J. Appl. Probab. **59,** 652–669 (2022)
*doi:*10.1017/jpr.2021.82

LOGARITHMIC HEAVY TRAFFIC ERROR BOUNDS IN GENERALIZED SWITCH AND LOAD BALANCING SYSTEMS

DANIELA HURTADO-LANGE,*
SUSHIL MAHAVIR VARMA ,** *** AND
SIVA THEJA MAGULURI,** *Georgia Institute of Technology*

Abstract

Motivated by applications to wireless networks, cloud computing, data centers, etc., stochastic processing networks have been studied in the literature under various asymptotic regimes. In the heavy traffic regime, the steady-state mean queue length is proved to be $\Theta(1/\epsilon)$, where ϵ is the heavy traffic parameter (which goes to zero in the limit). The focus of this paper is on obtaining queue length bounds on pre-limit systems, thus establishing the rate of convergence to heavy traffic. For the generalized switch, operating under the MaxWeight algorithm, we show that the mean queue length is within $O(\log(1/\epsilon))$ of its heavy traffic limit. This result holds regardless of the complete resource pooling (CRP) condition being satisfied. Furthermore, when the CRP condition is satisfied, we show that the mean queue length under the MaxWeight algorithm is within $O(\log(1/\epsilon))$ of the optimal scheduling policy. Finally, we obtain similar results for the rate of convergence to heavy traffic of the total queue length in load balancing systems operating under the 'join the shortest queue' routeing algorithm.

Keywords: Drift method; state space collapse; MaxWeight; generalized switch; load balancing

2020 Mathematics Subject Classification: Primary 60K25; 68M20; 90B22; 60H99

1. Introduction

Resource allocation and load balancing problems arise frequently in a wide variety of applications such as wireless networks, data centers, ride hailing systems (e.g. Uber and Lyft), routeing and congestion control of traffic, manufacturing, telecommunications, etc. Performance analysis of these systems is frequently addressed by modeling them as stochastic processing networks (SPNs) [12], and essential performance measures are delay and queue lengths. Exact analysis of these measures usually becomes intractable, so a common practice is to study asymptotic regimes. Heavy traffic is a popular regime, where one studies the behavior of the system as the load grows to the maximum capacity. The heavy traffic limit of queue lengths and delay provides meaningful insights about the actual performance of the systems, but an essential question is whether the limiting behavior is close to the behavior in all traffic. In [2], [4], and [7], Eryilmaz, Hurtado-Lange, Maguluri, and Srikant obtain the heavy

Received 4 April 2020; revision received 22 August 2021.
* Postal address: Department of Mathematics, William & Mary, Jones Hall, Room 100, 200 Ukrop Way, Williamsburg, VA 23185, USA.
** Postal address: H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive, NW, Atlanta, GA 30332, USA.
*** Email address: sushil@gatech.edu

traffic limit of linear combinations of the expected queue lengths. However, the rate of convergence to this limit is not studied. In other words, they compute error bounds on the queue lengths that vanish in heavy traffic, but they are not optimized. In this paper we show that these error bounds grow logarithmically, as opposed to the polynomial bounds obtained in [2], [4], and [7].

Most of the literature on heavy traffic analysis is for systems that satisfy the so-called complete resource pooling (CRP) condition. Under this condition, the system exhibits state space collapse (SSC) onto a line, and hence it behaves as a single-server queue in the heavy traffic limit. There are several methodologies to study systems that satisfy CRP, such as diffusion limits [11], transform methods [5], and Lyapunov drift-based arguments [2]. The latter has also been used to show that the steady-state mean of a linear combination of the queue lengths is of the form $K_1/\epsilon + o(1/\epsilon)$, where K_1 is an appropriately defined constant and ϵ is a parameter representing how far away the arrival rate vector is from the boundary of the capacity region.

In this paper we study a generalized switch model, which was first introduced in [11] to study several SPNs with control on the service process, such as input-queued switches, *ad hoc* wireless networks, cloud computing, data centers, etc. We consider the MaxWeight algorithm, and, using a tighter variant of the drift argument in [2], [4], and [7], we show that MaxWeight is within $K_2 \log(1/\epsilon)$ of the optimal policy under the CRP condition (see Corollary 2.1). This is the first contribution of this paper.

We also study a generalized switch without assuming that the CRP condition is satisfied, and we improve the bounds presented in [4] without adding any assumption. Specifically, we compute an upper bound of the form $K_1/\epsilon + K_2 \log(1/\epsilon)$ for linear combinations of the expected queue lengths (see Theorem 2.1). This establishes a logarithmically growing error bound with respect to the limiting queue length, which is of the form K_1/ϵ. This is the second contribution of this paper.

In addition to systems where the service is controlled, we look at load balancing systems, where the control is on the arrivals. We consider the popular 'join the shortest queue' (JSQ) algorithm, which is known to exhibit a one-dimensional SSC [2]. We show that the mean sum of the queue lengths is $K_1'/\epsilon + K_2' \log(1/\epsilon)$ (see Theorem 3.1) which, in conjunction with the universal lower bound (ULB) shown in [2], establishes that JSQ is within $K_2' \log(1/\epsilon)$ of the optimal routeing policy. This is the third contribution of this paper. Similar results can be obtained for other routeing algorithms, such as power-of-d choices. However, we focus on JSQ in this paper, for brevity.

1.1. Literature review

The work closest to ours in the literature is that of Meyn [8], who studied a general resource allocation problem under the CRP condition. Meyn showed that a variation of the MaxWeight algorithm, called h-MaxWeight, achieves logarithmic optimality. Our result is similar in flavor but has two main distinctions. Firstly, our result is valid for the non-CRP case as well. Secondly, in [8] Meyn worked with h-MaxWeight, where h is a function that needs to be computed by analyzing the first-order approximation of the system. In this paper we use drift-based arguments that are easier to generalize to other SPNs. We showcase this generalization by analyzing the load balancing system under JSQ. Singh and Stolyar [10] and Sharifnassab, Tsitsiklis, and Golestani [9] studied systems where the CRP condition is not satisfied. In particular, Singh and Stolyar [10] also studied a generalized switch operating under the MaxWeight algorithm in heavy traffic. However, in contrast to the present work, the focus in [10] is to show that the service provided to each of the queues is smooth across time, i.e. there are not large gaps in

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

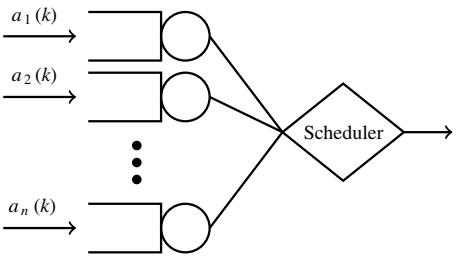


FIGURE 1. Generalized switch model.

service. This is done by characterizing the deviations of the queue lengths from the region of SSC. While we also obtain a bound on such deviations, the motivation, technique, and results are different.

Sharifnassab, Tsitsiklis, and Golestani [9] studied the more general setting of a multi-hop switched network with a general arrival process, operating under the MaxWeight scheduling algorithm. They analyzed the first-order approximation of the system (typically known as the 'fluid model') and bounded the gap between the fluid and the stochastic model in terms of the arrival process. Their results are applicable to a broader class of systems than the current paper, but they did not study the heavy traffic steady-state behavior.

1.2. Notation

We denote the set of integers from 1 to n by $[n]$. We denote the set of real and integer numbers by $\mathbb{R}$ and $\mathbb{Z}$, respectively, and we add a subscript $+$ to denote the subset of non-negative numbers. We define the greatest integer smaller than or equal to $x \in \mathbb{R}_+$ by $\lfloor x \rfloor$. All the vectors in the paper are boldface. The sets of n-dimensional vectors with real components and non-negative real components are denoted by $\mathbb{R}^n$ and $\mathbb{R}^n_+$, respectively. We denote the dot product between two vectors by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and the Euclidean norm of a vector by $\|\boldsymbol{x}\|$. We denote the ith canonical vector by $\boldsymbol{e}^{(i)}$, the vector of ones by $\mathbf{1}$, and the vector of zeros by $\mathbf{0}$. We denote the transpose of a matrix by $A^\top$, and the Hadamard product between two matrices by $A \circ B$. The expectation and variance of a random variable X are given by $\mathbb{E}[X]$ and $\mathrm{Var}[X]$, respectively, and the covariance between two random variables X and Y by $\mathrm{Cov}(X, Y)$. The probability of an event E is denoted by $\mathbb{P}[E]$, and the indicator function of an event E by $\mathbb{1}_{\{E\}}$. For a set S we use $\mathrm{Int}(S)$ and $\mathrm{Bo}(S)$ to denote its relative interior and its relative boundary, respectively.

2. Logarithmic error bounds in the generalized switch

2.1. Model

In this section we present the generalized switch model in detail. Consider n queues operating in discrete time, with time indexed by $k \in \mathbb{Z}_+$. A pictorial example is presented in Figure 1.

2.1.1. *Arrival process.* We define a sequence of i.i.d. random variables $\{a_i(k) \colon k \in \mathbb{Z}_+\}$ for all $i \in [n]$, where $a_i(k)$ denotes the number of jobs that arrive at the ith queue at time k. Denote the mean arrival rate vector by $\boldsymbol{\lambda} \stackrel{\triangle}{=} \mathbb{E}[\boldsymbol{a}(1)]$ and the covariance matrix of the random vector $\boldsymbol{a}(1)$ by Σ_a. Assume $a_i(1) \leq A_{\max}$ with probability 1 for all $i \in [n]$, where $A_{\max}$ is a finite constant.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

2.1.2. *Service process.* Let $s_i(k)$ be the potential service that can be offered by server i in time slot k. If there are not enough jobs to serve in the queue, there is unused service in that time slot, and we denote it by $u_i(k)$. Then the actual number of served jobs in queue i at time k is $s_i(k) - u_i(k)$. We allow interference among the servers, which requires them to satisfy a set of feasibility constraints in each time slot. The scheduler is allowed to pick any service rate vector that satisfies these constraints in each time slot. Additionally, the environment of the servers can affect the interference constraints. We capture this by a sequence of i.i.d. random variables $\{M(k)\colon k \in \mathbb{Z}_+\}$, where $M(k)$ is the 'channel state' in time slot k. We assume that the channel state has a finite state space, denoted by $\mathcal{M}$, and that the interference constraints in time slot k are completely determined by the value of $M(k)$. Let the pmf of $M(1)$ be $\psi_m \stackrel{\triangle}{=} \mathbb{P}[M(1) = m]$ for all $m \in \mathcal{M}$. Finally, $\mathcal{S}^{(m)}$ denotes the set of feasible service rate vectors in channel state m. Observe that $\mathcal{S}^{(m)}$ contains the potential (not necessarily actual) service rates that satisfy the interference constraints in channel state m and, therefore, for any $\boldsymbol{x} \in \mathcal{S}^{(m)}$, all the non-negative vectors that are dominated by $\boldsymbol{x}$ are also feasible. In other words, if $\boldsymbol{0} \le \boldsymbol{y} \le \boldsymbol{x}$, then $\boldsymbol{y}$ is also a feasible service rate vector. For simplicity, we only consider maximal feasible service rate vectors in each set $\mathcal{S}^{(m)}$ and their projection on the coordinate axes, and we assume that $\mathcal{S}^{(m)}$ is a finite set for each $m \in \mathcal{M}$. Thus there exists a finite constant $S_{\max}$ such that $s_i(1) \le S_{\max}$ with probability 1 for all $i \in [n]$.

2.1.3. *Queueing process.* The following steps are followed in each time slot (in this order).

- Observe the channel state and queue length vector.
- A scheduling problem is solved to determine which queues are served and the service rates are determined according to the channel state.
- Arrivals occur in the system.
- Jobs are processed according to the selected schedule.

Then the queue dynamics follow the recursion

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k) \quad \text{for all k} \in \mathbb{Z}_+,\ i \in [n]. \tag{2.1}$$

Thus $\{\boldsymbol{q}(k)\colon k \in \mathbb{Z}_+\}$ is a discrete-time Markov chain with countable state space. If the unused service is positive, then the queue length at the start of the next time slot should be zero and *vice versa*. Thus we have

$$q_i(k+1)u_i(k) = 0 \quad \text{for all } k \in \mathbb{Z}_+,\ i \in [n]. \tag{2.2}$$

The scheduling problem is solved using the MaxWeight scheduling algorithm, which selects the schedule with the maximum total weighted queue length. Mathematically, provided that $M(k) = m$, we have

$$\boldsymbol{s}(k) \in \arg\max_{\boldsymbol{x} \in \mathcal{S}^{(m)}} \langle \boldsymbol{q}(k), \boldsymbol{x} \rangle, \tag{2.3}$$

and the ties are broken randomly. Observe that, unless there are ties, the potential service vector is deterministic after observing the channel state and the queue length vector.

2.1.4. *Capacity region.* It is proved in [2] that the capacity region of this system is $\mathcal{C} = \sum_{m \in \mathcal{M}} \psi_m \,\text{ConvexHull}(\mathcal{S}^{(m)})$. Thus it is a coordinate convex polytope. We describe it as the intersection of finitely many half-spaces, i.e. we write

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}_+^n \colon \left\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \right\rangle \le b^{(\ell)},\ \ell = 1, \ldots, L\right\}.$$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Without loss of generality, we assume $\boldsymbol{c}^{(\ell)} \geq \boldsymbol{0}$, $\|\boldsymbol{c}^{(\ell)}\| = 1$ and $b^{(\ell)} > 0$ for all $\ell \in [L]$. We also denote the ℓth facet as $\mathcal{F}^{(\ell)} \triangleq \{\boldsymbol{x} \in \mathcal{C} \colon \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle = b^{(\ell)}\}$. In addition, we denote the maximum $\boldsymbol{c}^{(\ell)}$-weighted service rate by $b^{(m,\ell)}$. Mathematically, we have

$$b^{(m,\ell)} = \max_{\boldsymbol{x} \in \mathcal{S}^{(m)}} \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \quad \text{for all } \ell \in [L].$$

To capture the randomness in the service process due to the channel state, we define a sequence of i.i.d. random variables $\{B_\ell(k) \colon k \in \mathbb{Z}_+\}$ (independent of queue lengths and arrival process) with pmf given by $\mathbb{P}[B_\ell(1) = b^{(m,\ell)}] = \psi_m$. Let the covariance matrix of the vector $\{B_\ell(1)\}_{\ell \in [L]}$ be Σ_B.

2.1.5. *Heavy traffic and state space collapse.* In heavy traffic, we take the limit as the vector of arrival rates approaches the boundary of the capacity region. Formally, we fix a vector $\boldsymbol{\nu}$ on the boundary of $\mathcal{C}$. For $\epsilon \in (0, 1)$, we let $\boldsymbol{\lambda}^{(\epsilon)} \triangleq (1 - \epsilon)\boldsymbol{\nu}$ be the mean arrival rate vector, and we take the limit as $\epsilon \downarrow 0$. Specifically, we analyze a sequence of generalized switches parametrized by ϵ and denote the queue length, arrival process, service process, and unused service for the ϵth system by $\{\boldsymbol{q}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$, $\{\boldsymbol{a}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$, $\{\boldsymbol{s}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$, and $\{\boldsymbol{u}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$, respectively. The parametrization is such that $\mathbb{E}[\boldsymbol{a}^{(\epsilon)}(1)] = \boldsymbol{\lambda}^{(\epsilon)}$. Then the heavy traffic regime is observed as $\epsilon \downarrow 0$.

For every $\epsilon \in (0, 1)$, observe that $\boldsymbol{\lambda}^{(\epsilon)} \in \mathrm{Int}(\mathcal{C})$ and therefore the queue length process $\{\boldsymbol{q}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$ of the generalized switch operating under MaxWeight is a positive recurrent discrete-time Markov chain. Thus the steady-state vector of queue lengths is well-defined. A proof of positive recurrence is presented in [2]. We denote all the steady-state vectors with a bar on top of the variable. In particular, $\overline{\boldsymbol{q}}^{(\epsilon)}$ is a steady-state random vector that is the limit in distribution of $\boldsymbol{q}^{(\epsilon)}(k)$ as $k \to \infty$. In addition, let $\overline{\boldsymbol{a}}^{(\epsilon)}$, $\overline{M}$, $\overline{B}_\ell$ be the steady-state random vector/variable with the same distribution as $\boldsymbol{a}^{(\epsilon)}(1)$, $M(1)$, $B_\ell(1)$, respectively. We have $\mathbb{E}[\overline{\boldsymbol{a}}^{(\epsilon)}] = \boldsymbol{\lambda}^{(\epsilon)}$, and denote the covariance matrix of $\overline{\boldsymbol{a}}^{(\epsilon)}$ by $\Sigma_a^{(\epsilon)}$. Also, denote the steady-state offered service by $\overline{\boldsymbol{s}}^{(\epsilon)}$ and the steady-state unused service by $\overline{\boldsymbol{u}}^{(\epsilon)}$. Finally, let $(\overline{\boldsymbol{q}}^{(\epsilon)})^+ \triangleq \overline{\boldsymbol{q}}^{(\epsilon)} + \overline{\boldsymbol{a}}^{(\epsilon)} - \overline{\boldsymbol{s}}^{(\epsilon)} + \overline{\boldsymbol{u}}^{(\epsilon)}$ be the vector of queue lengths one time slot after $\overline{\boldsymbol{q}}^{(\epsilon)}$.

Define the cone $\mathcal{K}$ spanned by the normal to the facets $\mathcal{F}^{(\ell)}$ that intersect at $\boldsymbol{\nu}$, i.e. the facets such that $\boldsymbol{\nu} \in \mathcal{F}^{(\ell)}$. Let $P \triangleq \{\ell \in [L] \colon \boldsymbol{\nu} \in \mathcal{F}^{(\ell)}\}$. It was shown in [4] that, as ϵ decreases to zero, the vector of queue lengths concentrates around the cone $\mathcal{K}$. In other words, it was shown that the projection of the vector of queue lengths on the cone $\mathcal{K}$ approximates the actual vector of queue lengths, and the error of approximation is bounded by a finite (but unknown) constant. Therefore this result is a notion of SSC. In Proposition 2.1 we prove an explicit expression for this upper bound. Observe that the cone $\mathcal{K}$ can also be represented as

$$\mathcal{K} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^n \colon \boldsymbol{x} = \sum_{\ell \in P} \xi_\ell \boldsymbol{c}^{(\ell)}, \ \xi_\ell \geq 0 \text{ for all } \ell \in P \right\}.$$

In addition, define $\mathcal{H}$ as the affine hull of $\mathcal{K}$, and let $\tilde{P} \subset P$ be the maximal set of indices in P such that $\{\boldsymbol{c}^{(\ell)} \colon \ell \in \tilde{P}\}$ is a set of linearly independent vectors. Let $C \triangleq [\boldsymbol{c}^{(\ell)}]_{\ell \in \tilde{P}}$ be a matrix with columns $\boldsymbol{c}^{(\ell)}$ with $\ell \in \tilde{P}$, and observe that $\mathcal{H}$ is the column space of C.

2.2. Logarithmic error bounds

In this section we present the main result of this paper. Specifically, we provide error bounds of linear combinations of the expected queue lengths as $\epsilon \downarrow 0$. After stating the result, we

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

discuss two applications of the result. Then in Section 2.3 we prove SSC, which is an essential step in the proof of Theorem 2.1, and in Section 2.4 we prove the theorem.

Theorem 2.1. *Consider a set of generalized switches operating under the MaxWeight scheduling policy, parametrized by the heavy traffic parameter* $\epsilon \in (0, 1)$ *as described in Section 2.1. Then there exists* $\epsilon_0 \in (0, 1)$ *such that for any* $\epsilon < \epsilon_0$ *and any vector* $\boldsymbol{w} \in \bigcap_{\ell \in P} \mathcal{F}^{(\ell)}$*, we have*

$$\left| \mathbb{E}\left[\langle \bar{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w} \rangle\right] - \frac{1}{2\epsilon} \mathbf{1}^\top \left(H \circ \Sigma_a^{(\epsilon)}\right) \mathbf{1} - \frac{1}{2\epsilon} \mathbf{1}^\top \left(\left(C^\top C\right)^{-1} \circ \Sigma_B\right) \mathbf{1} \right| \leq \beta \log\left(\frac{1}{\epsilon}\right), \tag{2.4}$$

where $H \triangleq C(C^\top C)^{-1} C^\top$ *is the projection matrix into* $\mathcal{H}$ *and* β *is a constant independent of* ϵ *and* $\boldsymbol{w}$*.*

A similar result establishing the heavy traffic behavior of a generalized switch when the CRP condition is not necessarily satisfied, was presented in [4]. The main difference is that the result in [4] shows that the right-hand side term in (2.4) is $o(1/\epsilon)$, and we obtain a tighter bound.

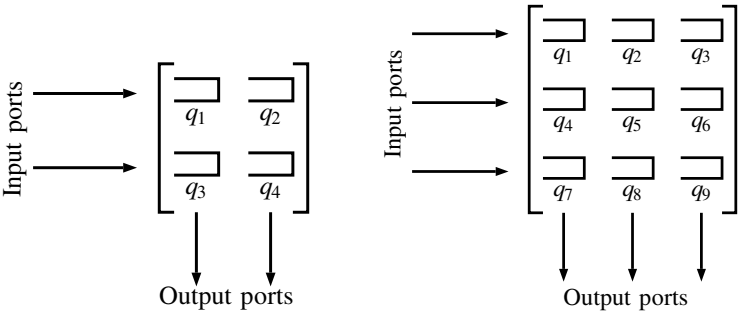
Theorem 2.1 presents a logarithmic error bound of the queue length behavior in light traffic (positive ϵ) with respect to the heavy traffic behavior. However, the result is not sufficient to claim optimality of MaxWeight because we are not comparing MaxWeight against any other scheduling policy in this paper. One way to prove such a result for a scheduling algorithm $\mathcal{A}$ would be to obtain a ULB (i.e. a lower bound for the linear combination of queue lengths that is satisfied by all scheduling policies), and then prove that this ULB is achieved by the generalized switch operating under $\mathcal{A}$. One can compute such a ULB for the generalized switch [4, Proposition 1], but this ULB is not necessarily for the same linear combination of queue lengths as presented in Theorem 2.1. Hence we cannot conclude optimality of MaxWeight. Further, there are counterexamples that prove that MaxWeight is not optimal. In [6], Lu et al. show the existence of a gap between the performance of a 2×2 switch (which is a particular case of the generalized switch) and the ULB computed in [7] (which is a particular case of the ULB computed in [4]). Specifically, they show that the ULB and the performance of MaxWeight differ by a multiplicative constant. Hence, in general, MaxWeight need not be within an additive error of $O(\log(1/\epsilon))$ from the optimal policy.

Such a logarithmic optimality of MaxWeight can be obtained from Theorem 2.1 when the CRP condition is satisfied and SSC occurs in a one-dimensional subspace. In this case the heavy traffic limit is known to be the same as the ULB of the scaled expected linear combination of the queue length. Specifically, if we fix $\ell \in [L]$ and assume $\boldsymbol{\nu} \in \mathrm{Int}\left(\mathcal{F}^{(\ell)}\right)$, the CRP condition is satisfied and SSC occurs in the line generated by $\boldsymbol{c}^{(\ell)}$. Then, as shown in [4, Proposition 1], for any scheduling algorithm, we have

$$\mathbb{E}\left[\langle \bar{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)} \rangle\right] \geq \mathrm{ULB} \triangleq \frac{1}{2\epsilon b^{(\ell)}} \left(\left(\boldsymbol{c}^{(\ell)}\right)^\top \Sigma_a^{(\epsilon)} \boldsymbol{c}^{(\ell)} + \sigma_{B_\ell}^2\right) - \frac{(1-\epsilon) b_{\max}}{2},$$

where $b_{\max} \triangleq \max_{m \in \mathcal{M}, \ell \in [L]} \left\{b^{(m,\ell)}\right\}$ and $\sigma_{B_\ell}^2 \triangleq (\Sigma_B)_{\ell,\ell}$. By Theorem 2.1, we know that MaxWeight approaches the above lower bound as $\epsilon \downarrow 0$. Thus Theorem 2.1 establishes that MaxWeight is within $O(\log(1/\epsilon))$ of the optimal policy under the CRP condition. We formally present this result in the next corollary.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press



(a) 2 × 2 switch. The feasible schedules are to serve queues q_1, q_4 or q_2, q_3.

(b) 3 × 3 switch. The feasible schedules are to serve q_1, q_5, q_9 or q_1, q_6, q_8 or q_2, q_4, q_9 or q_2, q_6, q_7 or q_3, q_4, q_8 or q_3, q_5, q_7.

FIGURE 2. Illustration of queue length vector in input-queued switch.

Corollary 2.1. *For the generalized switch operating under MaxWeight, as described in Theorem 2.1, fix $\ell \in [L]$ and assume $\boldsymbol{\nu} \in \mathrm{Int}\big(\mathcal{F}^{(\ell)}\big)$. Then, for any $\epsilon < \epsilon_0$, we have*

$$\mathrm{ULB} \le \mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)} \rangle\big] \le \mathrm{ULB} + \tilde{\beta} \log\left(\frac{1}{\epsilon}\right),$$

where $\tilde{\beta}$ is a constant independent of ϵ, which we compute in (A.2). Hence the MaxWeight algorithm is heavy traffic optimal and the error bound to the optimal value is $\mathrm{O}(\log(1/\epsilon))$.

We present the proof of Corollary 2.1 in Appendix A. An immediate corollary of Theorem 2.1 is to compute the bounds in the case of an input-queued switch. An input-queued switch is a generalized switch where the number of queues is a perfect square, say $n = N^2$, the channel state is fixed over time, and the feasibility constraints are known. Specifically, the input-queued switch can be thought of as an $N \times N$ matrix, where the (i, j)th component of the matrix is the queue of packets at ith input port, waiting to be processed at the jth output port. Thus rows are queues at input ports and columns are queues at output ports. All jobs take exactly one time slot to be processed and, in each time slot, at most one input/output pair can be served in each row and column. Then the set of feasible service rate vectors is the set of $N \times N$ permutation matrices. In Figure 2 we present a pictorial example of a 2 × 2 switch (Figure 2a) and a 3 × 3 switch (Figure 2b).

Below we present the performance bound for this system under the assumption that the arrival rate to all of the queues is $(1 - \epsilon)/N$, and hence all the queues are saturated in the heavy traffic limit.

Corollary 2.2. *For the input-queued switch defined above with independent arrivals, heavy traffic parameter $\epsilon \in (0, 1)$, and $\big(\sigma_{a_i}^{(\epsilon)}\big)^2 \triangleq \Sigma_{i,i}^{(\epsilon)}$, there exists a constant $\overline{\beta}$ and $\epsilon_0 \in (0, 1)$ such that for all $\epsilon < \epsilon_0$*

$$\left| \mathbb{E}\left[\sum_{i=1}^{N^2} \overline{q}_i^{(\epsilon)} \right] - \left(\frac{1}{\epsilon} - \frac{1}{2N\epsilon} \right) \sum_{i=1}^{N^2} \big(\sigma_{a_i}^{(\epsilon)}\big)^2 \right| \le \overline{\beta} \log\left(\frac{1}{\epsilon}\right).$$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

The proof involves simplifying the left-hand side of (2.4) and is omitted as it is similar to [4, Corollary 1].

2.3. State space collapse

We start by introducing some notation. For each $\epsilon \in (0, 1)$, let $\boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{K}}(k)$ and $\boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{H}}(k)$ be the projection of $\boldsymbol{q}^{(\epsilon)}(k)$ on $\mathcal{K}$ and $\mathcal{H}$ respectively, and

$$\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{K}}(k) \triangleq \boldsymbol{q}^{(\epsilon)}(k) - \boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{K}}(k), \quad \boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{H}}(k) \triangleq \boldsymbol{q}^{(\epsilon)}(k) - \boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{H}}(k).$$

Finally, we denote the steady-state vectors by $\overline{\boldsymbol{q}}^{(\epsilon)}_{\parallel\mathcal{K}}$, $\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{K}}$, $\overline{\boldsymbol{q}}^{(\epsilon)}_{\parallel\mathcal{H}}$, and $\overline{\boldsymbol{q}}^{(\epsilon)}_{\perp\mathcal{H}}$, which are the limit in distribution of $\boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{K}}(k)$, $\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{K}}(k)$, $\boldsymbol{q}^{(\epsilon)}_{\parallel\mathcal{H}}(k)$, and $\boldsymbol{q}^{(\epsilon)}_{\perp\mathcal{H}}(k)$ as $k \to \infty$, respectively. The steady-state vectors are well-defined as the above Markov chains are positive recurrent by the definition of projection and by the fact that $\{\boldsymbol{q}^{(\epsilon)}(k) \colon k \in \mathbb{Z}_+\}$ is positive recurrent for all $\epsilon \in (0, 1)$ [11, Proposition 2].

It was proved in [4] that $\|\overline{\boldsymbol{q}}_{\perp\mathcal{K}}\|$ has bounded moments, where the bounds do not depend on ϵ. Here we explicitly compute a bound, and later we use it to obtain the heavy traffic error bounds.

Proposition 2.1. *For the generalized switch model operating under MaxWeight parametrized by $\epsilon \in (0, 1)$ described in Section 2.1, consider a vector $\boldsymbol{\nu} \in \mathrm{Bo}(\mathcal{C})$. Let $\delta > 0$ be such that $\delta \le b^{(\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle$ for all $\ell \in [L] \setminus P$ if $P \subsetneq [L]$ and $\delta = 1$ if $P = [L]$. Let $\alpha \triangleq \max\{A_{\max}, S_{\max}\}$. If $\epsilon < \delta/(2\|\boldsymbol{\nu}\|)$, then for each $r = 1, 2, \ldots$ we have*

$$\mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp\mathcal{H}}\|^r\big] \le \mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp\mathcal{K}}\|^r\big] \le R_r \triangleq \left(\frac{8n\alpha^2}{\delta}\right)^r + \left(8\sqrt{n}\alpha\right)^r \left(\frac{8\sqrt{n}\alpha + \delta}{\delta}\right)^r r!$$

We present the proof in Appendix B.

2.4. Proof of Theorem 2.1.

The first part of our proof is similar to the proof of [4, Theorem 1], so we omit some steps. In the second part we show how to obtain logarithmic error bounds with respect to the heavy traffic limit. These bounds are tighter than the bounds presented in [4, Theorem 1].

Proof of Theorem 2.1. We omit the dependence on ϵ of the variables for ease of exposition.

It suffices to show the result for $\boldsymbol{w} = \boldsymbol{\nu}$, for the following reason. For any $\boldsymbol{w} \in \bigcap_{\ell \in P} \mathcal{F}^{(\ell)}$, let $\boldsymbol{w}_\perp = \boldsymbol{w} - \boldsymbol{\nu}$. Then for every $\ell \in P$ we have $\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{w}_\perp \rangle = 0$. Hence, since $\overline{\boldsymbol{q}}_{\parallel\mathcal{H}} = \sum_{\ell \in P} \tilde{\xi}_\ell \boldsymbol{c}^{(\ell)}$ for some $\tilde{\xi}_\ell \in \mathbb{R}$ by definition of the subspace $\mathcal{H}$, we have

$$\langle \overline{\boldsymbol{q}}, \boldsymbol{w} \rangle = \langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \boldsymbol{w} \rangle = \langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \boldsymbol{\nu} + \boldsymbol{w}_\perp \rangle = \langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \boldsymbol{\nu} \rangle = \langle \overline{\boldsymbol{q}}, \boldsymbol{\nu} \rangle.$$

We start by defining the Lyapunov function $V_{\parallel\mathcal{H}}(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_{\parallel\mathcal{H}}\|^2$. To set the drift of this Lyapunov function to zero in the steady state, we first verify that $\mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\parallel\mathcal{H}}\|^2\big] < \infty$. We omit this step for brevity, but it can be shown using the moment bounds obtained from the Foster–Lyapunov theorem [3, Proposition 6.16] with Lyapunov function $V(\boldsymbol{q}) = \|\boldsymbol{q}\|^2$, and

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

non-expansivity of the projection onto a convex set. Setting the drift of $V_{\|\mathcal{H}}(\boldsymbol{q})$ to zero in the steady state, we obtain

$$\begin{aligned}
0 &= \mathbb{E}\big[\|\overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}}\|^2 - \|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\|^2\big] \\
&= \mathbb{E}\big[\|\overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}} - \overline{\boldsymbol{u}}_{\|\mathcal{H}} + \overline{\boldsymbol{u}}_{\|\mathcal{H}}\|^2 - \|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\|^2\big] \\
&= \mathbb{E}\big[\|\overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}} - \overline{\boldsymbol{u}}_{\|\mathcal{H}}\|^2 + \|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\|^2 + 2\langle \overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}} - \overline{\boldsymbol{u}}_{\|\mathcal{H}},\ \overline{\boldsymbol{u}}_{\|\mathcal{H}}\rangle - \|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\|^2\big] \\
&\overset{(a)}{=} \mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\|\mathcal{H}} + \overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\|^2 - \|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\|^2 + 2\langle \overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}}, \overline{\boldsymbol{u}}_{\|\mathcal{H}}\rangle - \|\overline{\boldsymbol{q}}_{\|\mathcal{H}}\|^2\big] \\
&\overset{(b)}{=} \underbrace{\mathbb{E}\big[\|\overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\|^2\big]}_{\mathcal{T}_2} + \underbrace{2\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}},\ \overline{\boldsymbol{a}}_{\|\mathcal{H}} - \overline{\boldsymbol{s}}_{\|\mathcal{H}}\rangle\big]}_{-\mathcal{T}_1} - \underbrace{\mathbb{E}\big[\|\overline{\boldsymbol{u}}_{\|\mathcal{H}}\|^2\big]}_{\mathcal{T}_3} \\
&\quad + \underbrace{2\mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{+}_{\|\mathcal{H}},\ \overline{\boldsymbol{u}}_{\|\mathcal{H}}\rangle\big]}_{\mathcal{T}_4},
\end{aligned} \tag{2.5}$$

where (a) holds by definition of norm and dot product, and by the queue dynamics in (2.1), and (b) holds by expanding the first norm square and reorganizing terms. Thus we have

$$\mathcal{T}_1 = \mathcal{T}_2 - \mathcal{T}_3 + \mathcal{T}_4. \tag{2.6}$$

Observe that in the computation of (2.5) we only used properties of the Euclidean norm, the dot product, and the dynamics of the queues (2.1). Then (2.5) is valid for any SPN that satisfies (2.1).

We compute each $\mathcal{T}_i$ for $i = 1, 2, 3, 4$ separately. The terms $\mathcal{T}_1$ and $\mathcal{T}_4$ are the bottlenecks for the optimal error bounds, so we compute them at the end of this proof. We start with $\mathcal{T}_2$ and $\mathcal{T}_3$, which we borrow from [4].

The term $\mathcal{T}_2$ is related to the square of the arrivals and services, and therefore the covariance matrix of both processes is involved. Also, the arrival and service rate vectors are projected on $\mathcal{H}$. Then, using the least-squares problem, we can compute the desired norm. The details of this computation can be verified in [4, equations (38)–(42)], and we omit them for brevity. We obtain that there exists a constant β_1 such that

$$\left|\mathcal{T}_2 - \mathbf{1}^\top \big(H \circ \Sigma_a^{(\epsilon)}\big)\mathbf{1} - \mathbf{1}^\top \left(\left(C^\top C\right)^{-1} \circ \Sigma_B \mathbf{1}\right)\right| \leq \beta_1 \epsilon. \tag{2.7}$$

Now we compute $\mathcal{T}_3$. Observe that the unused service vector is bounded, because the potential service rate vector is also bounded. Additionally, the more loaded the system, the least unused service we should expect. Translating these intuitions into mathematics yields (2.8), where β_2 is a positive constant. A formal proof can be found in [4, equations (43)–(44)]. We omit the details for brevity:

$$|\mathcal{T}_3| \leq \beta_2 \epsilon. \tag{2.8}$$

Now we focus on the terms $\mathcal{T}_1$ and $\mathcal{T}_4$. We start with $\mathcal{T}_1$:

$$\mathcal{T}_1 = 2\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}}_{\|\mathcal{H}} - \overline{\boldsymbol{a}}_{\|\mathcal{H}}\rangle\big] \overset{(a)}{=} 2\epsilon\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \boldsymbol{\nu}\rangle\big] + 2\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\|\mathcal{H}}, \overline{\boldsymbol{s}} - \boldsymbol{\nu}\rangle\big],$$

where (a) follows by first using the orthogonality principle and then substituting $\mathbb{E}[\overline{\boldsymbol{a}}] = (1-\epsilon)\boldsymbol{\nu}$ and observing that $\overline{\boldsymbol{a}}$ is independent of $\overline{\boldsymbol{q}}_{\|\mathcal{H}}$. Observe that the first term is part of (2.4), which is the expression we want to obtain. Then we only need to bound the second term. We present the result in Claim 2.1, and we prove it at the end of the section.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Claim 2.1. *Consider the system described in Theorem 2.1. Then there exist $\epsilon_0' > 0$ and a finite constant β_3 such that*

$$\left|\mathbb{E}\left[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}\rangle\right]\right| \le \beta_3 \epsilon \log\left(\frac{1}{\epsilon}\right) \quad \text{for all } \epsilon < \epsilon_0'.$$

For $\mathcal{T}_4$ we have the following result.

Claim 2.2. *Consider the system described in Theorem 2.1. Then there exist $\epsilon_0'' > 0$ and a finite constant β_4 such that*

$$\mathcal{T}_4 \le \beta_4 \epsilon \log\left(\frac{1}{\epsilon}\right) \quad \text{for all } \epsilon < \epsilon_0''.$$

The proofs of both claims are presented at the end of the section. Now, using (2.7), (2.8), and Claims 2.1, 2.2 in (2.6), we obtain that for any $\epsilon < \epsilon_0 \stackrel{\triangle}{=} \min\{\epsilon_0', \epsilon_0''\}$

$$\left|\mathbb{E}\left[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{w}\rangle\right] - \frac{1}{2\epsilon}\mathbf{1}^\top\left(H \circ \Sigma_a^{(\epsilon)}\right)\mathbf{1} - \frac{1}{2\epsilon}\mathbf{1}^\top\left(\left(C^\top C\right)^{-1} \circ \Sigma_B\right)\mathbf{1}\right| \le \beta \log\left(\frac{1}{\epsilon}\right),$$

where $\beta \stackrel{\triangle}{=} \max\{\beta_1, \beta_2, \beta_3, \beta_4\}$. □

Now we prove the claims. The main idea is to use Hölder's inequality and Proposition 2.1 with the right choice of the parameter r. We additionally use the following result, which is proved in [4, Lemmas 2 and 3], and we intuitively explain below.

Lemma 2.1. *Let $\ell \in P$ and $m \in \mathcal{M}$.*

(i) *Then there exists $\boldsymbol{v}^{(m)} \in \mathcal{S}^{(m)}$ such that $b^{(m,\ell)} = \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{v}^{(m)}\rangle$. This implies that, for each $\ell \in P$,*

$$b^{(\ell)} = \mathbb{E}\left[B_\ell\right] = \sum_{m\in\mathcal{M}} \psi_m b^{(m,\ell)}.$$

(ii) *Define $\pi^{(m,\ell)} \stackrel{\triangle}{=} \mathbb{P}\left[\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle = b^{(m,\ell)} \mid \overline{M} = m\right]$. Then $1 - \pi^{(m,\ell)} \le \epsilon b^{(m,l)}/\gamma^{(m)}$, where*

$$\gamma^{(m)} \stackrel{\triangle}{=} \min\left\{b^{(m,\ell)} - \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle \colon \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x}\rangle < b^{(m,\ell)},\ \ell \in P,\ \boldsymbol{x} \in \mathcal{S}^{(m)}\right\}$$

is positive and finite.

One difficulty of the generalized switch model is that the vector of potential service rates obtained from MaxWeight (see (2.3)) does not necessarily belong to the capacity region $\mathcal{C}$. Given the channel state $\overline{M} = m$, we know that the vector of potential service satisfies $\bar{\boldsymbol{s}} \in \mathcal{S}^{(m)}$. However, the capacity region $\mathcal{C}$ is a convex combination of the sets ConvexHull($\mathcal{S}^{(m)}$) for all $m \in \mathcal{M}$. Hence there is no guarantee that the feasible service rate vectors belong to $\mathcal{C}$.

The first part of the lemma shows that the location parameter $b^{(\ell)}$ of each facet of $\mathcal{C}$ is a convex combination of the location parameter of hyperplanes that pass through the boundary of each set ConvexHull($\mathcal{S}^{(m)}$), and the weights associated to this convex combinations correspond to the probability of observing each channel state, i.e. ψ_m. The second part of the lemma shows that, given the channel state, the feasible service vector selected by the MaxWeight algorithm achieves the maximum $\boldsymbol{c}^{(\ell)}$ weighted service rate, $b^{(m,\ell)}$, with high probability. These results are important because in the proof of Theorem 2.1 we work with $\overline{\boldsymbol{q}}_{\parallel\mathcal{H}}$, and by definition, $\overline{\boldsymbol{q}}_{\parallel\mathcal{H}}$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

is a linear combination of the vectors $\boldsymbol{c}^{(\ell)}$ with $\ell \in P$. These two results imply that, intuitively, the expected vector of potential service rate behaves as if it belonged to the capacity region.

Proof of Claim 2.1. Conditioning on the channel state, we get

$$\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}\rangle\big] \overset{(a)}{=} \sum_{m\in\mathcal{M}} \psi_m \mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big],$$

where $\boldsymbol{v}^{(m)}$ is defined as in Lemma 2.1 and (a) follows from Lemma 2.1 part (i), and because $\overline{\boldsymbol{q}}_{\parallel\mathcal{H}}$ is a linear combination of the vectors $\boldsymbol{c}^{(\ell)}$ with $\boldsymbol{c}^{(\ell)} \in \tilde{P}$. It remains to show that $\mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]$ is $\mathrm{O}(\epsilon \log(1/\epsilon))$.

Observe that $\overline{\boldsymbol{q}} = \overline{\boldsymbol{q}}_{\parallel\mathcal{H}} + \overline{\boldsymbol{q}}_{\perp\mathcal{H}} = \overline{\boldsymbol{q}}_{\parallel\mathcal{K}} + \overline{\boldsymbol{q}}_{\perp\mathcal{K}}$, and thus

$$\begin{aligned}
&\mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \\
&\quad = \mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{K}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \qquad (2.9)\\
&\qquad + \mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\perp\mathcal{K}} - \overline{\boldsymbol{q}}_{\perp\mathcal{H}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]. \qquad (2.10)
\end{aligned}$$

Now we show that the terms in (2.9) and (2.10) are $\mathrm{O}(\epsilon \log(1/\epsilon))$. From (2.9), we have

$$\mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{K}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \leq 0$$

by the definition of projection on the cone $\mathcal{K}$ and by definition of $\boldsymbol{v}^{(m)}$ and $b^{(m,\ell)}$ in Lemma 2.1 part (i). Now we have

$$\begin{aligned}
0 &\geq \mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\parallel\mathcal{K}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \\
&\overset{(a)}{\geq} -\mathbb{E}_m\big[\langle \overline{\boldsymbol{q}}_{\perp\mathcal{K}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \\
&\overset{(b)}{\geq} -\mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp\mathcal{K}}\|^r\big]^{1/r} \mathbb{E}_m\big[\|\bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\|^p \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]^{1/p} \\
&\overset{(c)}{\geq} -R_r^{1/r}\, \mathbb{E}_m\big[\|\bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\|^p \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]^{1/p},
\end{aligned}$$

where (a) holds because $\overline{\boldsymbol{q}}_{\parallel\mathcal{K}} = \overline{\boldsymbol{q}} - \overline{\boldsymbol{q}}_{\perp\mathcal{K}}$, and because $\langle \overline{\boldsymbol{q}}, \bar{\boldsymbol{s}} - \boldsymbol{v}^{(m)}\rangle \geq 0$ by the definition of MaxWeight in (2.3) and since $\boldsymbol{v}^{(m)} \in \mathcal{S}^{(m)}$, (b) holds using Hölder's inequality for some $p, r > 1$ such that $1/p + 1/r = 1$, and (c) holds by SSC in Proposition 2.1. Now, by the definition of R_r, we have

$$R_r^{1/r} = \left(\left(\frac{8n\alpha^2}{\delta}\right)^r + \left(8\sqrt{n}\alpha\right)^r \left(\frac{8\sqrt{n}\alpha + \delta}{\delta}\right)^r r!\right)^{1/r} \leq \beta_5 (r!)^{1/r} \overset{(a)}{\leq} \beta_5 \mathrm{e}^{1/r-1} r^{1+1/2r},$$

where

$$\beta_5 \triangleq \frac{8n\alpha^2}{\delta} + 8\sqrt{n}\alpha\left(\frac{8\sqrt{n}\alpha + \delta}{\delta}\right).$$

Here (a) follows from Stirling's approximation for the factorial.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Now we bound the remaining term $\mathbb{E}_m\big[\|\bar{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\|^p \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]^{1/p}$ as follows:

$$\begin{aligned}
0 &\le \mathbb{E}_m\big[\|\bar{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\|^p \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big]^{1/p} \\
&\overset{(a)}{=} \mathbb{E}_m\big[\|\bar{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\|^p \mid \langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\big]^{1/p} \big(1 - \pi^{(m,\ell)}\big)^{1/p} \\
&\overset{(b)}{\le} n\big(S_{\max} + V_{\max}\big)\big(1 - \pi^{(m,\ell)}\big)^{1/p} \overset{(c)}{=} \beta_6 \epsilon^{1/p},
\end{aligned} \tag{2.11}$$

where (a) holds by definition of $\pi^{(m,\ell)}$ in Lemma 2.1 part (ii), (b) holds with $V_{\max} = \max_{m\in\mathcal{M}, i\in[n]} \nu_i^{(m)}$, and (c) holds by Lemma 2.1 part (ii) for

$$\beta_6 \triangleq n\big(S_{\max} + V_{\max}\big) \max \frac{b^{(m,\ell)}}{\gamma^{(m)}},\, 1.$$

Putting everything together, we obtain

$$\begin{aligned}
0 &\ge \mathbb{E}_m\big[\langle \bar{\boldsymbol{q}}_{\|\mathcal{K}}, \bar{\boldsymbol{s}} - \boldsymbol{\nu}^{(m)}\rangle \mathbb{1}_{\{\langle \boldsymbol{c}^{(\ell)}, \bar{\boldsymbol{s}}\rangle \neq b^{(m,\ell)}\}}\big] \\
&\ge -\beta_5\beta_6 \mathrm{e}^{1/r-1} r^{1+1/2r} \epsilon^{1/p} \\
&\overset{(a)}{=} -\beta_5\beta_6 \mathrm{e}^{\frac{1}{\lfloor \log(1/\epsilon)\rfloor}-1} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{1+1/2\lfloor \log(1/\epsilon)\rfloor} \epsilon^{-\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \epsilon \\
&\overset{(b)}{\ge} -2\beta_5\beta_6 \epsilon \log\left(\frac{1}{\epsilon}\right) \quad \text{for all } \epsilon < \epsilon_0',
\end{aligned}$$

where (a) holds after choosing $r \triangleq \lfloor \log(1/\epsilon)\rfloor$, and (b) follows for ϵ_0' as defined below, and because by the definition of floor function we have

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0} \mathrm{e}^{\frac{1}{\lfloor \log(1/\epsilon)\rfloor}-1} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\lfloor \log(1/\epsilon)\rfloor}} \epsilon^{-\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \\
&\quad \le \lim_{\epsilon\downarrow 0} \mathrm{e}^{\frac{1}{\log(1/\epsilon)-1}-1} \lim_{\epsilon\downarrow 0} \log\left(\frac{1}{\epsilon}\right)^{\frac{1}{2\log(1/\epsilon)-2}} \lim_{\epsilon\downarrow 0} \epsilon^{-\frac{1}{\log(1/\epsilon)}} = \frac{1}{e} \times 1 \times e = 1.
\end{aligned}$$

By definition of limit, there exists $\epsilon_0' > 0$ such that for all $\epsilon < \epsilon_0'$ we have

$$\mathrm{e}^{\frac{1}{\lfloor \log(1/\epsilon)\rfloor}-1} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\lfloor \log(1/\epsilon)\rfloor}} \epsilon^{-\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \le 2.$$

The proof that the term (2.10) is $\mathrm{O}(\epsilon \log(1/\epsilon))$ follows similarly by linearity of dot product, Hölder's inequality with $r = \lfloor \log(1/\epsilon)\rfloor$ and (2.11). We omit the details for brevity. □

We end this section with the proof of Claim 2.2.

Proof of Claim 2.2. In this proof we use ideas and notation from [2, equation (56)]. For each $\ell \in P$, let $\mathcal{L}_+^{(\ell)} \triangleq \big\{i \in [n] \colon c_i^{(\ell)} > 0\big\}$ and define

$$\widetilde{\boldsymbol{c}}^{(\ell)} = \big[c_i^{(\ell)}\big]_{i\in\mathcal{L}_+^{(\ell)}},\ \widetilde{\bar{\boldsymbol{q}}}^{(\ell)} = \big[\bar{q}_i\big]_{i\in\mathcal{L}_+^{(\ell)}} \quad \text{and} \quad \widetilde{\bar{\boldsymbol{u}}}^{(\ell)} = \big[\bar{u}_i\big]_{i\in\mathcal{L}_+^{(\ell)}}.$$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Then

$$0 \le \left|\frac{\mathcal{T}_4}{2}\right| = \left|\mathbb{E}\left[\langle \overline{\boldsymbol{q}}^+_{\parallel\mathcal{H}},\, \overline{\boldsymbol{u}}_{\parallel\mathcal{H}}\rangle\right]\right| \overset{(a)}{=} \left|\mathbb{E}\left[-\langle (\widetilde{\overline{\boldsymbol{q}}}^{(\ell)}_{\perp\mathcal{H}})^+,\, \widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\rangle\right]\right| \overset{(b)}{\le} \mathbb{E}\left[\left\|(\widetilde{\overline{\boldsymbol{q}}}^{(\ell)}_{\perp\mathcal{H}})^+\right\|^r\right]^{1/r} \mathbb{E}\left[\left\|\widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\right\|^p\right]^{1/p},$$

where (a) follows using the definition of projection on the subspace to substitute

$$\overline{\boldsymbol{q}}^+_{\parallel\mathcal{H}} = \sum_{\ell\in P} \langle \boldsymbol{c}^{(\ell)}, \overline{\boldsymbol{q}}^+\rangle \boldsymbol{c}^{(\ell)},$$

then the key property (2.2), and that

$$(\widetilde{\overline{\boldsymbol{q}}}^{(\ell)})^+ = (\widetilde{\overline{\boldsymbol{q}}}^{(\ell)}_{\parallel\mathcal{H}})^+ + (\widetilde{\overline{\boldsymbol{q}}}^{(\ell)}_{\perp\mathcal{H}})^+.$$

Then (b) holds by Hölder's inequality with $p, r > 1$ integers such that $1/r + 1/p = 1$.

Now we bound each of the terms. For the first term we use Proposition 2.1, and we obtain

$$\mathbb{E}\left[\left\|(\widetilde{\overline{\boldsymbol{q}}}^{(\ell)}_{\perp\mathcal{H}})^+\right\|^r\right]^{1/r} \le \mathbb{E}\left[\left\|\overline{\boldsymbol{q}}^+_{\perp\mathcal{H}}\right\|^r\right]^{1/r} \le R_r^{1/r} \overset{(a)}{\le} \beta_5 \mathrm{e}^{1/r-1} r^{1+1/2r},$$

where (a) holds by Stirling's approximation for the factorial. For the second term we obtain

$$0 \le \mathbb{E}\left[\left\|\widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\right\|^p\right] \overset{(a)}{\le} \sum_{\ell\in P} \sum_{i\in\mathcal{L}^{(\ell)}_+} \frac{\tilde{c}^{(\ell)}_i}{\tilde{c}^{(\ell)}_i} \mathbb{E}\left[\tilde{u}_i^p\right] \overset{(b)}{\le} \frac{S_{\max}^{p-1}}{\widetilde{c}_{\min}} \sum_{\ell\in P} \mathbb{E}\left[\langle \widetilde{\boldsymbol{c}}^{(\ell)}, \widetilde{\overline{\boldsymbol{u}}}^{(\ell)}\rangle\right] \overset{(c)}{\le} \beta_7 |P| \frac{S_{\max}^{p-1}}{\widetilde{c}_{\min}} \epsilon,$$

where (a) follows as all the terms in the summation are non-negative, (b) holds by defining $\widetilde{c}_{\min} = \min_{\ell\in P, i\in[n]}\{\tilde{c}^{(\ell)}_i\}$ and by definition of dot product, and (c) follows from [4, equation (43)] for a finite constant β_7, using a similar argument to the properties used to obtain (2.8). Now pick $r \overset{\triangle}{=} \lfloor \log(1/\epsilon)\rfloor$ to get

$$\begin{aligned}
0 &\le \left|\frac{\mathcal{T}_4}{2}\right| \\
&\le \beta_5 \beta_7^{1/p} \frac{S_{\max}^{1-1/p}}{\widetilde{c}_{\min}^{1/p}} |P|^{1/p} \mathrm{e}^{1/r-1} r^{1+1/2r} \epsilon^{1/p} \\
&= \beta_5 \beta_7^{1/p} S_{\max}^{\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \left(\frac{|P|}{\widetilde{c}_{\min}}\right)^{1-\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \mathrm{e}^{\frac{1}{\lfloor\log(1/\epsilon)\rfloor}-1} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{1+\frac{1}{2\lfloor\log(1/\epsilon)\rfloor}} \epsilon^{-\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \epsilon \\
&\overset{(a)}{\le} 2\beta_5 \max\left\{\beta_7,\, 1\right\} \frac{|P|}{\widetilde{c}_{\min}} \epsilon \log\left(\frac{1}{\epsilon}\right) \quad \text{for all } \epsilon < \epsilon_0'',
\end{aligned}$$

where (a) follows as

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0} S_{\max}^{\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \left(\frac{|P|}{e\widetilde{c}_{\min}}\right)^{1-\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\lfloor\log(1/\epsilon)\rfloor}} \epsilon^{-\frac{1}{\lfloor\log(1/\epsilon)\rfloor}} \\
&\qquad \le 1 \times \frac{|P|}{e\widetilde{c}_{\min}} \times 1 \times e = \frac{|P|}{\widetilde{c}_{\min}}.
\end{aligned}$$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Thus there exists $\epsilon_0'' > 0$ such that for all $\epsilon < \epsilon_0''$ we have

$$S_{\max}^{\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \left(\frac{|P|}{e\widetilde{c}_{\min}}\right)^{1-\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \left\lfloor \log\left(\frac{1}{\epsilon}\right)\right\rfloor^{\frac{1}{2\lfloor \log(1/\epsilon)\rfloor}} \epsilon^{-\frac{1}{\lfloor \log(1/\epsilon)\rfloor}} \leq \frac{2|P|}{\widetilde{c}_{\min}}. \qquad \square$$

The key idea in obtaining a logarithmic error bound is in picking the right exponent r in Hölder's inequality while bounding terms $\mathcal{T}_1$ and $\mathcal{T}_4$. We do this by minimizing the upper bound over r (for a fixed ϵ), which gives $r = \lfloor \log(1/\epsilon)\rfloor$. The idea of optimizing over the exponent in Hölder's inequality is motivated by the paper [1].

3. Logarithmic error bounds in the load balancing system

While the generalized switch models many different SPNs with control on the service process, there are many systems where the control is on the arrivals, such as the load balancing system. In this section we show that a similar methodology to the proof of Theorem 2.1 can be used in load balancing systems. Specifically, we study a load balancing system operating under 'join the shortest queue' (JSQ) as an illustrative example. Similar results can be obtained for other routeing algorithms such as power-of-d choices. We first define the model.

3.1. Load balancing model

Consider an SPN with n queues, each of them with a separate server. Arrivals occur in a single stream, and a dispatcher routes them according to JSQ (i.e. to the server with the smallest number of jobs in line). After routeing, jobs cannot commute lines. We model the system in discrete time, and we track the number of jobs in each queue. Then the service policy is irrelevant. In each time slot, all the arrivals are routed to the same queue.

Let $\{a(k)\colon k \in \mathbb{Z}_+\}$ be the arrival process to the system, which is a sequence of i.i.d. random variables, and let $\boldsymbol{a}(k)$ be the vector of arrivals to the queues after routeing at time k. By definition of JSQ, we have

$$i^* \in \arg\min_{i\in[n]}\{q_i(k)\}, \quad \boldsymbol{a}(k) = a(k)\boldsymbol{e}^{(i^*)}.$$

If there are multiple minimizers, any can be chosen at random. Note that $a(k) = \sum_{i=1}^n a_i(k)$ by definition. The potential service is a sequence of i.i.d. random vectors, which we denote by $\{\boldsymbol{s}(k)\colon k \in \mathbb{Z}_+\}$, and it is independent of the arrival and queue length processes. We assume there exist finite constants $A_{\max}$ and $S_{\max}$ such that $a(1) \leq A_{\max}$ and $s_i(1) \leq S_{\max}$ for all $i \in [n]$ with probability 1. We use $\boldsymbol{u}(k)$ to denote the unused service vector in time slot k, which is defined similarly to the generalized switch model. The dynamics of the queues occur according to (2.1), and (2.2) is satisfied for all $i \in [n]$.

Let $\boldsymbol{\mu} \triangleq \mathbb{E}[\boldsymbol{s}(1)]$, $\sigma_{s_i}^2 \triangleq \mathrm{Var}[s_i(1)]$ for each $i \in [n]$, and let $\mu_\Sigma \triangleq \sum_{i=1}^n \mu_i$. We assume $\mu_i > 0$ for all $i \in [n]$ because otherwise the jobs routed to the server with zero service rate will never be processed. It is well known that the capacity region of the load balancing model is $\mathcal{C} = \{\lambda \in \mathbb{R}_+\colon \lambda \leq \mu_\Sigma\}$, and that JSQ is a throughput optimal routeing algorithm [2, Lemma 2]. Then, to model heavy traffic, we parametrize the arrival process by $\epsilon \in (0, \mu_\Sigma)$, letting $\lambda^{(\epsilon)} \triangleq \mathbb{E}\big[a^{(\epsilon)}(1)\big] = \mu_\Sigma - \epsilon$ and $\big(\sigma_a^{(\epsilon)}\big)^2 \triangleq \mathrm{Var}\big[a^{(\epsilon)}(1)\big]$.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

It is also known that SSC occurs in the line where all the queues are equally long. Specifically, denoting

$$\overline{\boldsymbol{q}}_{\parallel}^{(\epsilon)} \triangleq \left(\frac{1}{n}\sum_{i=1}^{n}\overline{q}_i^{(\epsilon)}\right)\mathbf{1} \quad \text{and} \quad \overline{\boldsymbol{q}}_{\perp}^{(\epsilon)} \triangleq \overline{\boldsymbol{q}}^{(\epsilon)} - \overline{\boldsymbol{q}}_{\parallel}^{(\epsilon)},$$

we have that $\mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^r\big]$ is bounded for all $r \geq 1$, as proved in [2, Proposition 1]. Recall that we add a bar on top of the variables to denote steady state.

3.2. Logarithmic error bounds

The goal of this section is to prove the following result.

Theorem 3.1. *Consider a set of load balancing systems operating under JSQ, parametrized by the heavy traffic parameter $\epsilon \in (0, \mu_\Sigma)$, as described above. Then there exists a constant β_{JSQ} and $\epsilon_0 \in (0, \mu_\Sigma)$ such that for all $\epsilon < \epsilon_0$*

$$\left|\mathbb{E}\left[\sum_{i=1}^{n}\overline{q}_i^{(\epsilon)}\right] - \frac{1}{2\epsilon}\left(\big(\sigma_a^{(\epsilon)}\big)^2 + \sum_{i=1}^{n}\sigma_{s_i}^2\right)\right| \leq \beta_{\mathrm{JSQ}}\log\left(\frac{1}{\epsilon}\right).$$

Similarly to the generalized switch, an essential step in the proof of Theorem 3.1 is to find explicit upper bounds for the moments of $\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|$. We present them in the next proposition.

Proposition 3.1. *For the load balancing system operating under JSQ, parametrized by $\epsilon \in (0, \mu_\Sigma)$ as described in Section 3.1, let*

$$\mu_{\min} = \min_{i\in[n]}\mu_i, \qquad \delta \in (0, \mu_{\min}), \qquad \alpha_{\mathrm{JSQ}} \triangleq \max\{A_{\max}, S_{\max}\}.$$

Then, for any choice of $\epsilon \in (0, (\mu_{\min} - \delta)n)$, and all $r = 1, 2, \ldots$, we have

$$\mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp}^{(\epsilon)}\|^r\big] \leq R_r^{(\mathrm{JSQ})} \triangleq \left(\frac{6n\alpha_{\mathrm{JSQ}}^2}{\delta}\right)^r + \big(8\alpha_{\mathrm{JSQ}}\sqrt{n}\big)^r\left(\frac{4\alpha_{\mathrm{JSQ}}+\delta}{\delta}\right)^r r!$$

The proof of Proposition 3.1 is very similar to the proof of [2, Proposition 1], so we omit it.

The proof of Theorem 3.1 follows from the computation of the upper bound in [2], similarly to the proof of Theorem 2.1. We include a sketch proof for completeness.

Proof of Theorem 3.1. In this proof we omit the dependence on ϵ of the variables, for ease of exposition. We set the drift of $V_{\parallel}(\boldsymbol{q}) \triangleq \|\boldsymbol{q}_{\parallel}\|^2$ to zero. First observe that in the computation of (2.5) we only use properties of projection and norm, and we did not use properties of the generalized switch itself. Therefore the same steps can be followed for the load balancing system. We obtain

$$\begin{aligned} 0 &= \mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\parallel}^{+}\|^2 - \|\overline{\boldsymbol{q}}_{\parallel}\|^2\big] \\ &= \underbrace{\mathbb{E}\big[\|\overline{\boldsymbol{a}}_{\parallel} - \bar{\boldsymbol{s}}_{\parallel}\|^2\big]}_{\mathcal{T}_2} + \underbrace{2\mathbb{E}\big[\langle\overline{\boldsymbol{q}}_{\parallel},\ \overline{\boldsymbol{a}}_{\parallel} - \bar{\boldsymbol{s}}_{\parallel}\rangle\big]}_{-\mathcal{T}_1} - \underbrace{\mathbb{E}\big[\|\overline{\boldsymbol{u}}_{\parallel}\|^2\big]}_{\mathcal{T}_3} + \underbrace{2\mathbb{E}\big[\langle\overline{\boldsymbol{q}}_{\parallel}^{+},\ \overline{\boldsymbol{u}}_{\parallel}\rangle\big]}_{\mathcal{T}_4}. \end{aligned}$$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

We analyze term by term. For $\mathcal{T}_1$ we obtain

$$\frac{\mathcal{T}_1}{2} = \mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\parallel}, \bar{\boldsymbol{s}}_{\parallel} - \overline{\boldsymbol{a}}_{\parallel}\rangle\big] \overset{(a)}{=} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} \overline{q}_i\right]\mathbb{E}\left[\sum_{i=1}^{n} \bar{s}_i - \overline{a}\right] \overset{(b)}{=} \frac{\epsilon}{n}\mathbb{E}\left[\sum_{i=1}^{n} \overline{q}_i\right], \tag{3.1}$$

where (a) follows by definition of the projection, because the total arrivals and the potential service are independent of the queue lengths, and rearranging terms, and (b) holds by definition of ϵ. For $\mathcal{T}_2$ we obtain

$$\begin{aligned}\mathcal{T}_2 = \mathbb{E}\big[\|\overline{\boldsymbol{a}}_{\parallel} - \bar{\boldsymbol{s}}_{\parallel}\|^2\big] &\overset{(a)}{=} \frac{1}{n}\left(\mathbb{E}\big[\overline{a}^2\big] + \mathbb{E}\left[\left(\sum_{i=1}^{n} \bar{s}_i\right)^2\right] - 2\mathbb{E}[\overline{a}]\mathbb{E}\left[\sum_{i=1}^{n} \bar{s}_i\right]\right) \\ &\overset{(b)}{=} \frac{1}{n}\left(\big(\sigma_a^{(\epsilon)}\big)^2 + \sum_{i=1}^{n} \sigma_{s_i}^2 + \epsilon^2\right),\end{aligned} \tag{3.2}$$

where (a) holds by definition of the projection and Euclidean norm, and (b) holds by the definition of variance and of ϵ. For $\mathcal{T}_3$ we obtain

$$|\mathcal{T}_3| = \mathbb{E}\big[\|\overline{\boldsymbol{u}}_{\parallel}\|^2\big] = \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n} \overline{u}_i\right)^2\right] \overset{(a)}{\leq} S_{\max}\mathbb{E}\left[\sum_{i=1}^{n} \overline{u}_i\right] \overset{(b)}{=} S_{\max}\epsilon, \tag{3.3}$$

where (a) holds because, by definition of unused service, we have $\overline{u}_i \leq \bar{s}_i \leq S_{\max}$ with probability 1 for all $i \in [n]$, and (b) holds because $\mathbb{E}\big[\sum_{i=1}^{n} \overline{u}_i\big] = \epsilon$, which can be easily proved by setting the drift of the function $V_\ell(\boldsymbol{q}) = \sum_{i=1}^{n} q_i$ to zero. A proof of this fact can be found in [5, Lemma 5].

For the term $\mathcal{T}_4$ we follow a similar approach to the proof of Theorem 2.1. We obtain

$$\begin{aligned}\left|\frac{\mathcal{T}_4}{2}\right| = \mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\parallel}^{+}, \overline{\boldsymbol{u}}_{\parallel}\rangle\big] &\overset{(a)}{=} -\mathbb{E}\big[\langle \overline{\boldsymbol{q}}_{\perp}^{+}, \overline{\boldsymbol{u}}\rangle\big] \\ &\overset{(b)}{\leq} \mathbb{E}\big[\|\overline{\boldsymbol{q}}_{\perp}^{+}\|^r\big]^{1/r}\mathbb{E}\big[\|\overline{\boldsymbol{u}}\|^p\big]^{1/p} \\ &\overset{(c)}{\leq} \big(R_r^{(\mathrm{JSQ})}\big)^{1/r} S_{\max}^{1/r}\epsilon^{1-1/r},\end{aligned} \tag{3.4}$$

where (a) holds by definition of the projection, and reorganizing terms, (b) holds by Hölder's inequality with $p, r > 1$ integers such that $1/r + 1/p = 1$, and (c) holds by Proposition 3.1, because $\overline{\boldsymbol{u}} \leq \bar{\boldsymbol{s}} \leq S_{\max}\mathbf{1}$ component-wise with probability 1, and because $\mathbb{E}\big[\sum_{i=1}^{n} \overline{u}_i\big] = \epsilon$.

Then, putting (3.1), (3.2), (3.3), and (3.4) together, and letting $r = \lceil \log(1/\epsilon) \rceil$, we obtain the result. □

4. Conclusions

In this paper we study the performance of generalized switch operating under MaxWeight both when the CRP condition is satisfied and when it is not. We show that MaxWeight is within $\mathrm{O}(\log(1/\epsilon))$ from its heavy traffic performance. Additionally, when the CRP condition is satisfied, we show that it is within $\mathrm{O}(\log(1/\epsilon))$ from the optimal policy.

We also analyze the load balancing system operating under JSQ and prove that the rate of convergence of JSQ to the optimal heavy traffic performance under heavy traffic is $\mathrm{O}(\log(1/\epsilon))$. Similar results can be obtained for other routeing algorithms.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

A possible line of future work is to explore if the $O(\log(1/\epsilon))$ error is tight. At this point it is not known if the log error is an artifact of our proof or if there is indeed a log error in the heavy traffic prelimit.

Appendix A. Proof of Corollary 2.1

Proof. In this case we have $\mathcal{K} = \{\xi \boldsymbol{c}^{(\ell)} : \xi \geq 0\}$, i.e. the cone $\mathcal{K}$ is a half-line. This implies that $\mathcal{H}$ is the entire line defined by $\boldsymbol{c}^{(\ell)}$. Then in Theorem 2.1 we can take $\boldsymbol{w} = b^{(\ell)}\boldsymbol{c}^{(\ell)}$, and we have $H = \boldsymbol{c}^{(\ell)}(\boldsymbol{c}^{(\ell)})^\top$ because we assumed $\|\boldsymbol{c}^{(\ell)}\| = 1$. Then we obtain

$$\mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)}\rangle\big] \leq \frac{1}{2\epsilon b^{(\ell)}}\big((\boldsymbol{c}^{(\ell)})^\top \Sigma_a^{(\epsilon)} \boldsymbol{c}^{(\ell)} + \sigma_{B_\ell}^2\big) + \beta \log\left(\frac{1}{\epsilon}\right),$$

where $\sigma_{B_\ell}^2 \triangleq (\Sigma_B)_{\ell,\ell}$ and β is a constant that does not depend on ϵ.

To obtain a ULB, we use [4, Proposition 1]. We obtain that, under any scheduling algorithm (not necessarily MaxWeight), the expected queue length vector in the steady state satisfies

$$\mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)}\rangle\big] \geq \frac{1}{2\epsilon b^{(\ell)}}\big((\boldsymbol{c}^{(\ell)})^\top \Sigma_a^{(\epsilon)} \boldsymbol{c}^{(\ell)} + \sigma_{B_\ell}^2\big) - \frac{b_{\max}(1-\epsilon)}{2} \triangleq \mathrm{ULB}, \tag{A.1}$$

where $b_{\max} \triangleq \max_{m\in\mathcal{M},\ell\in[L]}\{b^{(m,\ell)}\}$. Then, putting the two results together, we obtain

$$\begin{aligned}\mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)}\rangle\big] &\leq \mathrm{ULB} + \left(\frac{b_{\max}(1-\epsilon)}{2} + \beta\log\left(\frac{1}{\epsilon}\right)\right)\\ &\overset{(a)}{\leq} \mathrm{ULB} + \left(\frac{b_{\max}}{2} + \beta\right)\log\left(\frac{1}{\epsilon}\right),\end{aligned}$$

where (a) holds because $1 - x \leq \log(1/x)$ for all $x > 0$. Therefore, defining

$$\tilde{\beta} \triangleq \frac{b_{\max}}{2} + \beta \tag{A.2}$$

and using (A.1), we obtain

$$\mathrm{ULB} \leq \mathbb{E}\big[\langle \overline{\boldsymbol{q}}^{(\epsilon)}, \boldsymbol{c}^{(\ell)}\rangle\big] \leq \mathrm{ULB} + \tilde{\beta}\log\left(\frac{1}{\epsilon}\right). \qquad \square$$

Appendix B. Proof of Proposition 2.1

Proof of Proposition 2.1. The proof is similar to [4, Proposition 2], so we only present a sketch. The first inequality holds because $\mathcal{K} \subset \mathcal{H}$ and by definition of $\overline{\boldsymbol{q}}_{\perp\mathcal{K}}$ and $\overline{\boldsymbol{q}}_{\perp\mathcal{H}}$. Now we bound the second inequality. From the definition of projection, the triangle inequality and the queue dynamics presented in (2.2), we obtain

$$|\|\boldsymbol{q}_{\perp\mathcal{K}}(k+1)\| - \|\boldsymbol{q}_{\perp\mathcal{K}}(k)\||\mathbb{1}_{\{\boldsymbol{q}(k)=\boldsymbol{q}\}} \leq 2\sqrt{n}\alpha,$$

with probability 1, and that for all $\boldsymbol{q}$ such that $\|\boldsymbol{q}_{\perp\mathcal{K}}\| \geq (4n\alpha)/\delta$ we have

$$\mathbb{E}\big[\|\boldsymbol{q}_{\perp\mathcal{K}}(k+1)\| - \|\boldsymbol{q}_{\perp\mathcal{K}}(k)\| \mid \boldsymbol{q}(k) = \boldsymbol{q}\big] \leq -\frac{\delta}{4}.$$

Using these results in [7, Lemma 3], we obtain the result. $\square$

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press

Funding information

This work was partially supported by the National Science Foundation grants CCF-1850439 and EPCN-2144316. Daniela Hurtado-Lange has partial funding from the Chilean National Agency for Research and Development ANID/DOCTORADO BECAS CHILE/2018-72190413.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] CHEN, Z., MAGULURI, S. T., SHAKKOTTAI, S. AND SHANMUGAM, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *In Advances in Neural Information Processing Systems* **33**, pp. 8223–8234. Curran Associates.

[2] ERYILMAZ, A. AND SRIKANT, R. (2012). Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* **72**, 311–359.

[3] HAJEK, B. (2015). *Random Processes for Engineers*. Cambridge University Press.

[4] HURTADO-LANGE, D. A. AND MAGULURI, S. T. (2022). Heavy-traffic analysis of queueing systems with no complete resource pooling. To appear in *Math. Operat. Res.*

[5] HURTADO-LANGE, D. AND MAGULURI, S. T. (2020). Transform methods for heavy-traffic analysis. *Stochastic Systems* **10**, 275–390.

[6] LU, Y., MAGULURI, S. T., SQUILLANTE, M., SUK, T. AND WU, X. (2018). An optimal scheduling policy for the 2 x 2 input-queued switch with symmetric arrival rates. *ACM SIGMETRICS Performance Evaluation Rev.* **45**, 217–223.

[7] MAGULURI, S. T. AND SRIKANT, R. (2016). Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stoch. Syst.* **6**, 211–250.

[8] MEYN, S. (2009). Stability and asymptotic optimality of generalized maxweight policies. *SIAM J. Control Optimization* **47**, 3259–3294.

[9] SHARIFNASSAB, A., TSITSIKLIS, J. N. AND GOLESTANI, S. J. (2020). Fluctuation bounds for the max-weight policy with applications to state space collapse. *Stochastic Systems.* **10**, 193–273.

[10] SINGH, R. AND STOLYAR, A. (2015). MaxWeight scheduling: Asymptotic behavior of unscaled queue-differentials in heavy traffic. Available at arXiv:1502.03793.

[11] STOLYAR, A. (2004). MaxWeight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Prob.* **14**, 1–53.

[12] WILLIAMS, R. (2016). Stochastic processing networks. *Ann. Rev. Statist. Appl.* **3**, 323–345.

https://doi.org/10.1017/jpr.2021.82 Published online by Cambridge University Press