52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

On Optimal Routing in Overloaded Parallel Queues

Bin Li, Atilla Eryilmaz, R. Srikant, and Leandros Tassiulas

Abstract— **We consider the problem of routing Bernoulli arrivals to parallel queues, where each queue provides service according to an independent Bernoulli process. We assume that the total arrival rate exceeds the sum of the service rates of the queues. Since such a queueing system is unstable, the vector of queue lengths does not have a well-defined stationary distribution. However, one metric which can be used to compare routing policies is the amount of unused service in the system. To lower-bound the cumulative unused service in the system, we present a "queue reversal" theorem for a single-server queue with independent and identically distributed (i.i.d.) arrivals and i.i.d. services: assuming that the queue is initially empty, the expected cumulative unused service is equal to the expected queue length in a queue where the arrivals and services are reversed. Thus, the expected cumulative unused service in the unstable system is equal to the expected queue length in a stable system, which can be calculated. Using this result for a single-server queue, we obtain a lower bound on the expected unused service in the parallel queueing system for any feasible routing policy. We then compare this lower bound to the performance of two simple routing policies: Randomized and Join-the-Shortest Queue Routing.**

I. Introduction

The classical problem of routing random arrivals to parallel queues with random services has received a lot of research interest (e.g., [1], [2], [3]). So far, we have a better understanding when the system is *under-loaded*, that is, when the arrival rate is less than the sum of the service rates. In under-loaded conditions, Join-the-Shortest-Queue (JSQ) policy that forwards all arrivals to a queue with the minimal queue-length, and the Randomized-Routing (RR) policy that routes all arrivals to a queue with a probability proportional to its service rate have been shown to stabilize the queueing system (i.e., they keep the mean queue-lengths finite). In fact, the JSQ policy can be regarded as a special case of the more general class of backpressure-based Max-Weight (MW) policies studied in [4]. Moreover, it minimizes the total mean queue-length in the *heavy-traffic regime* ([5], [2]), where the arrival rate approaches to the total sum of service rates from below.

In contrast to the rich and tight results on the performance of state-based policies (including JSQ routing and MW scheduling) in under-loaded conditions, their performance in overloaded regimes (i.e., when the arrival rate is greater than the total sum of service rates) is less understood. There are several interesting works (e.g., [6], [7], [8], [9]) dealing with over-loaded queueing systems. Under various conditions, these works study the performance and optimization of the metrics of *queue overflow rates* (i.e., the rates at which the queues grow in the overloaded regime), and the related metric of *the departure rates* of served packets. For example, in [7], [8], the authors focus on the design and analysis of MW type scheduling policies to minimize the queue overflow rates or to maximize the total departure rate from the system.

One caveat with the performance metrics of overflow rate and departure rate is that, being long-term time averages, they may not be able to differentiate between policies in terms of their *convergence rates* to the same limit. In fact, as will be noted in Section II, even suitably selected randomized decisions can achieve optimal overflow or departure rate levels in overloaded queueing systems. With this motivation, in this paper we propose and analyze the metric of *cumulative unused service* over time to analyze the performance of routing policies in overloaded systems. This metric not only measures the amount of *under-utilization* in the multi-server system over time, but also captures the speed with which the running-average of the departure rates converges to their limiting value (cf. Section II).

The proposed cumulative unused service process is difficult to analyze in both stable and unstable systems due to its non-stationary nature. To tackle this challenge, we establish a novel "queue reversal" result (cf. Theorem 1) that equates the expected cumulative unused service in the unstable system to the expected queue-length of a related (in fact, reversed) stable system. With this connection, we can obtain the mean cumulative unused service metric by studying the mean queue-length of a stable Markov chain, for which a rich set of tools and results exists. Based on this fundamental result, we are able to obtain a nontrivial lower bound (cf. Section III) on the expected cumulative unused service for any feasible routing policy serving N parallel queues under overloaded conditions.

This lower bound motivates us to study the performance of two well-known policies, namely Randomized Routing (RR) and Join-the-Shortest Queue (JSQ) policies, with respect to this fundamental limit (cf. Section IV). It is easy to observe that both RR and JSQ policies are departure-rate-optimal, in that, they both achieve the maximum total departure rate. After utilizing the queue-reversal theorem once again, we establish tight upper and lower bounds on the total mean cumulative unused service under the RR policy (cf. Proposition 2). This result reveals that the cumulative unused service performance under the RR policy deviates significantly from

B. Li (celibin@gmail.com) and A. Eryilmaz (eryilmaz.2@osu.edu) are with the ECE Department at the Ohio State University, USA; R. Srikant (rsrikant@illinois.edu) is with the ECE Department at the University of Illinois at Urbana-Champaign, USA; and L. Tassiulas (leandros@inf.uth.gr) is with the Computer Engineering and Telecommunications Department of the University of Thessaly, Greece.

This work is partly supported by QNRF Grant NPRP 09-1168-2-455, NSF grant CAREER-CNS-0953515, AFOSR MURI FA 9550-10-1-0573, and ARO MURI W911NF-12-1-0385.

the lower-bound, and suggests that the RR policy is sub-optimal. We also note that the JSQ policy minimizes the total cumulative unused service over all feasible policies under symmetric servers through a standard path-coupling argument (cf. Proposition 3).

To compare the unused service performance of the JSQ policy to the lower bound and the RR policy, we then perform numerical studies for both symmetric and asymmetric conditions (cf. Section V). These investigations show that for both conditions, the expected cumulative unused service performance of the JSQ policy approaches the lower bound, suggesting its optimality both in the scaling of the network size and the *critically overloaded regime* (where the total service rate approaches the arrival rate from above). Moreover, these numerical results demonstrate that the lower bound we derive through the queue-reversal theorem is indeed tight in these two scaling regimes.

II. System Model

We consider the discrete-time parallel queueing system depicted in Figure 1. Packets arrive according to an i.i.d. Bernoulli process[1] $\{A[t]\}_{t\geq 0}$ with mean rate of λ. Arriving packets are routed to one of N infinite-size queues, which provides service according to an i.i.d. Bernoulli process $\{S_n[t]\}_{t\geq 0}$ with mean rate of μ_n, $\forall n = 1, 2, ..., N$.

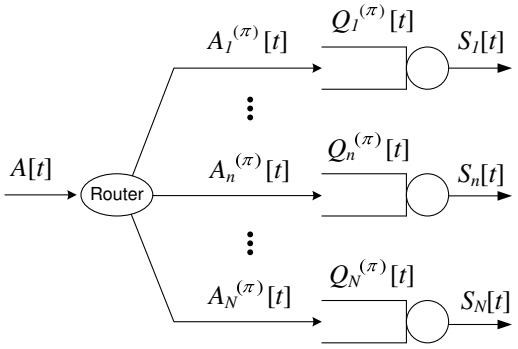


Fig. 1: Routing to parallel queues.

In each slot t, a *routing policy* π routes $A_n^{(\pi)}[t]$ of the incoming packets to the n^{th} queue such that $A[t] = \sum_{n=1}^{N} A_n^{(\pi)}[t]$. We let $Q_n^{(\pi)}[t]$ be the length of queue n in slot t under policy π, whose evolution is given by

$$\begin{aligned} Q_n^{(\pi)}[t+1] &= \max\left(0, Q_n^{(\pi)}[t] + A_n^{(\pi)}[t] - S_n[t]\right) \quad (1)\\ &= Q_n^{(\pi)}[t] + A_n^{(\pi)}[t] - S_n[t] + U_n^{(\pi)}[t], \end{aligned}$$

where $U_n^{(\pi)}[t] \triangleq \max(0, S_n[t] - Q_n^{(\pi)}[t] - A_n^{(\pi)}[t])$ denotes the amount of unused service at the n^{th} server in slot t. Accordingly, $S_n[t] - U_n^{(\pi)}[t]$ denotes the number of departures from queue n in slot t under policy π.

[1]In this work, we focus on the case of Bernoulli processes to simplify the exposition and analysis. However, many of our results can be extended to more general processes.

In this work, we are interested in the operation of the system in overloaded conditions, i.e., when the overload rate $\epsilon \triangleq \lambda - \sum_{n=1}^{N} \mu_n > 0$. As argued in Section I, an important metric of performance in such a scenario is the *expected total cumulative unused services until time T starting from zero initial state*[2]:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} U_{\Sigma}^{(\pi)}[t]\right], \quad (2)$$

under routing policy π, where $U_{\Sigma}^{(\pi)}[t] \triangleq \sum_{n=1}^{N} U_n^{(\pi)}[t]$.

This metric can characterize the convergence speed of the running-average of the expected departure rate by noting that

$$\begin{aligned} \mathbb{E}\left[\overline{d}^{(\pi)}[T]\right] &\triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[S_n[t] - U_n^{(\pi)}[t]\right]\\ &= \sum_{n=1}^{N} \mu_n - \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} U_{\Sigma}^{(\pi)}[t]\right]. \quad (3) \end{aligned}$$

This expression clearly shows that any policy π satisfying

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} U_{\Sigma}^{(\pi)}[t]\right] = 0$$

achieves the maximum departure rate of $\sum_{n=1}^{N} \mu_n$ that the system can provide. Yet, many policies, including randomized policy (see discussion following Definition 2), can possess this limiting behavior. The study of $\mathbb{E}\left[\sum_{t=0}^{T-1} U_{\Sigma}^{(\pi)}[t]\right]$ is important in extracting additional critical information about the convergence speed of the running-average of the expected departure rate to its limit. This motivates us to investigate the expected cumulative unused service performance of routing over parallel queues.

In addition to fundamental bounds for all feasible policies, in this work, we study the performance of two well-known routing policies: Join-the-Shortest-Queue (JSQ) policy ([1]), and Randomized Routing (RR) policy, described next.

Definition 1 (JSQ policy): In each time slot, the Join-the-Shortest-Queue (JSQ) policy forwards all incoming packets to the queue with the *shortest queue-length* in that time-slot. In case of ties, it selects a queue uniformly at random among the queues with the shortest queue-length.

The JSQ policy has been shown to possess many desirable properties under stable conditions (i.e., $\lambda < \sum_{n=1}^{N} \mu_n$), such as throughput-optimality, and mean delay optimality in heavy-traffic regimes, i.e., it minimizes the mean delay as the arrival rate approaches to the total sum of the service rates ([2]). We are interested in investigating whether the JSQ policy also performs well in terms of the metric in (2) in the over-loaded regime.

Another simple routing policy is the Randomized Routing (RR) policy, which is defined as follows:

[2]Throughout this work, we assume (unless stated otherwise) the initial condition of the system to be zero for all queues in order to capture the worst case cumulative unused service performance. Yet, the results are extendable to non-zero initial conditions.

Definition 2 (RR policy): In each time slot, the Randomized Routing (RR) policy forwards all incoming packets to the queue n with probability $q_n \triangleq \frac{\mu_n}{\sum_{i=1}^{N}\mu_i}$, $\forall n = 1, 2, ..., N$, i.e., in proportion to its service rate.

We note that the RR policy requires statistical information about the service rates, while the JSQ policy needs queue-length information in each slot. It is easy to see that both the JSQ and RR policies can stabilize the system under stable conditions. Under overloaded conditions, the system is unstable under both policies.Yet, it can also be seen that all queues will overflow under both policies, and the departure rate expression in (3) will converge to the optimal level of $\sum_n \mu_n$ under each policy. Thus, in the departure rate sense, both JSQ and RR policies are optimal. Yet, their performance in the new metric (2) will be significantly different, as we will observe.

Since the system is necessarily unstable in overloaded conditions, the traditional steady-state analysis does not apply, making it difficult to analyze the unused service metric. In the next section, we tackle this challenge by proposing a queue reversal theorem. This fundamental theorem seamlessly relates the unstable system to a related stable system, and helps us develop a lower bound on the expected cumulative unused service for any feasible routing policy.

III. Lower Bound Analysis

In this section, we establish a fundamental lower bound on the expected cumulative unused service (2) in the over-loaded system of Figure 1. To derive the lower bound, we first establish an interesting and surprising relationship between a single-server queue with i.i.d. arrivals and i.i.d. services, and a reverse queue in which the roles of the arrival and service processes are interchanged.

A. Queue Reversal Theorem

We assume that the arrival and service processes to a single-server queue are two independent sequences of i.i.d. nonnegative-valued random variables[3] $\{\alpha[t]\}_{t\geq 0}$ and $\{\beta[t]\}_{t\geq 0}$. Let $\Phi[t]$ be its queue-length in slot t, which evolves as

$$\Phi[t+1] = \Phi[t] + \alpha[t] - \beta[t] + \gamma[t], \quad t \geq 0, \tag{4}$$

where $\gamma[t] \triangleq \max\{0, \beta[t] - \Phi[t] - \alpha[t]\}$ denotes the amount of unused service in slot t.

Consider a hypothetical reversal of the single-server queue by exchanging the arrival and service processes. Let $\Phi^{(r)}[t]$ be the queue-length of the reverse queue in slot t. Then, the evolution of $\Phi^{(r)}[t]$ can be described as

$$\Phi^{(r)}[t+1] = \beta[t] - \alpha[t] + \gamma^{(r)}[t], \quad t \geq 0, \tag{5}$$

where $\gamma^{(r)}[t] \triangleq \max\{0, \alpha[t] - \Phi^{(r)}[t] - \beta[t]\}$ denotes the amount of unused service of the reverse queue in slot t. The single-server queue (also named the forward queue) and its reverse queue are shown in Figure 2.

[3]We note that these random variables need not be Bernoulli distributed as in the original system. This generality is necessary to apply the result to the multi-server system in the following subsection.

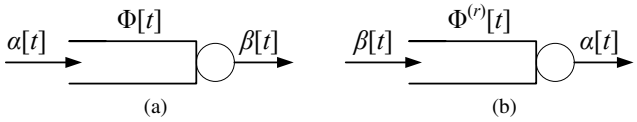


Fig. 2: (a) Forward queue; (b) Reverse queue.

We now provide a key relationship between the forward and reverse queues.

*Theorem 1 (**Queue Reversal Theorem**):* Suppose the forward and reverse queues introduced above (cf. Fig. 2) start from zero, i.e., $\Phi[0] = \Phi^{(r)}[0] = 0$. Then, for any $t \geq 0$, the total expected unused service until time t in the forward queue is equal to the expected queue-length at time t of the reverse queue. Similarly, for any $t \geq 0$, the expected queue-length at time t of the forward queue is equal to the total expected unused service until time t in the reverse queue.

In other words, given that $\Phi[0] = \Phi^{(r)}[0] = 0$, we have

$$\mathbb{E}[\Phi[t]] = \sum_{\tau=0}^{t-1} \mathbb{E}[\gamma^{(r)}[\tau]], \quad \forall t \geq 1 \tag{6}$$

$$\mathbb{E}[\Phi^{(r)}[t]] = \sum_{\tau=0}^{t-1} \mathbb{E}[\gamma[\tau]], \quad \forall t \geq 1. \tag{7}$$

Proof: Let $X[\tau] \triangleq \alpha[\tau] - \beta[\tau], \forall \tau \geq 0$. Then, we have

$$\begin{aligned}
\mathbb{E}[\Phi[t]] &\overset{(a)}{=} \mathbb{E}\left[\max_{0\leq m\leq t-1}\left\{\sum_{\tau=m}^{t-1} X[\tau], 0\right\}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\max_{0\leq m\leq t-1}\left\{\sum_{\tau=0}^{m} X[\tau], 0\right\}\right] \\
&\overset{(c)}{=} \sum_{k=1}^{t}\frac{1}{k}\mathbb{E}\left[\max\left\{\sum_{\tau=0}^{k-1} X[\tau], 0\right\}\right],
\end{aligned} \tag{8}$$

where step: (a) follows from the Lindley's equation; (b) uses the fact that $(X[0], X[1], ..., X[t-1])$ and $(X[t-1], ..., X[1], X[0])$ have the same distribution since $\{X[\tau]\}_{\tau\geq 0}$ are i.i.d.; (c) follows from the Spitzer's Identity (see [10]). Similarly, we can show that

$$\mathbb{E}[\Phi^{(r)}[t]] = \sum_{k=1}^{t}\frac{1}{k}\mathbb{E}\left[\max\left\{-\sum_{\tau=0}^{k-1} X[\tau], 0\right\}\right]. \tag{9}$$

By using the identity $\max\{x, y\} = x + y - \min\{x, y\}$, we have

$$\begin{aligned}
\mathbb{E}[\Phi[t]] &= \sum_{k=1}^{t}\frac{1}{k}\mathbb{E}\left[\sum_{\tau=0}^{k-1} X[\tau] - \min\left\{\sum_{\tau=0}^{k-1} X[\tau], 0\right\}\right] \\
&\overset{(a)}{=} \sum_{\tau=0}^{t-1}\mathbb{E}[X[\tau]] + \sum_{k=1}^{t}\frac{1}{k}\mathbb{E}\left[\max\left\{-\sum_{\tau=0}^{k-1} X[\tau], 0\right\}\right] \\
&\overset{(b)}{=} \sum_{\tau=0}^{t-1}\mathbb{E}[X[\tau]] + \mathbb{E}[\Phi^{(r)}[t]],
\end{aligned} \tag{10}$$

where (a) follows from the fact that $X[\tau], \tau \geq 0$ are i.i.d.; and (b) utilizes equation (9). By summing (4) over $\tau =$

$0, 1, ..., t-1$ and taking the expectation on both sides, we have

$$\mathbb{E}[\Phi[t]] = \sum_{\tau=0}^{t-1} \mathbb{E}[X[\tau]] + \sum_{\tau=0}^{t-1} \mathbb{E}[\gamma[\tau]]. \quad (11)$$

By comparing the equations (10) and (11), we have

$$\mathbb{E}[\Phi^{(r)}[t]] = \sum_{\tau=0}^{t-1} \mathbb{E}[\gamma[\tau]], \quad (12)$$

which proves (6). The proof of (7) follows the same steps, but with the roles of forward and reverse queues switched. ■

We note that this result is a bit surprising, since the cumulative unused service is non-decreasing in each sample path while the queue-length in the reverse queue may increase, decrease, or stay the same over time. Yet, their means are remains equal for all t. The key contribution of the Queue Reversal Theorem is that it relates the metric of interest in an overloaded (forward) queue (i.e., the mean arrival rate is strictly greater than the mean service rate) to the queue-length in a under-loaded (reverse) queue, for which we have mature and rich analytical tools. In fact, this theorem helps us establish a lower bound on the expected cumulative unused service under any feasible routing policy.

B. Lower Bound on the Cumulative Unused Service

We are ready to establish the necessary lower bound on the unused service in the over-loaded system of Figure 1. We construct a hypothetical single-server queue illustrated in Figure 3 with an infinite buffer, the same arrival process $\{A[t]\}_{t\geq 0}$ as before, and the service process $\{S_\Sigma[t]\}_{t\geq 0}$, where $S_\Sigma[t] \triangleq \sum_{n=1}^{N} S_n[t]$ is the total services of the original multi-server system. Thus, the single-server queueing

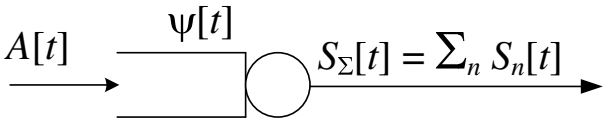


Fig. 3: Lower bounding system.

system stores all arrivals in a single queue for service at the combined service amount of its multi-server counterpart. Then, the queue-length process, $\{\Psi[t]\}_{t\geq 0}$, of the new system evolves as:

$$\begin{aligned} \Psi[t+1] &= \max\left(0, \Psi[t] + A[t] - S_\Sigma[t]\right) \\ &= \Psi[t] + A[t] - S_\Sigma[t] + W[t], \end{aligned} \quad (13)$$

where $W[t] \triangleq \max(0, S_\Sigma[t] - A[t] - \Psi[t])$ denotes the amount of unused service in slot t offered by the combined (also called *resource pooled*) server. The following lemma establishes the stochastic dominance relationship[4] between the original multi-server system and this single-server system.

Lemma 1: For any initial queue-length vector $\mathbf{Q}[0] = (Q_n[0])_{n=1}^N$, set $\Psi[0] = \sum_{n=1}^N Q_n[0]$. Then, under any feasible (possibly non-causal) routing policy π, the queue-length process $\{\mathbf{Q}^{(\pi)}[t]\}_{t\geq 0}$ and the unused service process $\{\mathbf{U}^{(\pi)}[t]\}_{t\geq 0}$ satisfies
(i) $\Psi[t] \preccurlyeq_{st} Q_\Sigma^{(\pi)}[t] \triangleq \sum_{n=1}^N Q_n^{(\pi)}[t]$, for all $t \geq 0$, and
(ii) $\sum_{\tau=0}^t W[\tau] \preccurlyeq_{st} \sum_{\tau=0}^t U_\Sigma^{(\pi)}[\tau]$, for all $t \geq 0$.

That is, $\Psi[t]$ and $\sum_{\tau=0}^t W[\tau]$ are stochastically dominated by the total queue-length and the total unused services of the N-queue system of Figure 1.

Proof: This follows from coupling the sample paths of queue-length process in the original system with that of the lower bounding system. ■

Next, we utilize the result (ii) of Lemma 1 that the total unused service of the original multi-server system of Figure 1 is lower-bounded by the cumulative unused service in the single-server system of Figure 3, and then apply the Queue Reversal Theorem to get a lower bound on the expected cumulative unused service. Note that our lower bound holds for more general arrival and service processes.

Assumption 1 (Basic Assumptions): We assume that:
(i) The service processes $\{(S_n[t])_n\}_{t\geq 0}$ are i.i.d. sequences of non-negative integer-valued and bounded random variables with $\mathbb{P}(S_n[1] \leq S^{max}) = 1$, for each n. We use the notations: $\mu_n \triangleq \mathbb{E}[S_n[1]]$, $\sigma_n^2 \triangleq var(S_n[1])$, and $\mu \triangleq \sum_{n=1}^N \mu_n$.
(ii) The arrival process $\{A[t]\}_{t\geq 0}$ is an i.i.d. sequence of non-negative integer-valued and bounded random variables with: $\mathbb{E}[A[1]] = \mu + \epsilon$, and $\mathbb{P}(A[1] \leq A^{max}) = 1$, where A^{max} is independent of ϵ. We use the notations: $\lambda \triangleq \mathbb{E}[A[1]]$, and $(\sigma^{(\epsilon)})^2 \triangleq var(A[1])$.

Proposition 1: Suppose Assumption 1 hold, and that the system starts from zero initial state $\mathbf{Q}[0] = \mathbf{0}$. Then, the unused service process $\{\mathbf{U}^{(\pi)}[t]\}_{t\geq 0}$ achieved by any feasible routing policy π satisfies

$$\mathbb{E}\left[\sum_{t=0}^{\infty} U_\Sigma^{(\pi)}[t]\right] \geq b_1^{(\epsilon)} \triangleq \frac{\zeta^{(\epsilon)}}{2\epsilon} - \frac{A^{\max}}{2}, \quad (14)$$

where $\zeta^{(\epsilon)} \triangleq \left(\sigma^{(\epsilon)}\right)^2 + \sum_{n=1}^N \sigma_n^2 + \epsilon^2$. Further, if $(\sigma^{(\epsilon)})^2$ converges to a constant σ^2 as $\epsilon \downarrow 0$, then,

$$\lim_{\epsilon\downarrow 0} \epsilon\mathbb{E}\left[\sum_{t=0}^{\infty} U_\Sigma^{(\pi)}[t]\right] \geq \frac{1}{2}\left(\sigma^2 + \sum_{n=1}^N \sigma_n^2\right). \quad (15)$$

Proof: The proof directly follows from Queue Reversal Theorem and Lemma 4 in [2]. ■

In the case of symmetric conditions, with Bernoulli arrivals with mean λ and symmetric Bernoulli services with mean $(\lambda - \epsilon)/N$ for each server, the lower bound becomes:

$$\lim_{\epsilon\downarrow 0} \epsilon\mathbb{E}\left[\sum_{t=0}^{\infty} U_\Sigma^{(\pi)}[t]\right] \geq \lambda - \frac{\lambda^2}{2}\left(1 + \frac{1}{N}\right). \quad (16)$$

We remark that this proposition not only shows that the total cumulative unused service over time for the unstable

[4] A random variable X is said to be stochastically dominated by another random variable Y, denoted as $X \preccurlyeq_{st} Y$, if $F_Y(z) \leq F_X(z)$ for all $z \in \mathbb{R}$.

system is lower-bounded, but also explicitly relates the bound to the degree of over-load ϵ and the number of servers N.

IV. Overload Analysis of RR and JSQ policies

In this section, we analyze the performance of the two well-known policies: the RR policy and the JSQ policy under the Bernoulli arrivals and services. We calculate lower and upper bounds on the expected cumulative unused service under the RR policy. The upper and lower bounds match as ϵ scales down to zero. Then, we show that the JSQ policy minimizes the cumulative unused service in the stochastic order sense under symmetric servers.

A. Lower and Upper Bounds under the RR Policy

In this subsection, we provide lower and upper bounds on the expected cumulative unused services under the RR policy in the system with Bernoulli arrivals and services. Recall that λ is the arrival rate, μ_n is the service rate of the n^{th} queue, and $\lambda = \sum_{n=1}^{N} \mu_n + \epsilon$, where $\epsilon > 0$. Then, we have the following result.

Proposition 2: Assume the system starts from the zero initial state $\mathbf{Q}[0] = 0$. Then, the unused service process $\{\mathbf{U}^{(RR)}[t]\}_{t\geq 0}$ under the RR policy satisfies

$$b_{LB}^{(RR)} \leq \epsilon \mathbb{E}\left[\sum_{t=0}^{\infty} U_{\Sigma}^{(RR)}[t]\right] \leq b_{UB}^{(RR)}, \tag{17}$$

where $b_{LB}^{(RR)} \triangleq \frac{\zeta_1}{2} - \frac{N\epsilon}{2}$, and $b_{UB}^{(RR)} \triangleq \frac{\zeta_1}{2}$, and also $\zeta_1 \triangleq \lambda\left(N-\lambda\right) + \left(\lambda - \epsilon\right)\left(N - \lambda + \epsilon\right) + \epsilon^2$. This implies

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{t=0}^{\infty} U_{\Sigma}^{(RR)}[t]\right] = \lambda(N-\lambda). \tag{18}$$

Proof: Under the RR policy, the n^{th} queue is equivalent to having the Bernoulli arrivals with the mean of $\lambda \frac{\mu_n}{\sum_{i=1}^{N} \mu_i}$ and the Bernoulli services with the mean of μ_n. Then, we apply the Queue Reversal Theorem to each individual queue and utilize Lemma 4 in [2] to get desired results. ∎

Since $\mathbb{E}\left[\sum_{t=0}^{\infty} U_{\Sigma}^{(RR)}[t]\right] < \infty$ for any $\epsilon > 0$, the running-average departure rate can converge to its service rate. Yet, for each $\epsilon > 0$, $\epsilon \mathbb{E}\left[\sum_{t=0}^{\infty} U_{\Sigma}^{(RR)}[t]\right]$ linearly increases with the number of queues N, which implies that the speed at which the running-average departure rate converges under the RR policy scales down linearly with the number of queues, which is undesirable in large-scale networks. In contrast, the limiting behavior of the lower bound in (16) is inversely related to N, which suggests that there is a potential for significant performance improvement over the RR policy in large scale systems. This motivates us to study the expected cumulative unused services under the JSQ policy, which possesses many good properties in under-loaded regimes.

B. Optimality of the JSQ Policy in Symmetric Conditions

In this subsection, we study the problem of optimal routing policy with respect to cumulative unused service under symmetric Bernoulli service processes. We show that the JSQ policy minimizes the process of the total cumulative unused service in the stochastic ordering sense.

Proposition 3: For the parallel queueing system with Bernoulli arrivals and symmetric Bernoulli services in Figure 1, let $\left\{\mathbf{U}^{(JSQ)}[t]\right\}_{t\geq 0}$ and $\left\{\mathbf{U}^{(\pi)}[t]\right\}_{t\geq 0}$ be the unused service processes under the JSQ policy and any feasible routing policy π, respectively. Then,

$$\sum_{\tau=0}^{t} U_{\Sigma}^{(JSQ)}[\tau] \preccurlyeq_{st} \sum_{\tau=0}^{t} U_{\Sigma}^{(\pi)}[\tau]. \tag{19}$$

Proof: This follows from coupling the queue-length realizations under JSQ and policy π appropriately. The proof is almost the same as in [11], and thus is omitted here for brevity. ∎

Even though the JSQ policy is optimal in terms of the cumulative unused service, its closeness to the lower bound we derived in Section III is unclear. In the next section, we demonstrate with numerical investigations that the expected cumulative unused service under the JSQ policy, for both symmetric and asymmetric service processes, matches the lower bound as ϵ scales down to zero, and more importantly, is independent of the number of queues when ϵ is sufficiently small.

V. Numerical Results

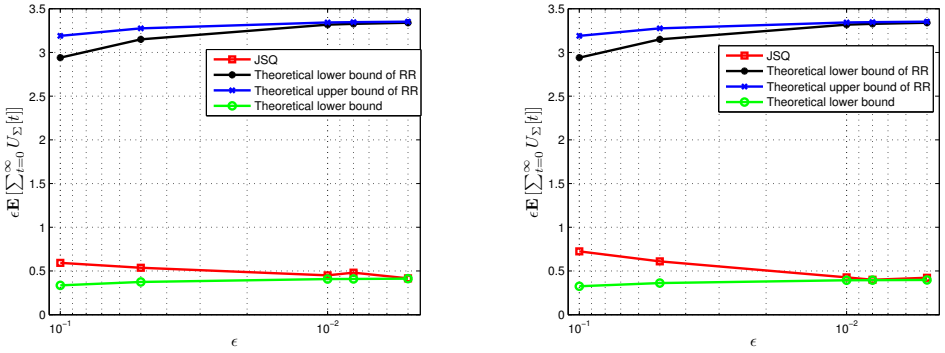
In this section, we provide simulations to compare the expected cumulative unused service performance of the JSQ policy with the RR policy. In the simulation, we take $\lambda = 0.8$. We consider both symmetric and non-symmetric Bernoulli service processes. In the symmetric setup, the service rate for the n^{th} queue is $\mu_n = (\lambda - \epsilon)/N$, while in the non-symmetric case, the service rate for n^{th} queue is $\mu_n = 2n(\lambda - \epsilon)/(N(N+1))$. We study the impact of the overload level $\epsilon = \lambda - \sum_n \mu_n > 0$ and the number of queues N on the expected cumulative unused service.

A. The Impact of Overload Level ϵ on Mean Unused Services

Figure 4 shows the impact of ϵ on the expected cumulative unused service of JSQ and RR policies when there are $N = 5$ queues. From Figure 4, we can observe that the expected cumulative unused service under the JSQ policy converges to the theoretical lower bound as ϵ scales down to zero, while the RR policy always keeps away from the theoretical lower bound. Thus, we conjecture that the JSQ policy is overload-optimal, i.e., it minimizes the expected cumulative unused service as ϵ diminishes, while the RR policy is sub-optimal. We leave the proof of this conjecture to future investigation.

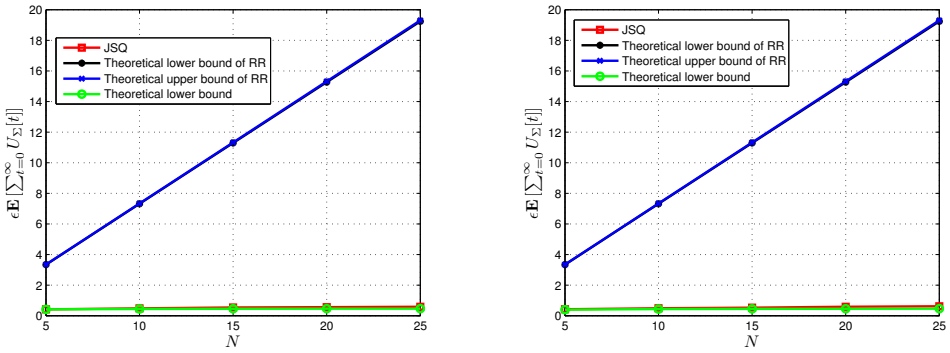
B. The Impact of Server Number N on Mean Unused Service

In this subsection, we study the impact of the number of queues on the expected cumulative unused service under the JSQ policy and the RR policy. Here, we fix ϵ to 0.005, and vary the number of queues from 5 to 25. From Figure 5, we can observe that the expected cumulative unused service under the JSQ policy stays close to the theoretical lower bound as the number of queues N increases, while it scales linearly with N under the RR policy, as derived in



(a) Symmetric services (b) Asymmetric services

Fig. 4: The impact of ϵ on the expected cumulative unused service.



(a) Symmetric Services (b) Asymmetric Services

Fig. 5: The impact of N on the expected cumulative unused service.

Proposition 2. This indicates that the performance of the JSQ policy is insensitive to the network size, which is desirable in large-scale networks.

VI. Conclusions

We considered a queueing system in which Bernoulli arrivals are routed to parallel servers with independent Bernoulli service processes. We studied the system performance in the overloaded regime, i.e., the total arrival rate is greater than the sum of the service rates of the queues. We proposed the use of cumulative unused service as a key metric in overloaded conditions, and provided a fundamental lower bound on it for any feasible routing policy. In the process of deriving the lower bound, we found a surprising result, which may be interesting in its own right: the expected cumulative unused service in a single-server queue with i.i.d. arrivals and i.i.d. services is equal to the expected queue-length in a queueing system where the roles of arrival and service processes are exchanged. Then, we compared the derived lower bound with the performance of two well-known routing policies, namely randomized and join-the-shortest queue, to observe their optimality and sub-optimality characteristics with respect to the new metric.

References

[1] G. J. Foschini and J. Salz, "A basic dynamic routing problem and diffusion," *IEEE Transactions on Communications*, vol. 26, no. 3, pp. 320–327, March 1978.

[2] A. Eryilmaz and R. Srikant, "Asymptotically tight steady-state queue-length bounds implied by drift conditions," *Queueing Systems Theory and Applications (QUESTA)*, vol. 72, no. 3-4, pp. 311–359, 2012.

[3] B. Li and A. Eryilmaz, "Optimal constant splitting for efficient routing over unreliable networks," in *In Proc. IEEE conference on Decision and Control (CDC)*, Maui, Hawaii, December 2012.

[4] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, pp. 1936–1948, December 1992.

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

[6] L. Georgiadis and L. Tassiulas, "Optimal overload response in sensor networks," *IEEE Transactions on Information Theory*, pp. 2684–2696, 2006.

[7] C. Chan, M. Armony, and N. Bambos, "Fairness in overloaded parallel queues," 2011, http://arxiv.org/pdf/1011.1237.pdf.

[8] D. Shah and D. Wischik, "Fluid models of congestion collapse in overloaded switched networks," *Queueing Systems Theory and Applications (QUESTA)*, vol. 69, pp. 121–143, October 2011.

[9] C. Li and E. Modiano, "Receiver-based flow control for networks in overload," in *to appear in the Proceedings of IEEE INFOCOM*, 2013.

[10] S. M. Ross, *Stochastic Processes*. Wiley, 1995.

[11] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory*, vol. 39, pp. 466–478, March 1993.