RC23414 (W0411-052) November 8, 2004
Mathematics

IBM Research Report

Optimal Probabilistic Routing in Distributed Parallel Queues

Xin Guo, Yingdong Lu, Mark S. Squillante
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Optimal Probabilistic Routing in Distributed Parallel Queues

Xin Guo, Yingdong Lu, Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

ABSTRACT

In this paper we consider the fundamental problem of routing customers among multiple distributed parallel queues to minimize an objective function based on equilibrium sojourn times, which arises in a wide variety of distributed computer systems, networks and applications. We derive optimal solutions to this theoretical scheduling problem under general assumptions for the arrival and service processes through stochastic-process limits. Our analysis extends previous studies by providing explicit solutions for the optimal scheduling problem and by considering general single-server queues, including correlated arrivals, under both first-come first-serve and processor-sharing queueing disciplines. In addition, we derive bounds for the variance of customer waiting times and exploit these results in order to obtain optimal solutions to the scheduling problem of interest based on equilibrium sojourn times subject to constraints on the waiting time variance, which have been ignored in previous studies. This collection of results allow us to cover risk factors and incorporate risk management within the context of our optimal scheduling problem. Numerical experiments with data from a real Web server system demonstrate the potential benefits of our theoretical results and methods in practice.

1. INTRODUCTION

The fundamental problem of scheduling a stream of customers among a set of distributed parallel queues to achieve some performance objective has received and continues to receive considerable attention in the research literature. This is motivated by the complexity of the theoretical problem and by the importance of the problem in practice where it arises in a wide variety of distributed computer applications and distributed computer system and communication network environments. A particular instance of the general problem is motivated by scalable Web server systems where incoming user requests are immediately routed to one of a set of computing nodes by a high-speed, load-balancing router and each node independently executes the requests routed to it.

We consider the theoretical problem of optimally routing customers among multiple distributed heterogeneous single-server parallel queues to minimize an objective function based on equilibrium sojourn times. This distributed scheduling problem involves two distinct issues: (i) the routing of customers to the parallel queues; and (ii) the queueing discipline for serving customers at each queue. The second issue is addressed by considering both first-come first-serve (FCFS) and processor sharing (PS) queueing disciplines, given the importance of these disciplines in the queueing theory literature and in the application areas motivating our study. We therefore derive optimal solutions to the theoretical scheduling problem of interest under general assumptions for the arrival and service processes and under the assumption that customers are routed to the parallel queues in a probabilistic manner. More specifically, we derive explicit solutions for the optimal vector of probabilities that control the routing of customers upon arrival among a set of heterogeneous general single-server queues through stochastic-process limits. Our assumption of probabilistic routing is consistent with previous theoretical studies of this fundamental optimization problem, and our solutions can be used for the parameter settings of other routing mechanisms found in practice, such as the weights of a weighted round-robin scheme. We also derive upper bounds for the variance of customer waiting times within this mathematical framework and exploit these results in order to obtain optimal solutions to the scheduling problem of interest based on equilibrium sojourn times subject to constraints on the waiting time variance.

Related scheduling problems have received considerable attention in the research literature. Our scheduling problem is consistent with or a generalization of the problems considered in [29, 30, 4, 23, 5, 9, 27, 8, 15, 25, 13] and the relevant references therein. A number of these studies [9, 8, 15, 13] have analyzed the performance of specific policies, as opposed to obtaining the optimal solution. Borst [5] and Sethuraman and Squillante [25] consider the problem of finding the optimal routing matrix in a multiclass variant of our scheduling problem, but under the restrictions of Poisson arrivals or fluid models for the individual queues. In addition, these studies provide results on the structural properties of the optimal solution, whereas we derive an explicit solution for the optimal scheduling problem under general assumptions for the queues. Our scheduling problem is also related to a global load-balancing optimization problem that has received considerable attention in the literature; e.g., see [29, 30, 4, 23] and the references cited therein. Ross and Yao [23] consider a problem that is similar to the problem studied in this paper, with the addition of a dedicated independent stream of customer arrivals to each server having non-preemptive priority over the other customers. Bonomi and Kumar [4] consider a model similar to that in [23] but with additional restrictions, and in both studies the objective is to minimize the expected response time taken over the two sets of customers where each arrival stream is a Poisson process. Our analysis addresses the single-class opti-

mal scheduling problem using different methods than those proposed in [4, 23] and eliminating the restriction of Poisson arrivals by focusing on general stochastic processes for arrivals. Shanthikumar and Xu [27] consider a scheduling optimization problem that is most related to our study, but they restrict their attention to independent and identically distributed (i.i.d.) arrival and service times and to the FCFS queueing discipline. Our analysis improves upon the results in [27] for the problem instance with renewal arrival and service processes and the FCFS discipline, extends these results to include the PS queueing discipline, and further considers an important form of correlated arrivals based on general regime-switching processes.

In contrast with [29, 30, 4, 23, 5, 9, 27, 8, 15, 25, 13] and other related studies, we also derive bounds for the customer waiting time variance in each of the problem instances considered and then exploit these results to obtain the solution that minimizes a function of the equilibrium sojourn times while satisfying constraints on the waiting time variance. This is important because the potential for high volatility in the sojourn time performance measure often observed in some system environments can seriously jeopardize the efficacy of the system if the optimum is only considered within a first-order context. Hence, it is quite natural to include such second-order risk factors and use them to address issues related to risk management in the formulation of the optimal scheduling problem. Furthermore, as part of our derivations of these extensions and other extensions noted above, we provide bounds to many important performance measures such as the variation of the multidimensional diffusion process and the moments of the running maximum of the regime-switching diffusion process.

The scheduling strategy considered in this paper is static in the sense that the routing probabilities do not change dynamically with time nor do they depend upon the states of the individual queues. While dynamic scheduling policies have the potential to outperform static policies [34, 11, 31, 10, 18, 32, 28, 14, 16], our focus in this paper is instead on the classical scheduling problem considered in previous studies (some of which were noted above) where the only information available to the routing policy concerns the overall customer arrival process and the customer service process at each single-server queue. In addition to its theoretical interest, this (static) scheduling problem is motivated by distributed system environments where the complexities and overheads of dynamic policies tend to outweigh their potential benefits over static policies. On the other hand, the use of our optimal scheduling solution in practice can also consist of repeated adjustments of the optimal routing vector with changes in the system environment, such as variations in the customer traffic.

The remainder of this paper is organized as follows. We first present the general model and framework for our analysis of the scheduling optimization problem. Then in Section 3 we derive the optimal solution for this problem when customer arrivals follow a renewal process. Section 4 extends this analysis to consider the scheduling optimization problem under a general form of correlated arrivals. The results of some numerical experiments are presented in Section 5, followed by our concluding remarks.

2. MATHEMATICAL MODEL AND FRAMEWORK

We consider a queueing system consisting of a high-speed router in front of N heterogeneous single-server parallel queues, as illustrated in Figure 1. Customers arrive to the system according to a general point process $\mathbf{A}(t)$ where the (marginal) distribution A of the corresponding increment process (i.e., interarrival distribution) on $\mathbb{R}^+$ has mean $\mathsf{E}[A] = \lambda^{-1}$ and variance $\mathsf{Var}[A] = \sigma_A^2$. Each customer is routed to one of the queues immediately upon its arrival according to a probability vector $\mathbf{P} \equiv [p_n]_{1 \le n \le N}$, independent of all else; i.e., a customer arrival is independently routed to queue n with probability p_n. The router is assumed to be sufficiently fast that customers essentially have no service demands and do not queue at the router. Customer service times are i.i.d. following general distributions S_n on $\mathbb{R}^+$ that depend upon the queue where the customer is served and have mean $\mathsf{E}[S_n] = \mu_n^{-1}$ and coefficient of variation $C_{S_n}^2$, $n = 1, \ldots, N$, mutually independent of the arrival and routing processes.

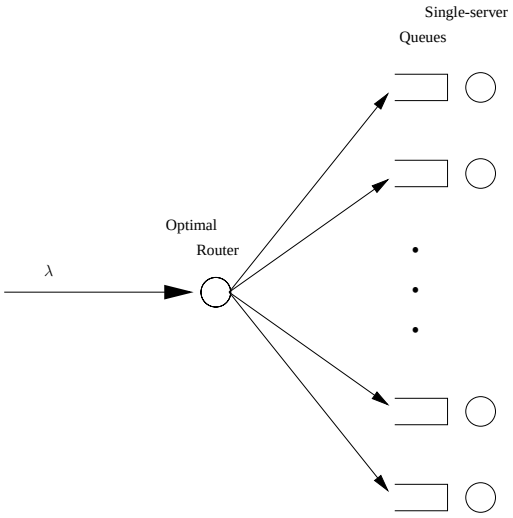

Figure 1: Queueing Network Model of Distibuted Parallel Queues

Each of the N single-server queues independently serves the customers routed to it under either an FCFS or PS queueing discipline. Let Z_n be an independent geometrically distributed random variable having mean p_n^{-1}. Then the customer arrival process for queue n is a general point process $\mathbf{A}_n(t)$ with (marginal) interarrival distribution A_n given by

$$A_n \;=\; \sum_{k=1}^{Z_n} X_k, \tag{1}$$

where the sequence of random variables $X_1, X_2, \ldots$ follow the increment process of the exogenous arrival point process $\mathbf{A}(t)$. Let $\lambda_n = \mathsf{E}[A_n]^{-1}$ be the mean arrival rate of customers to queue n, and let $\rho_n = \lambda_n/\mu_n$ be the traffic intensity for queue n.

In order to be able to handle general arrival and service processes and obtain explicit solutions to the scheduling optimization problems of interest, we consider stochastic-process limits for each of the N queues under the routing vector $\mathbf{P}$ as the basis of a general mathematical framework for our analysis. The core idea of this approach is to approximate the actual queueing processes by more mathematically tractable limiting processes, i.e., reflected diffusion processes in our case, each of which is exact in the limit as the queueing system approaches its critical value. Hence, the queueing analysis of our mathematical framework consists of two major components: the verification of the approximation, and the mathematical analysis of the approximation process. The verification is usually carried out through proving various types of weak convergence in probability measures induced by the queueing process; refer to Billingsley [3] and Whitt [33] for additional details. The

typical procedure consists of applying different scalings of time and space for the queueing process, and then the stochastic-process limits are obtained as the scaling parameters go to infinity. This is the approach taken in Section 3 to handle our queueing system under renewal arrival processes. When the random variables of the interarrival and service times have r finite higher moments, then the desired stochastic-process limit can be characterized by a pointwise estimation of the original process and the approximation process, where the pointwise estimation improves according to $o(T^{1/r})$ and converges to $o(\log(T))$ for finite moment generating functions. The pointwise estimation is termed the strong approximation, and it can be easily shown that strong approximation implies weak convergence; see, e.g., [6, Section 5.5.4] for details. This is the approach taken in Section 4, where our main contribution lies in the analysis of the approximation process. More specifically, we analyze a reflected Markov-modulated diffusion process that represents the stochastic-process limit for each single-server queue under a regime-switching arrival process. While this class of stochastic processes is important from a theoretical perspective, our choice is also motivated by recent studies showing this class of regime-switching processes to accurately model the type of correlated arrivals found in practice in system environments such as Web sites [21, 22]. Using integration equation techniques, we are able to derive a probabilistic characterization of the running maximum of the Markov-modulated diffusion process, which is the key ingredient in obtaining the statistics of our reflected diffusion process. In contrast to the rich literature on convergence results, the research literature on the analysis of diffusion proccesses as approximations for general queueing networks is very limited with few known results except for the simplest case, which is that of the reflected Brownian motion. Our study therefore provides an important step forward in an area that has many applications. Moreover, the result is of independent interest its own right, and also can be applied in areas such as stochastic control and mathematical finance.

The final major component of our mathematical framework is to exploit these diffusion approximations of the distributed parallel queues to calculate the vector of optimal routing probabilities. Let $\mathcal{T}$ be the random variable denoting the equilibrium sojourn time for customers in the queueing system. From the law of total probability we have $\mathsf{E}[\mathcal{T}] = \sum_{n=1}^{N} \mathsf{E}[\mathcal{T}_n] \cdot \mathsf{P}[\text{ customer served at queue } n\,]$, where $\mathsf{E}[\mathcal{T}_n]$ is the equilibrium sojourn time of customers served at queue n. Let $h_n < \infty$ be the holding cost, or weight, per customer per unit time at queue n. One of the scheduling optimization problems of interest in this paper is then given by

$$\text{(OR1)} \qquad \min \sum_{n=1}^{N} h_n \, \mathsf{E}[\mathcal{T}_n], \tag{2}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} p_n = 1, \qquad p_n \geq 0. \tag{3}$$

This formulation is consistent with the objective function considered in [27]. An alternative optimization problem of interest is obtained by replacing the objective function in (2) with

$$\text{(OR2)} \qquad \min \sum_{n=1}^{N} h_n \, \mathsf{E}[\mathcal{T}_n] \, p_n, \tag{4}$$

which is consistent with the objective function considered in [5, 25]. In both scheduling optimization problems, the decision variables are the routing probabilities p_n, $n = 1, \ldots, N$. The next two sections will determine the optimal routing probability vector both from among all possible solutions and from among those solutions whose variance satisfy certain constraints, under different assumptions for the exogenous arrival process.

3. RENEWAL ARRIVALS

In this section we consider the case where $\mathbf{A}(t)$ is a renewal process. Thus, under the independent probabilistic splitting of arrivals in our model, the arrival process to each of the N queues is also a renewal process with interarrival distribution as expressed in (1). From this expression and Wald's equation [24], we have

$$\begin{aligned} \mathsf{E}[A_n] &= \lambda_n^{-1} = \lambda^{-1} p_n^{-1}, \\ \mathsf{Var}[A_n] &= \frac{\sigma_A^2 p_n + \lambda^{-2}(1-p_n)}{p_n^2}, \\ C_{A_n}^2 &= \lambda^2 \, \sigma_A^2 \, p_n + 1 - p_n, \end{aligned}$$

where $C_{A_n}^2$ is the squared coefficient of variation for the interarrival distribution at queue n. Each queue n is therefore a GI/GI/1 queue with arrival and service processes having mean rates λ_n and μ_n and squared coefficients of variation $C_{A_n}^2$ and $C_{S_n}^2$, respectively.

3.1 FCFS queueing discipline

Our goals are to: (i) establish the stochastic-process limit for each GI/GI/1 FCFS queue n; (ii) analyze this diffusion process to obtain an approximation for the corresponding equilibrium sojourn time; and (iii) use this diffusion approximation to calculate the vector of optimal routing probabilities. Let $u_{n,k}$ represent the time between the $k-1^{st}$ and k^{th} customer arrivals at queue n, and let $v_{n,k}$ represent the service time of the k^{th} customer arrival, $k \geq 1$. Define $U_{n,k} \equiv u_{1,k} + \ldots + u_{n,k}$, $V_{n,k} \equiv v_{1,k} + \ldots + v_{n,k}$, $k \geq 1$, $N_n^U(t) \equiv \max\{\ell : U_{n,\ell} \leq t,\, \ell \geq 0\}$, $N_n^V(t) \equiv \max\{\ell : V_{n,\ell} \leq t,\, \ell \geq 0\}$, $t \geq 0$. Let $C_n(t) = \sum_{\ell=1}^{N_n^U(t)} V_{n,\ell}$ be the cumulative input process, and let $X_n(t) = C_n(t) - t$ be the associated net-input process, both for queue n, $t \geq 0$. We can define the workload process for queue n by $L_n(t) \equiv X_n(t) - \inf\{X(s) \wedge 0 : 0 \leq s \leq t\}$ and the corresponding queue length process by $Q_n(t) \equiv N_n^U(t) - N_n^V(C_n(t) - L_n(t))$, where $a \wedge b \equiv \min\{a, b\}$, $t \geq 0$.

Now let $u_{n,k}^m$ and $v_{n,k}^m$ represent the interarrival time and the service time of the k^{th} customer in the m^{th} instance of queue n in a sequence of instances of queue n, $k \geq 1$. The stochastic-process limits are then obtained for scaled versions of the stochastic processes associated with the m^{th} instance of queue n such that the traffic intensities for the sequence of queue n instances increase successively to the critical value of 1 as $m \to \infty$. In establishing these stochastic-process limits, we will be using the theory of weak convergence of probability measures on the space D of all right-continuous functions with finite left-limits on $[0, \infty)$; refer to Billingsley [3] and Whitt [33]. Assuming the sequence $\{u_{n,k}^m, v_{n,k}^m\}$ is stationary, we define a sequence of queues with $(\lambda_n^m)^{-1} = \mathsf{E}[u_{n,1}^m]$ and $(\mu_n^m)^{-1} = \mathsf{E}[v_{n,1}^m]$ which vary such that $\rho_n^m = (1 - m^{-1/2}) \to 1$ as $m \to \infty$, where ρ_n^m is the traffic intensity for the m^{th} instance of queue n. Let $\mathbf{L}_n^m(t)$ and $\mathbf{Q}_n^m(t)$ be the scaled random elements of D associated with the above workload and queue length processes defined as $\mathbf{L}_n^m(t) \equiv m^{-1/2} L_n^m(mt)$ and $\mathbf{Q}_n^m(t) \equiv m^{-1/2} Q_n^m(mt)$, $t \geq 0$, respectively. It then can be shown that

$$\mathbf{L}_n^m \;\Rightarrow\; \mathbf{L}_n \quad \text{as} \quad m \to \infty, \tag{5}$$

$$\mathbf{Q}_n^m \;\Rightarrow\; \mathbf{Q}_n \quad \text{as} \quad m \to \infty, \tag{6}$$

where $\Rightarrow$ denotes convergence in distribution, and $\mathbf{L}_n$ and $\mathbf{Q}_n$ are reflected Brownian motion (RBM) processes.

The stochastic-process limit $\mathbf{Q}_n$ in (6) for the GI/GI/1 FCFS queue n can be shown to be an RBM on $\mathbb{R}_+$ with drift $\lambda_n - \mu_n < 0$

and variance $\lambda_n(C_{S_n}^2 + C_{A_n}^2)$. More specifically, suppose $\mathbf{Q}_n$ to be an RBM defined in this manner and let $Q_n(t)$ be the length of queue n at time t as defined above. Upon applying Theorem 6.16 in [6], we have

$$\sup_{0\le t\le T} |Q_n(t) - \mathbf{Q}_n(t)| \;=\; o(T^{1/r}), \quad a.s. \tag{7}$$

When A and S_n have finite moment generating functions at a neighborhood of zero, this can be further improved to

$$\sup_{0\le t\le T} |Q_n(t) - \mathbf{Q}_n(t)| \;=\; o(\log(T)), \quad a.s., \tag{8}$$

or equivalently

$$\mathsf{P}[||Q_n(t) - \mathbf{Q}_n(t)||_T > C_1 \log T + x] \;\le\; C_2 e^{-C_3 x}, \tag{9}$$

where C_i, $i = 1, 2, 3$ are constants that are independent of T and x, and $||\cdot||_T$ denotes the supremum norm for functions on $[0, T]$.

We know that the RBM with (negative) drift $\lambda_n - \mu_n < 0$ and variance $\lambda_n(C_{S_n}^2 + C_{A_n}^2)$ has an invariant measure, which is an exponential distribution with rate $-2(\lambda_n - \mu_n)\lambda_n^{-1}(C_{S_n}^2 + C_{A_n}^2)^{-1}$. Using this diffusion approximation, the equilibrium number of customers at queue n can be approximated by

$$\mathsf{E}[\mathbf{Q}_n] \;\approx\; \rho_n \;+\; \frac{\lambda_n(C_{A_n}^2 + C_{S_n}^2)}{2(\mu_n - \lambda_n)}, \tag{10}$$

and upon applying Little's Law we obtain the corresponding diffusion approximation for the equilibrium sojourn time at queue n as

$$\begin{aligned} \mathsf{E}[\mathcal{T}_n] \;&\approx\; \frac{1}{\mu_n} \;+\; \frac{C_{A_n}^2 + C_{S_n}^2}{2(\mu_n - \lambda p_n)} \\ &\approx\; \frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{2(\mu_n - \lambda p_n)}. \end{aligned} \tag{11}$$

Substituting this into (2) and (4) respectively yields (OR1-IID-FCFS):

$$\min \;\sum_{n=1}^{N} h_n \left[\frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{2(\mu_n - \lambda p_n)}\right]; \tag{12}$$

and (OR2-IID-FCFS):

$$\min \;\sum_{n=1}^{N} h_n \left[\frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{2(\mu_n - \lambda p_n)}\right] p_n; \tag{13}$$

both subject to (3).

The solution for (OR1-IID-FCFS) can be obtained in closed form by applying the Lagrange method, which yields the optimal routing probability for queue n as follows

$$p_n \;=\; \frac{\mu_n}{\lambda} - \frac{\sum_{n=1}^{N}\mu_n - \lambda}{\lambda} \;\times\; \frac{\sqrt{h_n(\lambda^2\sigma_A^2 + \lambda^2 + C_{S_n}^2)\lambda - \lambda^2 h_n\mu_n}}{\sum_{n=1}^{N}\sqrt{h_n(\lambda^2\sigma_A^2 + \lambda^2 + C_{S_n}^2)\lambda - \lambda^2 h_n\mu_n}}. \tag{14}$$

We observe that the objective function in (OR2-IID-FCFS) is convex in the decision variables, and thus its solution can be efficiently computed using known methods in convex optimization; see, e.g., [2].

Remark: The above analysis holds for i.i.d. interarrival and service times that follow general distributions with light-tails. While this includes distributions with the so-called heavy-tailed property [1], it does not directly include heavy-tailed distributions for the i.i.d. interarrival and/or service times. However, our framework based on stochastic-process limits supports an analysis of such queues. Refer to [33, 17] for the technical details.

3.2 PS queueing discipline

Our goals in this section are the same as those of the previous section. An identical sequence of arguments could be made to show that the stochastic-process limit for the workload process of the GI/GI/1 PS queue n is as expressed in (5). In fact, it can be easily established that this result holds for the limiting workload process of any GI/GI/1 queue under a work conserving queueing discipline. The stochastic-process limit for the queue length process $\mathbf{Q}_n$ at the GI/GI/1 PS queue n also can be shown to be an RBM on $\mathbb{R}_+$, but the mean of the diffusion approximation for the distribution of this stochastic-process limit generally differs from that of the GI/GI/1 FCFS queue of the previous section. For our diffusion approximation under the PS queueing discipline, we rely on Grishechkin [12, 26]. The stochastic-process limit $\mathbf{Q}_n$ in (6) for the GI/GI/1 PS queue n can be shown to be an RBM on $\mathbb{R}_+$ with drift $[(1 + C_{S_n}^2)/2](\lambda_n - \mu_n) < 0$ and variance $\lambda_n(C_{S_n}^2 + C_{A_n}^2)$.

We know that the RBM with (negative) drift $[(1+C_{S_n}^2)/2](\lambda_n - \mu_n) < 0$ and variance $\lambda_n(C_{S_n}^2 + C_{A_n}^2)$ has an invariant measure, which is an exponential distribution with rate $-(1 + C_{S_n}^2)(\lambda_n - \mu_n)\lambda_n^{-1}(C_{S_n}^2 + C_{A_n}^2)^{-1}$. Using this diffusion approximation, the equilibrium number of customers at queue n can be approximated by

$$\mathsf{E}[\mathbf{Q}_n] \;\approx\; \rho_n \;+\; \frac{\lambda_n(C_{A_n}^2 + C_{S_n}^2)}{(1 + C_{S_n}^2)(\mu_n - \lambda_n)}, \tag{15}$$

and upon applying Little's Law we obtain the corresponding diffusion approximation for the equilibrium sojourn time at queue n as

$$\begin{aligned} \mathsf{E}[\mathcal{T}_n] \;&\approx\; \frac{1}{\mu_n} \;+\; \frac{C_{A_n}^2 + C_{S_n}^2}{(1 + C_{S_n}^2)(\mu_n - \lambda p_n)}, \\ &\approx\; \frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{(1 + C_{S_n}^2)(\mu_n - \lambda p_n)}. \end{aligned} \tag{16}$$

Note that the result in (16) is identical to the corresponding FCFS result in (11) when $C_{S_n}^2 = 1$. Substituting this result into (2) and (4) respectively yields (OR1-IID-PS):

$$\min \;\sum_{n=1}^{N} h_n \left[\frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{(1 + C_{S_n}^2)(\mu_n - \lambda p_n)}\right]; \tag{17}$$

and (OR2-IID-PS):

$$\min \;\sum_{n=1}^{N} h_n \left[\frac{1}{\mu_n} \;+\; \frac{\lambda^2\sigma_A^2 p_n + 1 - p_n + C_{S_n}^2}{(1 + C_{S_n}^2)(\mu_n - \lambda p_n)}\right] p_n; \tag{18}$$

both subject to (3).

The solution for (OR1-IID-PS) can be obtained in closed form by applying the Lagrange method, which yields the optimal routing probability for queue n as follows

$$p_n \;=\; \frac{\mu_n}{\lambda} - \frac{\sum_{n=1}^{N}\mu_n - \lambda}{\lambda} \;\times\; \frac{\sqrt{\frac{h_n[\mu_n(\lambda\sigma_A^2 - 1) + \lambda(1 + C_{S_n}^2)]}{\lambda(1 + C_{S_n}^2)}}}{\sum_{n=1}^{N}\sqrt{\frac{h_n[\mu_n(\lambda\sigma_A^2 - 1) + \lambda(1 + C_{S_n}^2)]}{\lambda(1 + C_{S_n}^2)}}}. \tag{19}$$

We observe that the objective function in (OR2-IID-PS) is convex in the decision variables, and thus its solution can be efficiently computed using known methods in convex optimization; see, e.g., [2].

3.3 Optimal routing with risk management

Both sets of solutions above minimize functions of the equilibrium sojourn times without any conditions beyond having the routing probabilities sum to 1. For reasons noted above, we also would like to determine the optimal solution when side constraints on the variance of customer waiting times are added. Specifically, we would like to find the optimal solution subject to some bound on the customer waiting time variance.

The first step is to derive an expression for the waiting time variance. In order to have our results apply to both sets of solutions above (and the set of solutions provided below), we consider for each queue n a generic RBM $\mathbf{R}_n$ having drift $\zeta_n < 0$ and variance ω_n. We know that the marginal distribution of the steady-state distribution of this limiting stochastic process is an exponential distribution with rate $-2\zeta_n/\omega_n$. However, it is extremely difficult to determine the correlation between different variables, which is required for an exact calculation of the variance of the customer waiting times. We therefore use an upper bound on the variance as follows:

$$\begin{aligned}
\mathsf{Var}\left[\sum_{n=1}^{N} p_n\mathbf{R}_n\right] &= \mathsf{E}\left[\sum_{n=1}^{N} p_n\mathbf{R}_n\right]^2 - \left(\mathsf{E}\left[\sum_{n=1}^{N} p_n\mathbf{R}_n\right]\right)^2 \\
&\leq N\sum_{n=1}^{N} p_n^2\mathsf{E}\left[\mathbf{R}_n\right]^2 - \left(\mathsf{E}\left[\sum_{n=1}^{N} p_n\mathbf{R}_n\right]\right)^2 \\
&= 2N\sum_{n=1}^{N}\frac{p_n^2\omega_n^2}{(-2\zeta_n)^2} - \left(\sum_{n=1}^{N}\frac{p_n\omega_n}{-2\zeta_n}\right)^2
\end{aligned}$$

Hence, the added condition based on the variation as a risk factor can be surrogated by the following side constraint

$$2N\sum_{n=1}^{N}\frac{p_n^2\omega_n^2}{(-2\zeta_n)^2} - \left(\sum_{n=1}^{N}\frac{p_n\omega_n}{-2\zeta_n}\right)^2 \leq \alpha. \qquad (20)$$

The optimization problems (OR1-IID-FCFS), (OR2-IID-FCFS), (OR1-IID-PS) and (OR2-IID-PS) are then solved as described above but with the additional constraint given in (20).

4. REGIME-SWITCHING MODEL

It has been frequently noted that many of the systems modeled by queueing networks exhibit various correlations and fluctuations in the arrival and service processes. One effective way to capture these correlations and fluctuations is to assume a regime-switching model for the processes. For the type of systems motivating our study, such correlations structures are especially important in the arrival process and thus this is our focus herein, although our analysis can be extended beyond this case. More concretely, we assume that there is an independent continuous Markov chain $\delta(t)$ taking on values $0, 1, \cdots, K$ with a time homogeneous transition probability, such that at each Markov state, the arrival rate is A_i. In most systems, a Markov chain of two state is sufficient to represent the "peak" and "off-peak" behavior. Hence, to elucidate the exposition, we shall assume in the sequel that the $\delta(t)$ is a continuous Markov chain on 2 states, i.e., 0 and 1. The analysis can be carried out with no difficulty for a general finite-state Markov chain.

We start with some strong approximation results for this Markov modulated queueing system. As in the previous section, we focus on one of the single-server queues, but instead drop the queue index n to reduce tedious notational issues. Consider a generic single-server queue with intervarrival process $\hat{A}$, whose mean is $1/\hat{\lambda}_\delta$ and coeffient of variation is $C^2_{\hat{a},\delta}$, and service time distribution S, whose mean is $\hat{\mu}_\delta$ and coefficient of variation is $C^2_{\hat{s},\delta}$. The subscript δ indicates the state of the Markov chain $\delta(t)$. Strong approximation results can be established for the queueing process. We then derive an expression for the steady state distribution of the approximation process, which is a Markov modulated diffusion process. In the end, we can apply the analysis to obtain the optimal routing probability for our distributed parallel system.

4.1 Strong approximation

Suppose that $\delta(t)$ has the following generator matrix $Q = (q_{ij})$ with

$$Q = \begin{pmatrix} -\gamma_0 & \gamma_0 \\ \gamma_1 & -\gamma_1 \end{pmatrix}.$$

We will show that the queue length of this δ-modulated queueing system can be approximated by a δ-modulated diffusion process. It should be noted that prior to our work, Choudhury, et al. [7] studied a similar system. However, their diffusion approximation result is more restrictive in that it was obtained by forcing the Markov state to change at the same rate. Our work is more along the lines of Massey and Mandelbaum [19], in which the queueing process is approximated on a pointwise basis, and the modulating process will be the same for the approximating process.

As we showed in the previous section, the queue length process in a generic $GI/GI/1$ queue can be approximated by a diffusion process through the strong approximation. For the Markov modulated queueing system, between two consecutive time epochs in which $\delta(t)$ changes its state, the queue process behaves like an $GI/GI/1$ queue, and thus it can be approximated in a pointwise manner by a diffusion process. More specifically, when $\delta(t) = 0$, the queue length process $Q^0(t)$ can be strongly approximated by the following reflected Brownian motion,

$$X^0(t) = \sigma_0 W^0(t) + \beta_0 t + \sup_{0\leq s\leq t}[-\sigma_0 W^0(t) - \beta_0 t]^+,$$

where $\beta_0 \equiv \hat{\lambda}_0 - \hat{\mu}_0$ and $\sigma_0 = \sqrt{\hat{\lambda}_0(C^2_{\hat{a},0} + C^2_{\hat{s},0})}$. Moreover, there exist constants C_1^0, C_2^0 and C_3^0 such that

$$\mathsf{P}[||Q^0 - X^0||_T > C_1^0 \log T + x] \leq C_2^0 e^{-C_3^0 x}$$

Similarly, the limiting process for $Q^1(t)$, i.e., the queue length process when $\delta(t) = 1$, is given by

$$X^1(t) = \sigma_1 W^1(t) + \beta_1 t + \sup_{0\leq s\leq t}[-\sigma_1 W^1(t) - \beta_1 t]^+,$$

where $\beta_1 \equiv \hat{\lambda}_1 - \hat{\mu}_1$ and $\sigma_1 = \sqrt{\hat{\lambda}_1(C^2_{\hat{a},1} + C^2_{\hat{s},1})}$, with

$$\mathsf{P}[||Q^1 - X^1||_T > C_1^1 \log T + x] \leq C_2^1 e^{-C_3^1 x}.$$

Upon combining them, we have the following result.

THEOREM 4.1. *Let $Q^\delta(t)$ be the queue length process of a Markov-modulated queueing process, and $\tilde{Z}^\delta(t)$ bethe following Markov modulated diffusion process,*

$$\tilde{Z}^\delta(t) \equiv \sigma_\delta W^\delta(t) + \beta_\delta t + \sup_{0\leq s\leq t}[-\sigma_\delta W^\delta(t) - \beta_\delta t]^+,$$

then $\tilde{Z}^\delta(t)$ is a strong approximation of $Q^\delta(t)$.

PROOF. See the appendix. □

4.2 Analysis of Markov modulated Brownian motion

Now we can turn our focus to the probabilistic analysis of the reflected $\delta(t)$-modulated diffusion process. To proceed withing the

desired optimization framework, several key quantities regarding the reflected Markov modulated Brownian motions need to be in place.

First, recall that the stationary average queue length is approximated by

$$\frac{1}{t}\int_0^t \mathsf{E}[\tilde{Z}^\delta(s)]ds,$$

where

$$\mathsf{E}[\tilde{Z}^\delta(s)] = \mathsf{E}[\sigma_\delta W^\delta(t) + \beta_\delta t] + \mathsf{E}[\sup_{0\le s\le t}[-\sigma_\delta W^\delta(t) - \beta_\delta t]^+]. \tag{21}$$

The first term in Equation (21) can be found via the following lemma. To simplify the presentation, we shall henceforth assume that $\sigma_\delta = 1$, noting that this assumption can of course be removed easily when applying our analysis to any real system.

LEMMA 4.2. *Let $m_\delta(t)$ be the mean of the diffusion process $W^\delta(t)+\beta_\delta t$ at time t with initial condition that $\delta(0) = \delta \in \{0, 1\}$. We then have*

$$\begin{aligned} m_0(t) &= \frac{\gamma_1\beta_0 + \gamma_0\beta_1}{\gamma_0+\beta_1}t + \frac{\gamma_0(\beta_0-\beta_1)}{(\gamma_0+\gamma_1)^2}[1-e^{-(\gamma_0+\gamma_1)t}], \\ m_1(t) &= \frac{\gamma_1\beta_0 + \gamma_0\beta_1}{\gamma_0+\gamma_1}t + \frac{\gamma_1(\beta_0-\beta_1)}{(\gamma_0+\gamma_1)^2}[1-e^{-(\gamma_0+\gamma_1)t}]. \end{aligned}$$

PROOF. See the appendix. □

The second term in Equation (21) can be derived via direct calculations on distributions of the running maximum of a Markov-modulated diffusion process. Letting $M^\delta(t)$ be the running maximum process with $\delta(0) = \delta \in \{0, 1\}$, and upon conditioning on the time of the first jump τ_δ of the Markov chain, we then have the following recursion for the distribution of $M^\delta(t)$.

LEMMA 4.3.

$$\begin{aligned} \mathsf{P}[M^0(t)\ge x] &= \mathsf{P}[M_0(t)\ge x]\mathsf{P}[\tau_0\ge t] \\ &+ \int_0^t \mathsf{P}[M_0(s)\ge x]F_{\tau_0}(ds) \\ &+ \int_0^t\int_{-\infty}^x \mathsf{P}[M^1(t-s)\ge x-y]\times \\ &\quad \mathsf{P}[M_0(s)\le x|X_0(s)\in dy]\times \\ &\quad F_{X_0}(dy)F_{\tau_0}(ds) \end{aligned} \tag{22}$$

and

$$\begin{aligned} \mathsf{P}[M^1(t)\ge x] &= \mathsf{P}[M_1(t)\ge x]\mathsf{P}[\tau_1\ge t] \\ &+ \int_0^t \mathsf{P}[M_1(s)\ge x]F_{\tau_1}(ds) \\ &+ \int_0^t\int_{-\infty}^x \mathsf{P}[M^0(t-s)\ge x-y]\times \\ &\quad \mathsf{P}[M_1(s)\le x|X_1(s)\in dy]\times \\ &\quad F_{X_1}(dy)F_{\tau_1}(ds), \end{aligned} \tag{23}$$

where $M_\delta(t)$ denotes the running maximum of a Brownian motion X_δ with drift β_δ and variance σ_δ, $F_{\tau_\delta}(ds)$ denotes the density function of the duration of the Markov chain $\delta(t)$ at state $\delta = 0, 1$, and $F_{X_\delta}(dy)$ denotes the density function of the value of the diffusion process $X(\delta)(t)$.

Furthermore, upon taking the Laplace transform on both sides of the equation in the above lemma and expressing

$$\int_0^\infty P(M^0(t)\ge x)e^{-\theta t}dt = \tilde{G}^0(x,\theta),\ \ \theta > 0,$$

we obtain

THEOREM 4.4.

$$\tilde{G}^0(x,\theta) = H_1(x,\theta) + \sum_{i=1}^4 \int_0^\infty C_i\tilde{G}^0(y,\theta)e^{K_i y}dy \tag{24}$$

$$\tilde{G}^1(x,\theta) = H_2(x,\theta) + \sum_{i=1}^4 \int_0^\infty D_i\tilde{G}^1(y,\theta)e^{L_i y}dy \tag{25}$$

where

$$\begin{aligned} H_1(x,\theta) &= F_1(x,\theta) + F_2(x,\theta) \\ &+ \frac{\gamma_0}{A_1(A_1-\mu_0)}\int_0^\infty \bar{F}_1(z,\theta)dz \\ &- \frac{\gamma_0}{A_2(A_2-\beta_0)}\int_0^\infty \bar{F}_2(z,\theta)dz \\ &- \frac{\gamma_0}{A_2(A_2-\beta_0)}\int_0^\infty \bar{F}_1(z,\theta)dz \\ &+ \frac{\gamma_0}{A_1(A_1-\beta_0)}\int_0^\infty \bar{F}_2(z,\theta)dz, \\ H_2(x,\theta) &= \bar{F}_1(x,\theta) + \bar{F}_2(x,\theta) \\ &+ \frac{\gamma_1}{\bar{A}_1(\bar{A}_1-\beta_1)}\int_0^\infty F_1(z,\theta)dz \\ &- \frac{\gamma_1}{\bar{A}_2(\bar{A}_2-\beta_1)}\int_0^\infty F_2(z,\theta)dz \\ &- \frac{\gamma_1}{\bar{A}_2(\bar{A}_2-\beta_1)}\int_0^\infty F_1(z,\theta)dz \\ &+ \frac{\gamma_1}{\bar{A}_1(\bar{A}_1-\beta_1)}\int_0^\infty F_2(z,\theta)dz, \end{aligned}$$

$$\begin{aligned} F_1(x,\theta) &= \int_0^\infty P(\tau_0\ge t)P(M_0(t)\ge x)e^{-\theta t}dt, \\ F_2(x,\theta) &= \int_0^\infty\int_0^t P(M_0(t)\ge x)\gamma_0 e^{-\gamma_0 s}e^{-\theta t}dt, \\ \bar{F}_1(x,\theta) &= \int_0^\infty P(\tau_1\ge t)P(M_1(t)\ge x)e^{-\theta t}dt, \\ \bar{F}_2(x,\theta) &= \int_0^\infty\int_0^t P(M_1(t)\ge x)\gamma_1 e^{-\gamma_1 s}e^{-\theta t}dt \end{aligned}$$

$$\begin{aligned} L_1 &= -(\beta_0-A_1)+\beta_1-\bar{A}_1, L_2 = \beta_1-\bar{A}_1-(\beta_0+\bar{A}_2), \\ L_3 &= -(\beta_0+A_1)+\beta_1-\bar{A}_2, L_4 = -(\beta_0+A_2)+\beta_1-\bar{A}_2; \\ K_1 &= -(\beta_1-\bar{A}_1)+\beta_0-A_1, K_2 = \beta_0-A_1-(\beta_1+\bar{A}_2), \\ K_3 &= -(\beta_1+\bar{A}_1)+\beta_0-A_2, K_4 = -(\beta_1+\bar{A}_2)+\beta_0-A_2; \end{aligned}$$

$$\begin{aligned} D_1 &= \frac{\gamma_1\gamma_0}{A_1\bar{A}_1(\bar{A}_1-\beta_0)}, D_2 = -\frac{\gamma_0\gamma_1}{\bar{A}_1A_2(A_2-\beta_0)}, \\ D_3 &= -\frac{\gamma_0\gamma_1}{\bar{A}_2A_1(2\bar{A}_2+\bar{A}_1-\beta_0)}, D_4 = \frac{\gamma_0\gamma_1}{A_2\bar{A}_2(2\bar{A}_2+A_2-\beta_0)}; \\ C_1 &= \frac{\gamma_1\gamma_0}{A_1\bar{A}_1(A_1-\beta_1)}, C_2 = -\frac{\gamma_0\gamma_1}{A_1\bar{A}_2(\bar{A}_2-\beta_1)}, \\ C_3 &= -\frac{\gamma_0\gamma_1}{A_2\bar{A}_1(2A_2+A_1-\beta_1)}, C_4 = \frac{\gamma_0\gamma_1}{A_2\bar{A}_2(2A_2+\bar{A}_2-\beta_1)}; \\ A_1 &= \sqrt{2\theta+2\gamma_0+\beta_0^2}, A_2 = \sqrt{2\theta+2\gamma_0+2\beta_0^2}, \\ \bar{A}_1 &= \sqrt{2\theta+2\gamma_1+\beta_1^2}, \bar{A}_2 = \sqrt{2\theta+2\gamma_1+2\beta_1^2}. \end{aligned}$$

PROOF. See the appendix. □

Observe that Equations (24) and (25) are standard integral equations of the second type, whose solutions take the following form (e.g., see [20])

$$\tilde{G}^0(x,\theta) = H(x,\theta) + \int_0^\infty \sum_{i=1}^4 C_i H(y,\theta) e^{K_i y} dy \qquad (26)$$

Thus, we obtain the Laplace transform of the distribution of the running maximum process with respect to time t. Various algorithms can be employed to obtain the inverse of the transform. Our interest is to compute the steady state distribution of the reflected diffusion process, for which we observe

$$\begin{aligned} & f(\beta_1,\beta_0,\gamma_0,\gamma_1) \qquad (27)\\ = & \lim_{t\to\infty} \frac{1}{t}\int_0^t m_0(s) - M^0(s) ds \\ = & \lim_{t\to\infty} \frac{1}{t}\int_0^t m_0(s) - [\int_0^\infty \mathsf{P}[M^0(s)\geq x]dx]ds \\ = & \lim_{t\to\infty}\lim_{\eta\to 0} \frac{1}{t}\int_0^t m_0(s) - [\int_0^\infty e^{\eta x}\mathsf{P}[M^0(s)\geq x]dx]ds \end{aligned}$$

The Laplace transform with respect to t for the function

$$\int_0^t \int_0^\infty e^{\eta x}\mathsf{P}[M^0(s)\geq x]dx$$

is $\frac{1}{\theta}\tilde{G}(\eta,\theta)$. Hence, following Karamata's Tauberian theorem and Equation (27), we conclude that

$$\begin{aligned} & f(\beta_0,\beta_1,\gamma_0,\gamma_1) \\ = & \tilde{m}_0(0) - \lim_{\eta\to 0}\lim_{\theta\to 0}\int_0^\infty e^{\eta x}[H(x,\theta) \\ & + \int_0^\infty \sum_{i=1}^4 C_i H(y,\theta) e^{K_i y} dy]dx \qquad (28) \end{aligned}$$

where $\tilde{m}_0(\theta)$ denotes the Laplace transform of $m_0(t)$ with respect to time t.

Finally, given the parameters for the particular Markov modulated queueing system of interest, the optimal routing probabilities can be obtained by solving the following optimization problem

$$\begin{aligned} \min \quad & \sum_{n=1}^N h_n f(p_n\lambda_0 - \mu_n, p_n\lambda_1 - \mu_n, \gamma_0, \gamma_1), \qquad (29)\\ s.t. \quad & \sum_{n=1}^N p_n = 1, \;\; p_n \geq 0 \end{aligned}$$

where $\lambda_0 = \pi_0\lambda$ and $\lambda_1 = \pi_1\lambda$ denote the arrival rate when $\delta(t)$ takes on the value of 0 and 1, respectively, π is the invariant probability vector of Q, and the function f is given in (28).

5. NUMERICAL EXPERIMENTS

We conducted many experiments to further investigate the optimal routing problem and our solution. In this section we present two sets of such results. The first set of experiments exploit our closed-form results in previous sections. We compare the different output of systems under FCFS or PS discipline, and also observe the impact of the variance of the arrival and service process upon the optimal routing probability. The second set of experiments are based on a comparison between the system performance obtained by simulating the access logs of a commercial Web site with a load-balancing router and the corresponding performance obtained by simulating the same access logs under our optimal routing solution.

5.1 Comparison of Different Disciplines and the Impact of Variance

Consider a system that consists of an arrival stream with unit arrival rate, two servers with mean service time 0.4 and 0.5. The holding costs, or weights, for the two queues are $h_1 = 2$ and $h_2 = 3$. In the first experiment, we set the coefficient of variation (CV) for the two service time distributions to be 2 and 1.5, respectively, and then we vary the CV for the interarrival distribution from 0.6 to 1.5. In the second experiment, we fix the CV for the interarrival distribution and that of the service time at the first server, and then we vary the CV of the service time at the second server from 0.6 to 1.5. The optimal routing probabilities for the queues with FCFS and PS disciplines are calculated for both experiments. In Figure 2, the optimal routing probability is plotted. We can observe that the differences between the two disciplines are quite visible. The FCFS queues are more sensitive to changes in variance. Moreover, such changes in variance can yield trends in opposite directions for the FCFS and PS queues.

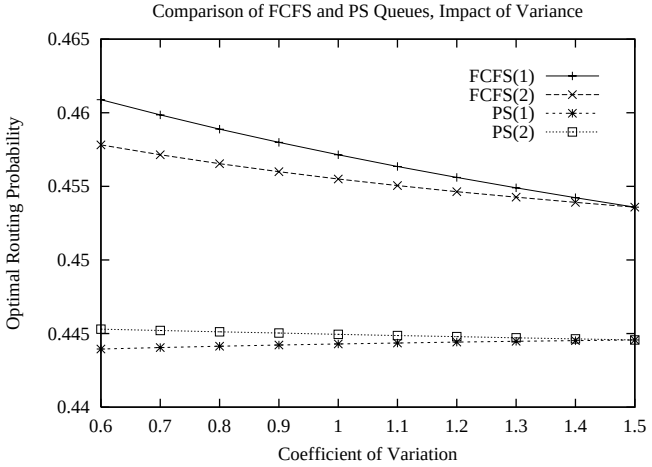

Figure 2: Impact of Variance on Optimal Solution under FCFS and PS Disciplines

5.2 Comparison with Web Site Data

While the main contribution of this paper is the foregoing collection of mathematical results, in this section we briefly consider the potential benefits of our optimal solution based on data from a real Web site. Specifically, we use the access log data from a large-scale international commercial Web site that consists of 3 complexes of servers at different locations, among which the overall offered traffic is geographically partitioned. That is, traffic originating from any given geographical area is directly routed to an assigned (primary) server-complex location (unless major portions of a server-complex location are taken off-line for some reason, in which case the other 2 locations will take over and serve this traffic). Each server-complex location consists of 4 server nodes to which incoming requests are routed by a front-end, high-speed router. This router attempts to balance the per-location load across the set of server nodes within a server-complex location, where each server node directly serves the client requests routed to it and operates independently of the other nodes.

Every server node comprising the Web site maintains its own access log of all client requests that are served by the node. Several months worth of data from 2003 was available to us. During this time period, the Web site as a whole served more than 200 requests per second on average, with a maximum of more than 700 requests per second during peak traffic intervals. Two stationary intervals of the access logs were used in our study, one being representative of stationary peak-traffic intervals and one being representative of stationary off-peak traffic intervals. Both intervals consist of a few hours worth of traffic at the Web site. For our purposes here, we are most interested in the arrival time and service time of the client requests served by each of the server nodes at all 3 locations. The access log contains the arrival time of each client request and the number of bytes comprising the request. We use a simple function of the byte size of each request (which has been validated against system measurements) to obtain the corresponding service time, although it is important to point out that changing the service time distribution did not change the trends of our results. By simulating each of these server nodes as a G/G/1 queue under the PS queueing discipline that is fed as input the sequence of arrival and service times from the corresponding access logs, we obtain an estimate of the performance exhibited at the Web site under the existing per-location routers. For comparison with our approach, we first used the access log data to compute the parameters of our results for the corresponding PS queue in Sections 3 and 4, and then we used our optimal solution for (OR2) with $h_1 = \ldots = h_4 = 1$ (and no side constraints) to obtain the optimal vector of routing probabilities. These probabilities were directly used as the weights of a weighted round-robin policy that was applied to the aggregate access log for each location in order to obtain the per-node traces for each of the 4 corresponding server nodes. We then used the same approach to simulate each server node as a G/G/1 PS queue, but instead with these per-node traces as input, in order to obtain an estimate of the performance under our optimal routing solution.

Table 1 provides the corresponding equilibrium sojourn time results for both peak and off-peak traffic intervals. In particular, we illustrate the performance of the 12 server nodes under our optimal routing solution relative to the performance under the routing policy employed at the existing Web site. Negative values imply that our optimal policy provides lower equilibrium sojourn times than those obtained under the existing-system routing policy and quantify such performance improvements, whereas positive values indicate improvements in the equilibrium sojourn time under the existing per-location load-balancing routers.

	Off-Peak Traffic	Peak Traffic
Node 1	-0.2219	0.0913
Node 2	-0.2972	0.7288
Node 3	-0.0535	-0.4046
Node 4	0.0177	-0.6385
Node 5	-0.2989	-0.9643
Node 6	-0.6589	-0.9460
Node 7	-0.8401	-0.9441
Node 8	-0.6710	-0.9673
Node 9	-0.0762	-0.2328
Node 10	-0.3178	-0.1804
Node 11	-0.0698	0.1984
Node 12	0.0119	-0.6373

Table 1: Relative per-node equilibrium sojourn time measures under optimal and existing-system routing policies.

The results for each server node in Table 1 show that, for the off-peak traffic intervals, the optimal routing solution always provides better equilibrium sojourn times than the per-location load-balancing router employed in the existing system. The only exceptions to this are node 4 and node 12, but the difference in each case is less than 2%. For the peak traffic intervals, we find even more significant improvements in performance under the optimal routing policy. The only exceptions to this are nodes 1, 2 and 11, but in these cases there are relatively larger performance improvements at other server nodes in the same location. We also observe that the optimal routing policy tends to provide fairly consistent equilibrium sojourn times at each of the server nodes of a given location, but this is not always the case under the existing system policy as illustrated in Table 1. Most importantly, the overall equilibrium sojourn time across all 3 locations (i.e., the objective function used in this set of experiments) is always better under our optimal routing solution than under the per-location load-balancing router employed in the existing system.

6. CONCLUSIONS

In this paper we studied the theoretical problem of optimally routing customers among multiple distributed heterogeneous single-server parallel queues to minimize an objective function based on equilibrium sojourn times. Our study makes the following contributions:

- We derived explicit solutions to the optimal routing problem under GI/GI/1 FCFS queues and different objective functions, extending and improving the results in [27];
- We derived explicit solutions to the optimal routing problem under GI/GI/1 PS queues and different objective functions;
- We derived a strong approximation for a reflected diffusion process, including a probabilistic characterization of the running maximum of the regime-switching diffusion process;
- We derived explicit solutions to the optimal routing problem under G/GI/1 queues with correlated arrivals and different objective functions, where the correlated arrivals are characterized by the general class of regime-switching processes;
- We derived upper bounds for the variance of waiting times that are used as side constraints for each of the stochastic routing optimization problems.

7. REFERENCES

[1] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 279–290, June 2001.

[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley and Sons, 2nd edition, 1993.

[3] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, second edition, 1999.

[4] F. Bonomi and A. Kumar. Adaptive optimal load balancing in a nonhomogeneous multiserver system with a central job scheduler. *IEEE Transactions on Computers*, 39(10):1232–1250, October 1990.

[5] S. C. Borst. Optimal probabilistic allocation of customer types to servers. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 116–125, 1995.

[6] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag, 2002.

[7] G. L. Choudhury, A. Mandelbaum, M. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models*, 13:121–146, 1997.

[8] M. E. Crovella, M. Harchol-Balter, and C. Murta. Task assignment in distributed systems: Improving performance by unbalancing load. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 268–269, June 1998.

[9] D. Dias, W. Kish, R. Mukherjee, and R. Tewari. A scalable and highly available Web server. In *Proceedings of the 1996 IEEE Computer Conference (COMPCON)*, February 1996.

[10] D. L. Eager, E. D. Lazowska, and J. Zahorjan. Adaptive load sharing in homogeneous distributed systems. *IEEE Transactions on Software Engineering*, SE-12(5):662–675, May 1986.

[11] G. J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, COM-26(3):320–327, March 1978.

[12] S. Grishechkin. GI/G/1 processor sharing queues in heavy traffic. *Advances in Applied Probability*, 26:539–555, 1994.

[13] M. Harchol-Balter. Task assignment with unknown duration. *Journal of the ACM*, 49(2), 2002.

[14] M. Harchol-Balter and A. Downey. Exploiting process lifetime distributions for dynamic load balancing. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 13–24, May 2001.

[15] G. Hunt, G. Goldszmidt, R. King, and R. Mukherjee. Network dispatcher: A connection router for scalable Internet services. In *Proceedings of the 7th International World Wide Web Conference*, April 1998.

[16] G. Koole, P. Sparaggis, and D. Towsley. Minimizing response times and queue lengths in systems of parallel queues. *Journal of Applied Probability*, 36:1185–1193, 1999.

[17] S. G. Kou and H. Wang. First passage times of a jump diffusion process. *Advances in Applied Probability*, 35:504–531, 2003.

[18] W. E. Leland and T. J. Ott. Load balancing heuristics and process behavior. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 56–69, May 1986.

[19] W. Massey and A. Mandelbaum. Strong approximations for time dependent queues. *Mathematics of Operations Research*, 20:33–64, 1995.

[20] A. D. Polyanin and A. V. Manzhirov. *Handbook of Integral Equations*. CRC Press, 1998.

[21] B. Ray and M. S. Squillante. A nonlinear model of Web server traffic and implications on long-range dependence. Technical report, IBM Research Division, 2002.

[22] A. Riska, M. S. Squillante, S.-Z. Yu, Z. Liu, and L. Zhang. Matrix-analytic analysis of a MAP/PH/1 queue fitted to Web server data. In *Advances in Algorithmic Methods for Stochastic Models, G. Latouche and P. Taylor (eds.)*. World Scientific, 2002.

[23] K. W. Ross and D. D. Yao. Optimal load balancing and scheduling in a distributed computer system. *Journal of the ACM*, 38(3):676–690, July 1991.

[24] S. M. Ross. *Stochastic Processes*. John Wiley and Sons, New York, second edition, 1997.

[25] J. Sethuraman and M. S. Squillante. Optimal stochastic scheduling in multiclass parallel queues. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 93–102, June 1999.

[26] J. Sethuraman, M. S. Squillante, and W. Whitt. Heavy-traffic limits for multiserver queues with limited timesharing. Preprint, May 2002.

[27] J. G. Shanthikumar and S. H. Xu. Asymptotically optimal routing and service rate allocation in a multiserver queueing system. *Operations Research*, 45(3):464–469, 1997.

[28] M. S. Squillante and R. D. Nelson. Analysis of task migration in shared-memory multiprocessors. In *Proceedings of ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 143–155, May 1991.

[29] A. N. Tantawi and D. Towsley. Optimal static load balancing in distributed computer systems. *Journal of the ACM*, 32(2):445–465, April 1985.

[30] Y. T. Wang and R. Morris. Load sharing in distributed systems. *IEEE Transactions on Computers*, C-34(3):204–217, 1985.

[31] R. W. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15:406–413, 1978.

[32] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34:55–62, 1986.

[33] W. Whitt. *Stochastic-Process Limits*. Springer-Verlag, New York, 2002.

[34] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181–189, 1977.

APPENDIX

A. APPENDIX

A.1 Proof Theorem 4.1

We define the random variable $N(t)$ to be the number of jumps that occur between time 0 and t. It then is clear that for each x, there exists an $M > 0$ such that $\mathsf{P}[N(T) > M] \leq C_2 e^{-C_3 x}$. Thus,

$$\begin{aligned}
& \mathsf{P}[\sup_{0\leq t\leq T} |Q^\delta(t) - \tilde{Z}^\delta(t)| > C_1 \log T + x] \\
\leq\ & \sum_{k=0}^{M} \mathsf{P}[\sup_{0\leq t\leq T} |Q^\delta(t) - \tilde{Z}^\delta(t) > C_1 \log T + x, N(T) = k] \\
& + \mathsf{P}[N(T) > M] \\
\leq\ & \sum_{k=0}^{M} \left\{ \mathsf{P}[\sup_{0\leq t\leq T} |Q^0(t) - \tilde{Z}^0(t)| > \frac{C_1 \log T + x}{k}, N(T) = k] \right. \\
& \left. + \mathsf{P}[\sup_{0\leq t\leq T} |Q^1(t) - \tilde{Z}^1(t)| > \frac{C_1 \log T + x}{k}, N(T) = k] \right\} \\
& + \mathsf{P}[N(T) > M] \\
\leq\ & \sum_{k=0}^{M} 2C_2 e^{-C_3 x} \mathsf{P}[N(T) = k] + \mathsf{P}[N(T) > M] \leq 3C_2 e^{-C_3 x}
\end{aligned}$$

The Theorem is now immediate by the equivalence property of strong approximations.

A.2 Proof of Lemma 4.2

Starting from time 0, suppose we are in state 0. Then between time 0 and Δt, we either see at least one jump from state 0 to 1, with probability $1 - e^{-\gamma_0 \Delta t}$, or we stay in state 0, with probability $e^{-\gamma_0 \Delta t}$ and the process starts again afresh, as a result of the

memoryless property of the exponential function. Since $\int_0^t dt = \int_0^{\Delta t} dt + \int_{\Delta t}^t dt$, we have

$$\begin{aligned} m_0(t) &= E\int_0^t \beta_\delta(s)ds\{1_{(jumps)} + 1_{(nojumps)}\} \\ &= P\{nojump\}\Delta t\beta_0 + m_0(t-\Delta) \\ &+ P\{jumps\}(1-e^{-\gamma_0\Delta t})m_0(t+\Delta t * \eta) \\ &+ O((\Delta t)^2) \quad |\eta| \le 1 \\ &= (1-\Delta t\gamma_0)\beta_0\Delta t + m_0(t) - \Delta t m_0'(t) \\ &+ \gamma_0\Delta t m_1(t) + O((\Delta t)^2). \end{aligned}$$

A Taylor's series expansion yields

$$\begin{aligned} m_0'(t) + \gamma_0 m_0(t) - \gamma_0 m_1(t) &= \beta_0, \\ m_1'(t) + \gamma_1 m_1(t) - \gamma_1 m_0(t) &= \beta_1, \end{aligned}$$

and after some simple calculations, we obtain

$$\begin{aligned} m_0''(t) + (\gamma_0+\gamma_1)m_0'(t) &= \gamma_1\beta_0 + \gamma_0\beta_1, \\ m_0(0) = 0, m_0'(0) &= \beta_0, \\ m_1''(t) + (\gamma_0+\gamma_1)m_1'(t) &= \gamma_1\beta_1 + \gamma_0\beta_0, \\ m_1(0) = 0, m_1'(0) &= \beta_1. \end{aligned}$$

Solving this second-order equation with different initial conditions then yields $(2.25), (2.26)$. □

A.3 Derivation of Theorem 4.4

First note that

LEMMA A.1. *Let $X(t) = \beta t + B(t)$, $M(t) = \sup_{s\le t} X(s)$, then for $x > y$*

$$\begin{aligned} & P(M(s) \le x, X(s) \in dy) \\ = \; & \frac{1}{\sqrt{2\pi s}}(\exp\frac{-(\beta s - y)^2}{2s} - \exp(\beta y - \frac{\beta^2 s}{2} - \frac{(2x-y)^2}{2s}))dy \end{aligned}$$

Upon taking the Laplace transform on both sides of Equation (22), we see that the left-hand side becomes

$$\int_0^\infty P(M^0(t) \ge x)e^{-\theta t}dt = \tilde{G}^0(x,\theta), \;\; \theta > 0.$$

Meanwhile, the last term on the right-hand side is given by

$$\begin{aligned} & \int_0^\infty \int_0^t \int_{-\infty}^x P(M^1(t-s) \ge x-y)e^{-\theta(t-s)}e^{-\theta s} \\ & \times \frac{1}{\sqrt{2\pi s}}(\exp\frac{-(\beta_0 s - y)^2}{2s} - \exp(\beta_0 y - \frac{\beta_0^2 s}{2} - \frac{(2x-y)^2}{2s})) \\ & \gamma_0 e^{-\gamma_0 s}dydsdt \\ = \; & \int_s^\infty \int_{-\infty}^x \int_0^\infty P(M^1(t-s) \ge x-y)e^{-\theta(t-s)} \\ & \times \frac{\gamma_0}{\sqrt{2\pi s}}(\exp(\frac{-(\beta_0 s - y)^2}{2s} - \theta s - \gamma_0 s) \\ & - \exp(\beta_0 y - \frac{\beta_0^2 s}{2} - \frac{(2x-y)^2}{2s} - \gamma_0 s))dsdydt \\ = \; & \int_0^\infty \int_{-\infty}^\infty P(M^1(t) \ge x-y)e^{-\theta t}dydt \\ & \times \int_0^\infty \frac{\gamma_0}{\sqrt{2\pi s}}(\exp(\frac{-(\beta_0 s - y)^2}{2s} - \theta s - \gamma_0 s) \\ & - \exp(\beta_0 y - \frac{\beta_0^2 s}{2} - \frac{(2x-y)^2}{2s} - \theta s - \gamma_0 s))ds \\ = \; & \int_0^\infty \int_{-\infty}^x P(M^1(t) \ge x-y)e^{-\theta t} \\ & \times (\frac{\gamma_0}{A_1}e^{(\beta_0 - A)y} - \frac{\gamma_0}{A_2}e^{\beta_0 y - A_2(2x-y)})dtdy, \end{aligned}$$

which is obtained with the help of

$$\int_0^\infty \exp(-as^2 + \frac{-b}{s^2})ds = \sqrt{\frac{\pi}{4a}}\exp(-2\sqrt{ab}), Re\, a > 0, Re\, b > 0.$$

Now, with a change of variable $x - y = z$, then the Laplace transform of Equation (22) becomes

$$\begin{aligned} \tilde{G}^0(x,\theta) &= \int_0^\infty \gamma_0\tilde{G}^1(z,\theta)(\frac{\exp((\beta_0 - A_1)(x-z))}{A_1} \\ &- \frac{\exp((\beta_0 - A_2)x - (\beta_0 + A_2)z)}{A_2})dz \\ &+ F_1(x,\theta) + F_2(x,\theta). \end{aligned} \tag{30}$$

Similarly,

$$\begin{aligned} \tilde{G}^1(z,\theta) &= \int_0^\infty \gamma_1\tilde{G}^0(y,\theta)(\frac{\exp((\beta_1 - \bar{A}_1)(z-y))}{\bar{A}_1} \\ &- \frac{\exp((\beta_1 - \bar{A}_2)x - (\beta_1 + \bar{A}_2)y)}{\bar{A}_2})dy + \\ &\bar{F}_1(z,\theta) + \bar{F}_2(z,\theta). \end{aligned} \tag{31}$$

Upon combining Equations (30) and (31), we have

$$\begin{aligned} & \tilde{G}^0(x,\theta) \\ = \; & \gamma_0\int_0^\infty (\int_0^\infty \gamma_1\tilde{G}^0(y,\theta)(\frac{\exp((\beta_1 - \bar{A}_1)(z-y))}{\bar{A}_1} \\ & - \frac{\exp((\beta_1 - \bar{A}_2)z - (\beta_1 + \bar{A}_2)y)}{\bar{A}_2})dy \\ & + \bar{F}_1(z,\theta) + \bar{F}_2(z,\theta)) \\ & \times \left[\frac{\exp((\beta_0 - A_1)(x-z))}{A_1} - \frac{\exp((\beta_0 - A_2)x - (\beta_0 + A_2)z)}{A_2}\right]dz \\ & + F_1(x,\theta) + F_2(x,\theta). \end{aligned}$$

Notice that $x - y = z$, and thus one can continue the calculation without much trouble from which Theorem 4.4 follows.