : x \leq \mu_\Sigma\}, \tag{3}$$

i.e., for each $\lambda \in \mathrm{Int}(\mathcal{C})$, 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 [4].

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

Definition 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 $i^{*\text{th}}$ 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.

3. Throughput optimality of power-of-d choices

In this section we state and prove the main theorem of this paper. Before presenting the result we formally define throughput optimality.

Definition 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 $\lambda \in \mathrm{Int}(\mathcal{C})$, where $\mathcal{C}$ is defined in (3).

Now we present the main theorem of this paper.

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

$$\mathcal{M}^{(d)} \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}_+^n : \frac{\sum_{i=1}^j y_{(i)}}{y_\Sigma} \geq \frac{\binom{j}{d}}{\binom{n}{d}} \ \forall d \leq j \leq n-1 \right\}, \tag{4}$$

where $y_\Sigma \triangleq \sum_{i=1}^n y_i$. Then, the power-of-d choices algorithm is throughput optimal for the load balancing system described in Section 2 if and only if $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$.

Remark 1. Observe that we can equivalently define $\mathcal{M}^{(d)}$ for all $d \in [n]$ as follows

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

where we use the convention $\binom{j}{d} = 0$ if $j < d$. Here we only added redundant constraints to $\mathcal{M}^{(d)}$, so we use the definition in (4) to avoid confusion.

Remark 2. An interpretation of Theorem 1 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 [15]. Equation (4) 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 (4) 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 (4) 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 3. For $i \in [n]$, define $\nu_i \triangleq \frac{\binom{i-1}{d-1}}{\binom{n}{d}}$, 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 [10]. This notion has been used in the study of balls and bins models in [1], and to prove optimality of routing and servicing algorithms in [11]. 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. Theorem 1 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 $\lambda \in \mathrm{Int}(\mathcal{C})$ 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 $\lambda \in \mathrm{Int}\left(\overline{\mathcal{C}}\right)$, where

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

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 1.

In the proof of Theorem 1 we use the Foster-Lyapunov theorem [17, Theorem 3.3.7] and a certificate that a DTMC is not positive recurrent [17, Theorem 3.3.10].

Proof of Theorem 1. Let $\epsilon \triangleq \mu_\Sigma - \lambda$, and observe that $\lambda \in \mathrm{Int}(\mathcal{C})$ 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 with Lyapunov function $V(\boldsymbol{q}) = \|\boldsymbol{q}\|^2$. We have

$$\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q}(k))] = \mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k+1)\|^2 - \|\boldsymbol{q}(k)\|^2\right]$$
$$\overset{(a)}{=} \mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{q}(k) + \boldsymbol{a}(k) - \boldsymbol{s}(k)\|^2 - \|\boldsymbol{u}(k)\|^2 - \|\boldsymbol{q}(k)\|^2\right]$$

$$\overset{(b)}{\leq} \mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\|^2\right] + 2\mathbb{E}_{\boldsymbol{q}}\left[\langle \boldsymbol{q}, \boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle\right], \tag{5}$$

where (a) holds after few algebraic steps, using (1) and (2); and (b) holds because $\|\boldsymbol{u}(k)\|^2 \geq 0$ and after expanding the first term. We analyze each of the terms in (5) separately. For the first term, we have

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

where (a) holds because all the arrivals in one time slot are routed to the same queue; and (b) holds by definition of variance. Define $K_1 \triangleq \lambda^2 + \sigma_a^2 + \sum_{i=1}^{n}\left(\mu_i^2+\sigma_{s_i}^2\right)$, and observe K_1 is a finite constant. Then,

$$\mathbb{E}_{\boldsymbol{q}}\left[\|\boldsymbol{a}(k)-\boldsymbol{s}(k)\|^2\right] \leq K_1. \tag{6}$$

Observe that the computation of the bound (6) does not use any properties of the routing algorithm. In other words, the bound (6) is valid for the load balancing system under any routing algorithm.

To compute the second term of (5), we first compute $\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle]$. 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

$$\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle] = \lambda \sum_{i=1}^{n-d+1} q_{(i)} \frac{\binom{n-i}{d-1}}{\binom{n}{d}} \tag{7}$$

because there are $\binom{n-i}{d-1}$ ways to sample d queues, and make sure that $q_{(i)}$ is the shortest; and there are $\binom{n}{d}$ ways to sample d queues uniformly at random. If there are ties on the queue lengths, power-of-d breaks them at random. Hence, the result in (7) 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 (5) is

$$\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle] = \mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k)\rangle] - \langle \boldsymbol{q}, \boldsymbol{\mu}\rangle$$
$$= \sum_{i=1}^{n-d+1} q_{(i)}\left(\frac{\lambda\binom{n-i}{d-1}}{\binom{n}{d}} - \mu_{\phi(i)}\right) - \sum_{i=n-d+2}^{n} q_{(i)}\mu_{\phi(i)}. \tag{8}$$

Define

$$\alpha_i \triangleq \begin{cases} \dfrac{\lambda\binom{n-i}{d-1}}{\binom{n}{d}} - \mu_{\phi(i)} & \text{, if } 1 \leq i \leq n-d+1 \\ -\mu_{\phi(i)} & \text{, if } n-d+1 < i \leq n. \end{cases} \tag{9}$$

Claim 2. *The parameters α_i defined in (9) satisfy*

1. $\alpha_n \leq -\mu_1$.
2. $\sum_{i=1}^{n}\alpha_i = -\epsilon$.
3. *For any $j \in \mathbb{Z}_+$ satisfying $2 \leq j \leq n-1$, we have $\sum_{i=j}^{n}\alpha_i \leq -K_2$, where $K_2 \triangleq \min\left\{\mu_1, \frac{\epsilon}{\binom{n}{d}}\right\}$.*

We prove Claim 2 in Section 6.1. Now we compute an upper bound for (8). We obtain

$$\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k)-\boldsymbol{s}(k)\rangle] = \sum_{i=1}^{n}\alpha_i q_{(i)}$$
$$= q_{(1)}\sum_{i=1}^{n}\alpha_i + \sum_{j=2}^{n}\left(\sum_{i=j}^{n}\alpha_i\right)\left(q_{(j)}-q_{(j-1)}\right)$$
$$\overset{(a)}{\leq} -\epsilon q_{(1)} - K_2\sum_{j=2}^{n}\left(q_{(j)}-q_{(j-1)}\right)$$
$$\overset{(b)}{=} q_{(1)}(K_2-\epsilon) - K_2 q_{(n)} \overset{(c)}{\leq} -K_2 q_{(n)}, \tag{10}$$

where (a) holds by properties 2 and 3 in Claim 2; (b) holds after solving the telescopic sum and rearranging terms; and (c) holds because $K_2 \leq \frac{\epsilon}{\binom{n}{d}}$ by definition, and $\binom{n}{d} \geq 1$. Using (6) and (10) in (5) we obtain

$$\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q}(k))] \leq K_1 - 2K_2 q_{(n)}.$$

This inequality is sufficient to prove the conditions of the Foster-Lyapunov theorem. 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 $\lambda \in \text{Int}(\mathcal{C})$ 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 $j \in \mathbb{Z}_+$ such that $d \leq j \leq n-1$ and $\frac{\sum_{i=1}^{j}\mu_i}{\mu_\Sigma} < \frac{\binom{j}{d}}{\binom{n}{d}}$. Let j^* be the smallest j satisfying this condition, and $\delta_{j^*} > 0$ satisfy

$$\frac{\sum_{i=1}^{j^*}\mu_i}{\mu_\Sigma} + \delta_{j^*} = \frac{\binom{j^*}{d}}{\binom{n}{d}}. \tag{11}$$

Using the Lyapunov function $V_{j^*}(\boldsymbol{q}) = \sum_{i=1}^{j^*} q_i$, we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{j^*}(\boldsymbol{q}(k))\right] = \sum_{i=1}^{j^*}\mathbb{E}_{\boldsymbol{q}}[a_i(k) - s_i(k) + u_i(k)]$$
$$\overset{(a)}{\geq} \sum_{i=1}^{j^*}\mathbb{E}_{\boldsymbol{q}}[a_i(k)] - \sum_{i=1}^{j^*}\mu_i$$
$$\overset{(b)}{\geq} \sum_{i=1}^{j^*}\mathbb{E}_{\boldsymbol{q}}\left[a_{\tilde{\phi}(i)}(k)\right] - \sum_{i=1}^{j^*}\mu_i \overset{(c)}{=} \mu_\Sigma\delta_{j^*} - \epsilon\frac{\binom{j^*}{d}}{\binom{n}{d}},$$

where (a) holds because $\mathbb{E}[s_i(k)] = \mu_i$ and $\mathbb{E}[u_i(k)] \geq 0$ 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; and (c) holds computing $\mathbb{E}_{\boldsymbol{q}}\left[a_{\tilde{\phi}(i)}(k)\right]$ similarly to (7), and reorganizing terms.

If $\epsilon > 0$ satisfies $\epsilon \leq \mu_\Sigma \min\left\{1, \delta_{j^*}\frac{\binom{j^*}{d}}{\binom{n}{d}}\right\}$, then we have $\mathbb{E}_{\boldsymbol{q}}\left[V_{j^*}(\boldsymbol{q}(k+1)) - V_{j^*}(\boldsymbol{q}(k))\right] \geq 0$ for all $\boldsymbol{q} \in \mathbb{R}_+^n$. Additionally, we need to prove that $\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{j^*}(\boldsymbol{q}(k))\right]$ is bounded. We have

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{j^*}(\boldsymbol{q}(k))\right] \overset{(a)}{\leq} \sum_{i=1}^{j^*}\mathbb{E}_{\boldsymbol{q}}[a_i(k)] \overset{(b)}{\leq} \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{q}}[a_i(k)] \overset{(c)}{=} \lambda,$$

where (a) holds because $u_i(k) \leq s_i(k)$ with probability 1, by definition of unused service; (b) holds because the number of arrivals to each queue is nonnegative; and (c) holds by definition of the routing algorithm and λ. This proves the theorem.

4. Heavy-traffic analysis

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 we find the limiting joint distribution of the vector of queue lengths scaled by the heavy-traffic parameter.

Heavy traffic means that we load the system close to its maximum capacity. To take the limit we parametrize the system as follows. 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 $\lambda^{(\epsilon)} \triangleq \mathbb{E}\left[a^{(\epsilon)}(1)\right] = \mu_\Sigma - \epsilon$. 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 a one-dimensional subspace. In other words, we show that, in the limit, the n-dimensional load balancing system operating under power-of-d choices behaves as a single server queue. Before showing the result we introduce the following notation. For any vector $\boldsymbol{x} \in \mathbb{R}^n$, define

$$\boldsymbol{x}_\parallel = \mathbf{1}\left(\frac{\sum_{i=1}^n x_i}{n}\right), \ \boldsymbol{x}_\perp \triangleq \boldsymbol{x} - \boldsymbol{x}_\parallel. \tag{12}$$

Then, $\boldsymbol{x}_\parallel$ is the projection of $\boldsymbol{x}$ on the line $\{\boldsymbol{z} \in \mathbb{R}^n : z_i = z_j \, \forall i, j \in [n]\}$ and $\boldsymbol{x}_\perp$ is the error of approximating $\boldsymbol{x}$ by $\boldsymbol{x}_\parallel$. Now we present the result.

Proposition 3. *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 $k \to \infty$. Let $\delta > 0$ be such that for all $j \in \mathbb{Z}_+$ satisfying $d \leq j \leq n-1$ we have*

$$\frac{\sum_{i=1}^j \mu_i}{\mu_\Sigma} - \delta \geq \frac{\binom{j}{d}}{\binom{n}{d}}. \tag{13}$$

If $\epsilon < \delta\mu_\Sigma$, then $\mathbb{E}\left[\|\overline{\boldsymbol{q}}_\perp^{(\epsilon)}\|^m\right] \leq M_m$ for each $m = 1, 2, \ldots$, where M_m is a finite constant (independent of ϵ).

Proposition 3 says that the error of approximating $\overline{\boldsymbol{q}}^{(\epsilon)}$ by $\overline{\boldsymbol{q}}_\parallel^{(\epsilon)}$ 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 $\overline{\boldsymbol{q}}_\parallel^{(\epsilon)}$ also becomes larger. However, the error of approximating $\overline{\boldsymbol{q}}^{(\epsilon)}$ by $\overline{\boldsymbol{q}}_\parallel^{(\epsilon)}$, denoted as $\overline{\boldsymbol{q}}_\perp^{(\epsilon)}$, has bounded moments. Then, as ϵ goes to zero it becomes negligible.

Observe that the vector $\overline{\boldsymbol{q}}^{(\epsilon)}$ is well defined, because $\boldsymbol{\mu} \in \text{Int}\left(\mathcal{M}^{(d)}\right) \subset \mathcal{M}^{(d)}$. Then, from Theorem 1, for all $\epsilon > 0$ the DTMC $\left\{\boldsymbol{q}^{(\epsilon)}(k) : k \in \mathbb{Z}_+\right\}$ is positive recurrent.

In the proof of Proposition 3 we use a result first presented in [4, Lemma 1], which is a corollary of a result proved in [7].

Proof of Proposition 3. For ease of exposition, we omit the dependence on ϵ of the variables. Define $V(\boldsymbol{q}) \triangleq \|\boldsymbol{q}\|^2$, $V_\parallel(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_\parallel\|^2$, $W_\perp(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_\perp\|$. We use the Lyapunov function $W_\perp(\boldsymbol{q})$. We start with a fact first used in [4]. 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,

$$\Delta W_\perp(\boldsymbol{q}) \leq \frac{1}{2\|\boldsymbol{q}_\perp\|}\left(\Delta V(\boldsymbol{q}) - \Delta V_\parallel(\boldsymbol{q})\right). \tag{14}$$

We need to show two conditions. In the first condition we show that $\mathbb{E}_{\boldsymbol{q}}[\Delta W_\perp(\boldsymbol{q}(k))]$ is negative if $\boldsymbol{q}$ lies outside a bounded set, and in the second condition we show that $\mathbb{E}_{\boldsymbol{q}}[\Delta W_\perp(\boldsymbol{q}(k))]$ is bounded.

To prove the first one, we find an upper bound to $\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q})]$ and a lower bound to $\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_\parallel(\boldsymbol{q})\right]$. We start with $\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q})]$. From the proof of Theorem 1, we know (6) is satisfied. We analyze the last term differently here. Defining $\phi(i)$ as in the proof of Theorem 1, we have

$$\mathbb{E}_{\boldsymbol{q}}[\langle \boldsymbol{q}, \boldsymbol{a}(k) - \boldsymbol{s}(k)\rangle] \overset{(a)}{=} -\epsilon\left(\frac{\sum_{i=1}^n q_i}{n}\right) + \sum_{i=1}^n q_{(i)}\beta_i,$$

where (a) holds reorganizing terms, and defining

$$\beta_i \triangleq \begin{cases} \frac{\binom{n-i}{d-1}}{\binom{n}{d}}\lambda + \frac{\epsilon}{n} - \mu_{\phi(i)} & \text{, if } 1 \leq i \leq n-d+1 \\ \frac{\epsilon}{n} - \mu_{\phi(i)} & \text{, if } n-d+1 < i \leq n. \end{cases} \tag{15}$$

Claim 4. *The parameters β_i defined in (15) satisfy*

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

We prove Claim 4 in Section 6.2. Observe that if $d = 1$, the second property is not satisfied. Using Claim 4 we obtain

$$\begin{aligned}\sum_{i=1}^n q_{(i)}\beta_i &= q_{(1)}\sum_{i=1}^n \beta_i + \sum_{j=2}^n \left(\sum_{i=j}^n \beta_i\right)\left(q_{(j)} - q_{(j-1)}\right) \\ &\leq \left(-\delta\mu_\Sigma + \epsilon\right)\left(q_{(n)} - q_{(1)}\right).\end{aligned} \tag{16}$$

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

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

where (a) holds because $q_i \leq q_{(n)}$ for all $i \in [n]$ and $\frac{1}{n}\sum_{j=1}^n q_j \geq q_{(1)}$ by definition of $q_{(1)}$ and $q_{(n)}$. Using this result in (16) we obtain that

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

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

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{q}}[\Delta V(\boldsymbol{q}(k))] \\ &\leq K_1 - 2\epsilon\left(\frac{\sum_{i=1}^n q_i}{n}\right) + 2\left(\frac{-\delta\mu_\Sigma + \epsilon_0}{\sqrt{n}}\right)\left\|\boldsymbol{q}_\perp\right\|.\end{aligned} \tag{17}$$

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

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{\parallel}(\boldsymbol{q}(k))\right] \geq -2\epsilon\left(\frac{\sum_{i=1}^{n} q_i}{n}\right) - K_3, \tag{18}$$

where $K_3 \triangleq 2nS_{\max}^2$, and $S_{\max}$ is a finite constant such that $s_i(1) \leq S_{\max}$ for all $i \in [n]$ with probability 1. Using (17) and (18) in (14) we obtain

$$\mathbb{E}_{\boldsymbol{q}}\left[\Delta W_{\perp}(\boldsymbol{q}(k))\right] \leq \frac{K_1 + K_3}{2\left\|\boldsymbol{q}_{\perp}\right\|} + \left(\frac{-\delta\mu_{\Sigma} + \epsilon_0}{\sqrt{n}}\right),$$

which proves the first condition of [4, Lemma 1]. The second condition is trivially satisfied because the arrival and service random variables are bounded.

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 5. *Consider a set of load balancing systems operating under power-of-d as described in Proposition 3. 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 $\frac{1}{2n}\left(\sigma_a^2 + \mathbf{1}^T \Sigma_s \mathbf{1}\right)$, and $\Longrightarrow$ denotes convergence in distribution.*

Remark 5. In Proposition 3 and Theorem 5 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 homogeneous service rates $\boldsymbol{\mu} = c\mathbf{1}$ (with $c > 0$) satisfies all the inequalities in (4), and none of them is tight. Then, such $\boldsymbol{\mu} = c\mathbf{1} \in \operatorname{Int}\left(\mathcal{M}^{(d)}\right)$. 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 5. We use the MGF method [8], which is a two-step procedure to compute the joint distribution of the scaled vector of queue lengths in heavy traffic, in queueing systems that experience one-dimensional SSC. In fact, our theorem is a corollary of [8, Theorem 2]. We only verify that three conditions are satisfied.

We first verify that the routing algorithm is throughput optimal, which holds from Theorem 1 because we assume $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$. The second condition is SSC to a one-dimensional subspace, which is satisfied by Proposition 3. The last condition is existence of the MGF of $\epsilon \sum_{i=1}^{n} \overline{q}_i$, which we formalize in Claim 6 and prove in Section 6.3.

Claim 6. *For the load balancing system described in Theorem 5, 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 $\theta \in [-\Theta, \Theta]$.*

5. Generalization to other routing policies

In this section we generalize the sufficient conditions in Theorem 1 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 7. *Given $\pi : 2^{[n]} \to [0, 1]$, consider a load balancing system as described in Section 2, 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*

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

where $\mathcal{S}_i^{\tau} \triangleq \{\mathcal{S} \subseteq [n] : \ \tau(n-i+1) \in \mathcal{S}, \ \tau(\ell) \notin \mathcal{S} \ \ \forall \ell < n-i+1\}$.
$\mathcal{R}^{\pi}$ 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 1, and we present a sketch in Section 6.4 for completeness.

6. Details of the proofs in Sections 3, 4 and 5

6.1. Proof of Claim 2

Proof of Claim 2. 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 $2 \leq j \leq n-d+1$ we have that the tail sums are

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

   where (a) holds because $\sum_{i=j}^{n-d+1} \binom{n-i}{d-1} = \binom{n+1-j}{d}$; (b) holds by definition of ϵ and because $\boldsymbol{\mu} \in \mathcal{M}^{(d)}$; and (c) holds because $\binom{n+1-j}{d} \geq 1$, and because $\sum_{i=1}^{n+1-j} \mu_i$ is the sum of the $n+j-1$ smallest elements of $\boldsymbol{\mu}$. If $n-d+1 < j \leq n-1$ we have

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

6.2. Proof of Claim 4

Proof of Claim 4. Properties 1 and 2 follow immediately from the fact that $\beta_i = \alpha_i + \frac{\epsilon}{n}$. To prove the third property we divide in cases. If $j \leq n-d+1$ we have

$$\sum_{i=j}^{n} \beta_i = \sum_{i=j}^{n-d+1} \frac{\binom{n-i}{d-1}}{\binom{n}{d}}\lambda + \sum_{i=j}^{n}\left(\frac{\epsilon}{n} - \mu_{\phi(i)}\right) \leq \frac{\binom{n+1-j}{d}}{\binom{n}{d}}\mu_{\Sigma} + \epsilon - \sum_{i=1}^{n-j+1} \mu_i \overset{(a)}{\leq} \epsilon - \delta\mu_{\Sigma},$$

where (a) holds by (13) and reorganizing terms.

If $j > n-d+1$ we have

$$\sum_{i=j}^{n} \beta_i = \sum_{i=j}^{n}\left(\frac{\epsilon}{n} - \mu_{\phi(i)}\right)$$

$$\leq \frac{n-j+1}{n}\epsilon - \sum_{i=1}^{n-j+1} \mu_i \overset{(a)}{\leq} \epsilon - \delta\mu_\Sigma,$$

where (a) holds by (13) and because $\frac{n-j+1}{n} \leq 1$.

6.3. Existence of MGF

Proof of Claim 6. The proof is similar to the proof of existence of MGF under JSQ routing, which was done in [8]. We write a sketch of the proof here for completeness. First observe that if $\theta \leq 0$, the proof holds trivially. Now, assume $\theta > 0$. Observe that $f(x) = e^x$ is a convex function. Then, by Jensen's inequality, we have

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

Then, it suffices to show that $\mathbb{E}\left[\sum_{i=1}^n e^{\theta\epsilon q_i}\right] < \infty$ for $\theta \in [-\Theta, \Theta]$, for all $i \in [n]$. We use the Foster-Lyapunov theorem [17, Theorem 3.3.7] with function $V_{MGF}(\boldsymbol{q}) = \sum_{i=1}^n e^{\theta\epsilon q_i}$. For each $i \in [n]$ we have

$$\left(e^{\theta\epsilon q_i(k+1)} - 1\right)\left(e^{-\theta\epsilon u_i(k)} - 1\right) = 0,$$

which holds by (2). Then, reorganizing terms and summing over $i \in [n]$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[\Delta V_{MGF}(\boldsymbol{q}(k))\right] \\
&= \sum_{i=1}^n \left(1 - \mathbb{E}\left[e^{-\theta\epsilon u_i(k)}\right]\right) \\
&+ \sum_{i=1}^n e^{\theta\epsilon q_{(i)}}\left(\mathbb{E}_{\boldsymbol{q}}\left[e^{\theta\epsilon\left(a_{\phi(i)}(k) - s_{\phi(i)}(k)\right)}\right] - 1\right),
\end{aligned} \tag{19}$$

where $\phi(i)$ is defined as in the proof of Theorem 1. Since $\boldsymbol{u}(k) \geq \boldsymbol{0}$ and $\theta > 0$, the first term is upper bounded by n. Now, for a bounded random variable X, define $M_X(\theta) \triangleq \mathbb{E}\left[e^{\theta\epsilon X}\right]$. Then, for each $i \in [n]$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{q}}\left[e^{\theta\epsilon\left(a_{\phi(i)}(k) - s_{\phi(i)}(k)\right)}\right] - 1 \\
&= M_{a_{\phi(i)} - s_{\phi(i)}}(\theta) - 1 \overset{(a)}{=} \theta M'_{a_{\phi(i)} - s_{\phi(i)}}(\xi_i),
\end{aligned}$$

where (a) holds by Taylor expansion, for a number ξ_i between 0 and θ. Observe that the MGF is continuously differentiable at $\theta = 0$ [14, p.78] and

$$M'_{a_{\phi(i)} - s_{\phi(i)}}(0) = \mathbb{E}_{\boldsymbol{q}}\left[a_{\phi(i)}(k) - s_{\phi(i)}(k)\right] = \alpha_i,$$

where α_i was defined in (9). For each $i \in [n]$, let $\tilde{\Theta}_i > 0$ be such that for all θ between 0 and $\tilde{\Theta}_i$ we have

$$M'_{a_{\phi(i)} - s_{\phi(i)}}(\xi_i) \leq \frac{1}{2}\alpha_i.$$

Let $\tilde{\Theta} = \min_{i\in[n]} \tilde{\Theta}_i$. Then, for all θ satisfying $\theta\epsilon < \tilde{\Theta}$

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

The rest of the proof is equivalent to the last steps of the proof of throughput optimality, so we omit it for brevity. The proof concludes by letting $\Theta = n\tilde{\Theta}$.

6.4. Proof of Theorem 7

Proof of Theorem 7. The proof is very similar to Theorem 1. 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 1 we obtain the simple expression presented in (7). 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.

7. Acknowledgments

This work was partially supported by the National Science Foundation [NSF-CCF: 1850439]. Daniela Hurtado-Lange has partial funding from ANID/DOCTORADO BECAS CHILE/2018 [72190413]

References

[1] Y. Azar, A.Z. Broder, A.R. Karlin, E. Upfal, Balanced allocations, SIAM J. Comput. 29 (1) (1999) 180–200.
[2] H. Chen, H. Ye, Asymptotic optimality of balanced routing, Oper. Res. 60 (1) (2012) 163–179.
[3] A. Ephremides, P. Varaiya, J. Walrand, A simple dynamic routing problem, IEEE Trans. Autom. Control 25 (4) (1980) 690–693.
[4] A. Eryilmaz, R. Srikant, Asymptotically tight steady-state queue length bounds implied by drift conditions, Queueing Syst. 72 (3–4) (2012) 311–359.
[5] G. Foschini, J. Salz, A basic dynamic routing problem and diffusion, IEEE Trans. Commun. 26 (3) (1978) 320–327.
[6] S. Foss, N. Chernova, On the stability of a partially accessible multi-station queue with state-dependent routing, Queueing Syst. 29 (1) (1998) 55–73.
[7] B. Hajek, Hitting-time and occupation-time bounds implied by drift analysis with applications, Adv. Appl. Probab. (1982) 502–525.
[8] D. Hurtado-Lange, S.T. Maguluri, Transform methods for heavy-traffic analysis, Stoch. Syst. 10 (4) (2020) 275–309.
[9] S.T. Maguluri, R. Srikant, L. Ying, Heavy traffic optimal resource allocation algorithms for cloud computing clusters, Perform. Eval. 81 (2014) 20–39.
[10] A.W. Marshall, I. Olkin, B.C. Arnold, Inequalities: Theory of Majorization and Its Applications, vol. 143, Springer, 1979.
[11] R. Menich, R.F. Serfozo, Optimality of routing and servicing in dependent parallel processing systems, Queueing Syst. 9 (4) (1991) 403–418.
[12] M. Mitzenmacher, Load balancing and density dependent jump Markov processes, in: FOCS, IEEE, 1996, p. 213.
[13] M. Mitzenmacher, The power of two choices in randomized load balancing, IEEE Trans. Parallel Distrib. Syst. 12 (10) (2001) 1094–1104.
[14] A. Mood, Introduction to the Theory of Statistics, McGraw-Hill, 1950.
[15] A. Mukhopadhyay, R.R. Mazumdar, Analysis of randomized Join-the-Shortest-Queue (JSQ) schemes in large heterogeneous processor-sharing systems, IEEE Trans. Control Netw. Syst. 3 (2) (2015) 116–126.
[16] A. Mukhopadhyay, A. Karthik, R.R. Mazumdar, Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay, Stoch. Syst. 6 (1) (2016) 90–131.
[17] R. Srikant, L. Ying, Communication Networks: An Optimization, Control and Stochastic Networks Perspective, Cambridge University Press, 2014.
[18] R. Weber, On the optimal assignment of customers to parallel servers, J. Appl. Probab. 15 (2) (1978) 406–413.
[19] W. Winston, Optimality of the shortest line discipline, J. Appl. Probab. 14 (1) (1977) 181–189, https://doi.org/10.1017/S0021900200104772.