Optimal Scheduling in a Multiserver Stochastic Network

Mohsen Bayati
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA
bayati@stanford.edu

Mayank Sharma and Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
{mxsharma,mss}@us.ibm.com

1. INTRODUCTION

We consider a fundamental scheduling problem in a multiserver stochastic network consisting of 2 classes of customers and 2 classes of servers. Customers of class k arrive to queue k according to a Poisson process with rate λ_k, $k = 1, 2$. The service times of class k customers at class ℓ servers are i.i.d. following an exponential distribution with mean $\mu_{k\ell}^{-1}$, $\forall k, \ell = 1, 2$, where $0 < \mu_{11}, \mu_{12}, \mu_{22} < \infty$ and $\mu_{21} = 0$. Hence, class 1 customers can be served at both classes of servers, but class 2 customers can only be served at class 2 servers. A FCFS queueing discipline is employed at each queue. The customer arrival and service processes are mutually independent of each other and of all resource allocation decisions.

Suppose the *base model* is comprised of one class 1 server and one class 2 server. Let Q_k denote the random variable for the number of class k customers in the system, $k = 1, 2$. The scheduling problem of interest consists of determining the allocation of the class 1 server and the class 2 server among the customers of classes 1 and 2 according to the *objective function* $Q^* \triangleq \min c_1\mathbb{E}[Q_1] + c_2\mathbb{E}[Q_2]$, with $c_1, c_2 > 0$. This includes scheduling decisions that can occur upon the arrival and departure of customers of each class.

The above problem arises in a wide range of applications in parallel and distributed computer systems; refer to, e.g., [5, 6]. Another important application area motivating our study is skill-based routing in call centers. Much of the research in this area has focused on canonical skill-based routing topologies that represent the core structural components of more complex systems; see, e.g., [4]. Indeed, the above stochastic network model is often called the canonical N-design for skill-based routing. A few studies have analyzed the performance of call centers with a N-design skill-based routing topology under specific (suboptimal) scheduling policies [4]. Even fewer studies have considered the optimal scheduling of customers with respect to Q^* in multiserver N-design stochastic networks. Squillante et al. [7] establish the optimality of a $c\mu$-type scheduling policy in the fluid limit and derive a dynamic threshold-based policy (tree-based in general) to address stability problems under the pure $c\mu$ policy. Bell and Williams [2] prove the asymptotic optimality of a dynamic threshold-based scheduling policy (tree-based in general) in the heavy-traffic limit.

The present study is most closely related to [2, 7]. However, our goal is to determine the properties of the optimal scheduling policy with respect to Q^* in the original stochastic network as opposed to the fluid or diffusion scaling of this network. Furthermore, our study actually considers a more general stochastic network in which there are $S_\ell \geq 1$ servers of class ℓ (homogeneous within each class) and each class k customer that waits in queue k for an i.i.d. amount of time, having an exponential distribution with mean γ_k^{-1}, will leave the system without being served (mutually independent of all other stochastic processes and resource allocation decisions). This general scheduling problem then consists of determining the allocation of the S_1 class 1 servers and the S_2 class 2 servers among the customers of classes 1 and 2 according to the objective function Q^*. Note that the customer abandonment (reneging) in our general model is not considered in [2, 7]. On the other hand, since the technical details of our results even for the simpler base model defined above are quite involved and require more space than is available to us here, we restrict our attention in this extended abstract to the simpler base model and refer the interested reader to [1] for our more general results and technical details.

2. MATHEMATICAL ANALYSIS

Our analysis starts by analyzing the stochastic dynamic program corresponding to the scheduling problem in the above two-server base model without customer abandonment. The associated Bellman optimality equation for all states $(i, j) \in \mathbb{Z}_+^2$ is given by

$$\theta + \nu J_{i,j} = \min_{x_{i,j} \in [0,1]} \Big\{ g_{i,j}(x_{i,j}) + \nu \sum_{(i',j')} p_{ij,i'j'}(x_{i,j}) J_{i',j'} \Big\}, \quad (1)$$

where: $x_{i,j}$ is the control variable for state (i, j); $g_{i,j}(x_{i,j})$ is the cost function for state (i, j) and control $x_{i,j}$; $p_{ij,i'j'}(x_{i,j})$ is the probability that the next state will be (i', j') given that the current state is (i, j) and the selected control is $x_{i,j}$; $J_{i,j}$ is the value function of the optimal policy starting at state (i, j) taken over the set of all non-anticipatory and preemptive scheduling policies; $\nu = \lambda_1 + \lambda_2 + \mu_{11} + \mu_{12} + \mu_{22}$ is the rate used in uniformizing the continuous-time Markov chain; and θ is the optimal average cost. Equation (1) is also known as the average cost optimality equation. Assume throughout $\mu_{11} > \mu_{12}$. Refer to [3] for additional details.

Our general approach consists of deriving structural properties of the value functions of the optimal policy $J_{i,j}$ from the optimality equation (1), and proving structural properties of the optimal policy under different model parameter values. We consider each in turn.

2.1 Structural properties of value functions

Define $\Delta_{i,j}^1 \triangleq J_{i,j} - J_{i-1,j}$, $\Delta_{i,j}^2 \triangleq J_{i,j} - J_{i,j-1}$. From the objective function Q^*, we have a linear holding cost $g_{i,j}(x_{i,j}) = c_1 i + c_2 j$. This together with the optimality equation (1) and direct calculations yield for states $(1, 0)$, $(i, 0)$, $i \geq 2$, $(0, j)$, $j \geq 1$, $(1, 1)$

$$\begin{aligned}
\theta &= c_1 + \lambda_1 \Delta_{2,0}^1 + \lambda_2 \Delta_{1,1}^2 + \min_{x_{1,0}} \{(\mu_{11} - \mu_{12}) \Delta_{1,0}^1 x_{1,0}\} \\
\theta &= c_1 i + \lambda_1 \Delta_{i+1,0}^1 + \lambda_2 \Delta_{i,1}^2 - M_{11}^{12} \Delta_{i,0}^1 + \min_{x_{i,0}} \{-\mu_{12} \Delta_{i,0}^1 x_{i,0}\} \\
\theta &= c_2 j + \lambda_1 \Delta_{1,j}^1 + \lambda_2 \Delta_{0,j+1}^2 + \min_{x_{0,j}} \{-\mu_{22} \Delta_{0,j}^2 x_{0,j}\} \\
\theta &= c_1 + c_2 + \lambda_1 \Delta_{2,1}^1 + \lambda_2 \Delta_{1,2}^2 - \mu_{12} \Delta_{1,1}^1 \\
&\qquad + \min_{x_{1,1}} \{((\mu_{12} - \mu_{11}) \Delta_{1,1}^1 - \mu_{22} \Delta_{1,1}^2) x_{1,1}\}
\end{aligned}$$

where $M_{ab}^{cd} = \mu_{ab} + \mu_{cd}$ and without loss of generality, we set

$J_{0,0} = 0$. It is easy to show that $\Delta^1_{i,0} \geq 0$, $\Delta^2_{0,j} \geq 0$, $\Delta^1_{1,1} \geq 0$ and $\Delta^2_{1,1} \geq 0$; further recall $\mu_{11} > \mu_{12}$. Therefore, the above minimums are achieved with $x_{i,0} = 0, \forall i \geq 1$, $x_{0,j} = 1, \forall j \geq 1$, and $x_{1,1} = 1$, independent of the model parameter values.

In light of these general results, we use the value iteration method to generate a sequence of value functions $J^k_{i,j}$, starting from $J^0_{i,j} = 0$, together with a collection of induction and other arguments to derive various structural properties of the value functions of the optimal policy $J_{i,j}$ for all remaining states from the optimality equation (1). We now concentrate on a particular set of results and refer to [1] for our more general results and derived structural properties.

From the linear holding cost function, the optimality equation (1), the above results and direct calculations, we have for all states

$$\nu J^{k+1}_{1,0} = c_1 + \Lambda^{J,k}_{1,0} + M^{22}_{12} J^k_{1,0} + \mu_{11} J^k_{0,0} - \theta \quad (2)$$

$$\nu J^{k+1}_{i,0} = c_1 i + \Lambda^{J,k}_{i,0} + M^{12}_{11} J^k_{i-1,0} + \mu_{22} J^k_{i,0} - \theta, \quad i \geq 2 \quad (3)$$

$$\nu J^{k+1}_{0,j} = c_2 j + \Lambda^{J,k}_{0,j} + \mu_{22} J^k_{0,j-1} + M^{12}_{11} J^k_{0,j} - \theta, \quad j \geq 1 \quad (4)$$

$$\nu J^{k+1}_{1,j} = c_1 + c_2 j + \Lambda^{J,k}_{1,j} + \mu_{11} J^k_{0,j} + \mu_{12} J^k_{1,j} + \mu_{22} J^k_{1,j-1} - \theta, \quad j \geq 1 \quad (5)$$

$$\nu J^{k+1}_{i,j} = c_1 i + c_2 j + \Lambda^{J,k}_{i,j} + M^{12}_{11} J^k_{i-1,j} + \mu_{22} J^k_{i,j} - h^k_{i,j} - \theta \quad (6)$$

where $\Lambda^{J,k}_{i,j} = \lambda_1 J^k_{i+1,j} + \lambda_2 J^k_{i,j+1}$ for all i, j, $h^k_{i,j} = \max_{x^k_{i,j}} x^k_{i,j} f^k_{i,j}$, $f^k_{i,j} = \mu_{22}\Delta^{2,k}_{i,j} - \mu_{12}\Delta^{1,k}_{i,j}$ for all $i, j \geq 1$. Here we have defined the $f_{i,j}$ so that, from (2) – (6), the decision rule of the optimal policy is $x_{i,j} = 1 \Leftrightarrow f_{i,j} \geq 0$ and $x_{i,j} = 0 \Leftrightarrow f_{i,j} < 0$ for all $(i,j) \in \mathcal{S} \equiv \{(i,j) : i \geq 1, j \geq 1, (i,j) \neq (1,1)\}$, regardless of the model parameter values. Note that $x_{i,j} = 1$ means server 2 focuses on serving queue 2 when non-empty, whereas $x_{i,j} = 0$ means server 2 focuses on serving queue 1 when non-empty.

First consider the states (i,j), $i \geq 3, j \geq 2$, for which we obtain

$$\nu\Delta^{1,k+1}_{i,j} = c_1 + \Lambda^{\Delta^{1,k}}_{i,j} + M^{12}_{11}\Delta^{1,k}_{i-1,j} + \mu_{22}\Delta^{1,k}_{i,j} - h^k_{i,j} + h^k_{i-1,j} \quad (7)$$

$$\nu\Delta^{2,k+1}_{i,j} = c_2 + \Lambda^{\Delta^{2,k}}_{i,j} + M^{12}_{11}\Delta^{2,k}_{i-1,j} + \mu_{22}\Delta^{2,k}_{i,j} - h^k_{i,j} + h^k_{i,j-1} \quad (8)$$

where $\Lambda^{\Delta^{\ell,k}}_{i,j} = \lambda_1\Delta^{\ell,k}_{i+1,j} + \lambda_2\Delta^{\ell,k}_{i,j+1}$. Multiplying (8) by μ_{22} and (7) by μ_{12} and taking the difference yields

$$\begin{aligned}\nu f^{k+1}_{i,j} = \eta + \Lambda^{f,k}_{i,j} + (\mu_{11}+\mu_{12}) f^k_{i-1,j} + \mu_{22} f^k_{i,j} \\ -(\mu_{22}-\mu_{12}) h^k_{i,j} + \mu_{22} h^k_{i,j-1} - \mu_{12} h^k_{i-1,j}\end{aligned} \quad (9)$$

where $\eta = \mu_{22}c_2 - \mu_{12}c_1$ and $\Lambda^{f,k}_{i,j} = \lambda_1 f^k_{i+1,j} + \lambda_2 f^k_{i,j+1}$.

Next consider the states $(i,1)$, $i \geq 3$, for which we have

$$\nu\Delta^{1,k+1}_{i,1} = c_1 + \Lambda^{\Delta^{1,k}}_{i,1} + M^{12}_{11}\Delta^{1,k}_{i-1,1} + \mu_{22}\Delta^{1,k}_{i,1} - h^k_{i,1} + h^k_{i-1,1} \quad (10)$$

$$\nu\Delta^{2,k+1}_{i,1} = c_2 + \Lambda^{\Delta^{2,k}}_{i,1} + M^{12}_{11}\Delta^{2,k}_{i-1,1} + \mu_{22}\Delta^{2,k}_{i,1} - h^k_{i,1}. \quad (11)$$

Again, multiply (11) by μ_{22}, (10) by μ_{12} and take the difference:

$$\begin{aligned}\nu f^{k+1}_{i,1} = \eta + \Lambda^{f,k}_{i,1} + (\mu_{11}+\mu_{12}) f^k_{i-1,1} + \mu_{22} f^k_{i,1} \\ -(\mu_{22}-\mu_{12}) h^k_{i,1} - \mu_{12} h^k_{i-1,1}.\end{aligned} \quad (12)$$

Similarly considering the states $(2,j)$, $j \geq 2$, we derive

$$\nu\Delta^{1,k+1}_{2,j} = c_1 + \Lambda^{\Delta^{1,k}}_{2,j} + \mu_{11}\Delta^{1,k}_{1,j} + \mu_{22}(\Delta^{1,k}_{2,j} + \Delta^{2,k}_{1,j}) - h^k_{2,j} \quad (13)$$

$$\nu\Delta^{2,k+1}_{2,j} = c_2 + \Lambda^{\Delta^{2,k}}_{2,j} + M^{12}_{11}\Delta^{2,k}_{1,j} + \mu_{22}\Delta^{2,k}_{2,j} - h^k_{2,j} + h^k_{2,j-1} \quad (14)$$

$$\begin{aligned}\nu f^{k+1}_{2,j} = \eta + \Lambda^{f,k}_{2,j} + \mu_{11} f^k_{1,j} + \mu_{22} f^k_{2,j} - (\mu_{22}-\mu_{12}) h^k_{2,j} \\ +\mu_{22} h^k_{2,j-1}.\end{aligned} \quad (15)$$

For the state $(2,1)$, $\nu\Delta^{1,k+1}_{2,1}$ is given by (13) with $j = 1$, whereas

$$\nu\Delta^{2,k+1}_{2,1} = c_2 + \Lambda^{\Delta^{2,k}}_{2,1} + M^{12}_{11}\Delta^{2,k}_{1,1} + \mu_{22}\Delta^{2,k}_{2,1} - h^k_{2,1}, \quad (16)$$

from which we obtain

$$\nu f^{k+1}_{2,1} = \eta + \Lambda^{f,k}_{2,1} + \mu_{11} f^k_{1,1} + \mu_{22} f^k_{2,1} - (\mu_{22}-\mu_{12}) h^k_{2,1}. \quad (17)$$

Finally, analogously considering the states $(1,j)$, $j \geq 2$, yields

$$\nu\Delta^{1,k+1}_{1,j} = c_1 + \Lambda^{\Delta^{1,k}}_{1,j} + \mu_{12}\Delta^{1,k}_{1,j} + \mu_{22}\Delta^{1,k}_{1,j-1} \quad (18)$$

$$\nu\Delta^{2,k+1}_{1,j} = c_2 + \Lambda^{\Delta^{2,k}}_{1,j} + \mu_{11}\Delta^{2,k}_{0,j} + \mu_{12}\Delta^{2,k}_{1,j} + \mu_{22}\Delta^{2,k}_{1,j-1} \quad (19)$$

$$\nu f^{k+1}_{1,j} = \eta + \Lambda^{f,k}_{1,j} + \mu_{11}\mu_{22}\Delta^{2,k}_{0,j} + \mu_{12} f^k_{1,j} + \mu_{22} f^k_{1,j-1}. \quad (20)$$

2.2 Structural properties of optimal policies

We now turn to establish structural properties of the optimal scheduling policy under different model parameters based on the results in §2.1. Let us first consider the case $c_2\mu_{22} \geq c_1\mu_{12}$, for which we obtain one of our main results showing $c\mu$ to be optimal.

THEOREM 1. *In the regime $c_2\mu_{22} \geq c_1\mu_{12}$, the optimal scheduling policy is for server 2 to always serve queue 2 while it is non-empty, i.e., the $c\mu$ policy according to $c_2\mu_{22} \geq c_1\mu_{12}$ is optimal.*

PROOF. Since we have $x_{i,0} = 0$ for all $i \geq 1$, $x_{0,j} = 1$ for all $j \geq 1$ and $x_{1,1} = 1$ from the results of the previous section, it suffices to prove that $f_{i,j} \geq 0$ for all $(i,j) \in \mathcal{S}$, which we establish by induction: Assume $f^k_{i,j} \geq 0$ and show that $f^{k+1}_{i,j} \geq 0$.

Starting from the induction hypothesis, we have $x^k_{i,j} = 1$ and thus $h^k_{i,j} = f^k_{i,j}$, for all $(i,j) \in \mathcal{S}$. Upon substituting this result into (9) and (15), and then simplifying we obtain, with $\eta \geq 0$,

$$\nu f^{k+1}_{i,j} = \eta + \Lambda^{f,k}_{i,j} + \mu_{11} f^k_{i-1,j} + \mu_{12} f^k_{i,j} + \mu_{22} f^k_{i,j-1} \geq 0, \quad (21)$$

where the inequality follows from the induction hypothesis. Similarly, substituting $h^k_{i,j} = f^k_{i,j}$ in (12),(17) and simplifying yields

$$\nu f^{k+1}_{i,1} = \eta + \Lambda^{f,k}_{i,1} + \mu_{11} f^k_{i-1,1} + \mu_{12} f^k_{i,1} \geq 0, \quad (22)$$

$$\nu f^{k+1}_{2,1} = \eta + \Lambda^{f,k}_{2,1} + \mu_{11} f^k_{1,1} + \mu_{12} f^k_{2,1} \geq 0, \quad (23)$$

with the inequalities following from the induction hypothesis. By the induction hypothesis and since $\Delta^{2,k}_{0,j} \geq 0$, we have from (20)

$$\nu f^{k+1}_{1,j} = \eta + \Lambda^{f,k}_{1,j} + \mu_{11}\mu_{22}\Delta^{2,k}_{0,j} + \mu_{12} f^k_{1,j} + \mu_{22} f^k_{1,j-1} \geq 0. \quad (24)$$

To complete the induction proof we need only assert that the hypothesis trivially holds when $k = 0$ given our choice of $J^0_{i,j} = 0 \Rightarrow f^0_{i,j} = 0$ for all i and j. Taking the limit as $k \to \infty$ yields the desired result that $f_{i,j} \geq 0$ for all $(i,j) \in \mathcal{S}$, since $\lim_{k\to\infty} J^k_{i,j} = J_{i,j}$. This then implies $x_{i,j} = 1$ for all $(i,j) \in \mathcal{S}$, which together with $x_{0,j} = 1$ and $x_{1,1} = 1$ is equivalent to $c\mu$ when $c_2\mu_{22} \geq c_1\mu_{12}$, and thus completes the proof. □

Turning to the case $c_2\mu_{22} < c_1\mu_{12}$, we exploit a larger set of structural properties for the value functions of the optimal policy, including another collection of induction hypotheses and various induction and other arguments, in order to prove structural properties of the optimal policy when $c_2\mu_{22} < c_1\mu_{12}$. Due to space restrictions, we omit the techncal details and briefly summarize some of our main results, refering the reader to [1] for more information. Namely, our analysis demonstrates that, when $c_2\mu_{22} < c_1\mu_{12}$, the optimal scheduling policy is state-dependent with switching curves that involve complex two-dimensional structures whose shapes are functionals of the corresponding stationary distribution, and thus dependent upon the model parameters including traffic intensity.

3. NUMERICAL EXPERIMENTS

Our results in §2 for the base two-server stochastic network model with parameters $c_2\mu_{22} \geq c_1\mu_{12}$ show that the $c\mu$ policy is optimal. We therefore consider the stochastic network with parameters

$c_2\mu_{22} < c_1\mu_{12}$, for which our theoretical results show that the optimal policy is generally state-dependent. To explore our results in §2 on the various structural properties of the corresponding switching curve, we used the standard value iteration algorithm to analyze the stochastic dynamic program under various scenarios and fixed buffer sizes n_1, n_2. As a representative example, Fig. 1(a) and Fig. 1(c) illustrate the optimal scheduling regions for the parameter settings $c_1 = 1, c_2 = 0.45, \mu_{11} = \mu_{22} = 1, \mu_{12} = 0.5$ with $\lambda_1 = \lambda_2 = 0.25$ (light traffic), $n_1 = n_2 = 3000$ and $\lambda_1 = \lambda_2 = 0.9$ (heavy traffic), $n_1 = 1800$, $n_2 = 3600$, respectively. The shaded region in these plots corresponds to the values of queue lengths for which server 2 should serve queue 2 under the optimal policy, i.e., $x_{i,j} = 1$, whereas outside of this region server 2 follows the $c\mu$ policy, i.e., $x_{i,j} = 0$. Notice that in both cases there is a degree of symmetry between the two server-queue pairs because we have chosen $\lambda_1 = \lambda_2,\ \mu_{11} = \mu_{22}$.

The plots in Fig. 1(a) and Fig. 1(c) illustrate that the optimal policy is indeed state-dependent in a more complex manner than a policy based on a fixed, one-dimensional threshold. Further, a clear qualitative difference can be seen between the two traffic settings which can be understood from our theoretical results on the trade-off that the optimal policy affects between minimizing cost and balancing load. When the traffic intensity is light, relatively small load imbalance between the queues can have a considerable impact on the objective function which makes it more important to balance the load and thus the point at which server 2 switches from serving queue 2 to helping queue 1 exhibits an almost linear increase, up to a given level before flattening. Conversely, when the traffic intensity is heavy, it is always better for server 2 to help queue 1 in order to minimize costs, unless queue 1 relative to queue 2 drops below a two-dimensional curve in which case server 2 is required by the optimal policy to serve queue 2. Observe that the value of Q_2 at which this switching curve becomes flat decreases as the traffic intensity increases while the switching value of Q_1 for small Q_2 tends to increase with the traffic intensity, suggesting that the switching curve becomes a one-dimensional threshold in the limit as the traffic intensity goes to 1, which is consistent with [2]. Moreover, the levels at which these switching curves flatten are functionals of the stationary distribution. Finally, the shape of the optimal regions for Q_2 approaching n_2 is the result of boundary effects due to the finite buffers (absent in Fig. 1(c) since $n_2 = 3600$), which is why our numerical experiments examine a wide range of values for n_1, n_2.

These numerical results together with our theoretical results on the various structural properties of the optimal policy suggest that we should also analyze the corresponding switching curve as a function of $(Q_1/Q_2, Q_2)$. Fig. 1(b) and Fig. 1(d) illustrate these relative switching curves for the scheduling regions in Fig. 1(a) and Fig. 1(c), respectively, where the x-axis is normalized by n_2 in order to examine the trends of the optimal relative switching curves as buffer sizes n_1, n_2 increase. Our results in §2 further suggest that the characteristics of the relative switching curves converge to a limit as $Q_1, Q_2 \to \infty$, which is consistent with the numerical solutions of the stochastic dynamic program for larger and larger values of n_1, n_2. In the infinite buffer limit, we can show that for large backlogs in queue 2, indeed a less complex threshold-based policy is optimal; e.g., in Fig. 1(d), the relative switching curves appear to tend to a limit with a $1/Q_2$-trend. The structures of these switching curves involve functionals of the corresponding stationary distribution. These results and trends also have important connections with previous limiting regime results [2, 7]. Note that the deviating shape of the optimal relative switching curves for Q_2/n_2 very close to 1 is the result of boundary effects due to the finite buffers, further illustrating important differences among finite and infinite buffer stochastic networks.

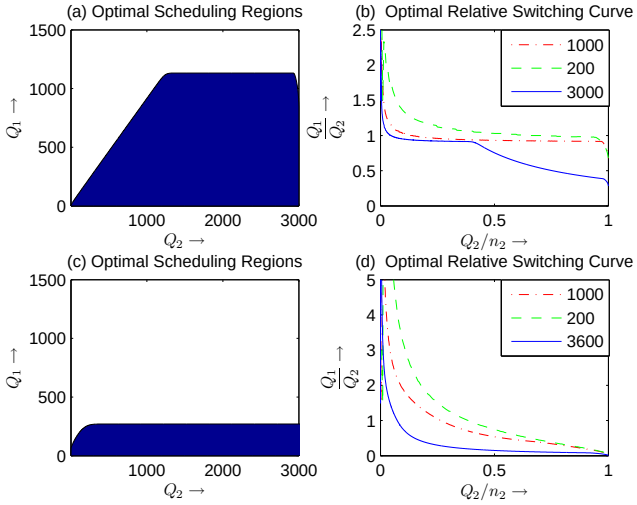

Figure 1: Regions (a,c) and switching curves (b,d) for the optimal policy under light (a,b) and heavy (c,d) traffic.

We note that while the optimal scheduling regions and the optimal relative switching curves can vary considerably under different parameter values $\lambda_k, \mu_{11}, \mu_{22}, \mu_{12}, c_k$, the above structural properties, basic results and trends continue to hold. Similarly, when $S_\ell > 1$, we continue to have an analogous $c\mu$ result in the general multi-server stochastic network for the states $(i,j) \in \mathcal{S} \equiv \{(i,j) : i \geq S_1, j \geq S_2, (i,j) \neq (S_1, S_2)\}$, where the boundary states are handled directly in a manner analogous to that of §2. In the remaining cases, we continue to have a state-dependent optimal scheduling policy with switching curves that involve complex two-dimensional structures which are functionals of the corresponding stationary distribution, and thus dependent upon the model parameters such as the traffic intensity.

4. REFERENCES

[1] M. Bayati, M. Sharma, M. S. Squillante. Optimal scheduling in a multiserver stochastic network. Technical Report, 2006.

[2] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Prob.*, 11:608–649, 2001.

[3] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volume II*. Athena Scientific, 2nd edition, 2001.

[4] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *M&SOM*, 5:79–141, 2003.

[5] M. S. Squillante and E. D. Lazowska. Using processor-cache affinity information in shared-memory multiprocessor scheduling. *IEEE Trans. Par. Dist. Sys.*, 4(2):131–143, 1993.

[6] M. S. Squillante and R. D. Nelson. Analysis of task migration in shared-memory multiprocessors. *Proc. ACM SIGMETRICS Conf. Meas. and Mod. Comp. Sys.*, pp. 143–155, 1991.

[7] M. S. Squillante, C. H. Xia, D. D. Yao, and L. Zhang. Threshold-based priority policies for parallel-server systems with affinity scheduling. *Proc. Am. Cont. Conf.*, pp. 2992–2999, 2001.