On the Improvement from Scheduling a Two-Station Queueing Network in Heavy Traffic

by

Jihong Ou and Lawrence M. Wein

OR 208-90 January 1990

On the Improvement from Scheduling a Two-Station Queueing Network in Heavy Traffic

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

Abstract

For a two-station multiclass queueing network in heavy traffic, we assess the improvement from scheduling (job release and priority sequencing) that can occur relative to Poisson input and first-come first-served (FCFS) sequencing. In particular, simple upper bounds are derived on the optimal objective function value (found in Wein 1989a) of a Brownian control problem that approximates (via Harrison's 1988 model) a two-station queueing network scheduling problem in heavy traffic. When the system is perfectly balanced, the Brownian analysis predicts that optimal scheduling will reduce the long run expected average number of customers in the network *by at least a factor of four* relative to the Poisson input, FCFS sequencing policy that achieves the same throughput rate. When the system is not perfectly balanced, the corresponding factor is slightly smaller than two.

January 1990

On the Improvement from Scheduling a Two-Station Queueing Network in Heavy Traffic

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

1. Introduction and Summary

In recent years, queueing networks have become a primary mathematical model of manufacturing systems, and thus the job-shop scheduling problem, for which there exists a vast literature, can be viewed as the problem of scheduling a multiclass queueing network. The two most important issues faced in scheduling problems are to find an effective scheduling policy, and to assess the improvement in performance that will take place. In particular, one would ideally like to find an optimal scheduling policy, and then measure the increase in performance of this policy relative to the currently used policy. By scheduling policy, we mean both a *customer release* policy (when are customers released into the network or shop) and a priority sequencing policy (which customer should be served next at each station in the network).

In this paper, we attempt to assess the improvement from scheduling that can occur relative to the commonly assumed policy of Poisson input, which represents a lack of systematic control over arrivals, and first-come first-served (FCFS) sequencing, which is often used in job shops. This assessment is made for a multiclass queueing network consisting of two single-server stations. Following Kelly's (1979) terminology, we consider a network populated by a variety of different *types* of customers, where each customer type has its own arbitrary route through the network. Then a different *class* of customer is defined for each combination of type and stage of completion along its route. In the context of a job-shop, each customer type corresponds to a different type of product that can be processed at the shop, and each customer class corresponds to a particular operation for a particular product type.

It will be assumed that each server has its own exponential service time distribution, and thus all customers, regardless of class, have the same service time distribution at a given station. Denote the desired *throughput rate* (number of departures per unit of time) of type j customers by λ_j, for $j = 1, ..., J$, and let $\bar{\lambda} = \sum_{j=1}^{J} \lambda_j$ be the total throughput rate. It is well known (see Baskett et al. 1975) that under the policy of Poisson input (customers of type j arrive according to independent Poisson processes with rate λ_j, for $j = 1, ..., J$) and FCFS sequencing, the long run expected average number of customers in the system is

$$L = \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}, \tag{1.1}$$

where ρ_i is the traffic intensity, or server utilization, for station i, which can be easily calculated from the arrival and service rates, and the routing information.

Unfortunately, there exist no exact results for scheduling two-station multiclass queueing networks. However, Harrison (1988) has developed a Brownian network model that allows one to approximate a queueing network scheduling problem by a control problem involving Brownian motion. This model was formulated under the *balanced heavy loading* conditions that there exists a large integer n such that

$$\sqrt{n}(1-\rho_i) \text{ is of moderate size for } i = 1, 2. \tag{1.2}$$

A representative example is to choose $n = 100$ when $\rho_1 = \rho_2 = .9$. The Brownian control problem is more tractable than its conventional counterpart, and Wein (1989a) has solved a Brownian control problem that approximates a scheduling problem for a more general version (for example, each class has their own general processing time distribution) of the two-station network described above. The scheduling problem was to choose a customer release policy (the timing was controllable, but the type of entering customer was deterministically chosen according to the fractions $\lambda_j, j = 1, ..., J$) and a priority sequencing policy (which class of customer to serve next at each station) to minimize the long run expected average number of customers in the network subject to the constraint that the long run expected average throughput rate was greater than or equal to $\bar{\lambda}$. In Wein (1989b), the solution to this Brownian control problem was interpreted in terms of the original queueing system in order to develop an effective scheduling policy. Readers are referred to that paper for a description of the motivating factory scheduling problem, and definitions of the workload regulating release policy and the workload balancing sequencing policy, which is based on dynamic reduced costs from a linear program.

This paper calculates simple upper bounds on the long run expected average number of customers in the network under the optimal solution of the idealized Brownian model, which we denote by L_{opt}. These bounds are calculated under the assumption that all service times are exponential. The main results are

$$L_{opt} \le \frac{\rho_1}{2(1-\rho_1)}, \quad \text{if } \rho_1 = \rho_2, \tag{1.3}$$

and

$$L_{opt} \le \max_{i=1,2}\left(\frac{\rho_i}{1-\rho_i}\right), \quad \text{if } \rho_1 \neq \rho_2. \tag{1.4}$$

These bounds offer both optimistic and pessimistic news with respect to the improvements provided by scheduling. On the optimistic side, inequalities (1.1) and (1.3) imply that the idealized Brownian solution cuts L by at least a factor of four relative to the Poisson, FCFS case. Thus scheduling can offer a dramatic improvement in system performance. When $\rho_1 \neq \rho_2$, (1.1), (1.2), and (1.4) suggest that L is cut by at least a factor that is slightly smaller than two. We believe the relative difference between these two cases has more to do with the quality of the derived bounds than with any inherent difficulty in obtaining scheduling improvements in the unbalanced case.

On the pessimistic side, we have not been able to eliminate the $(1-\rho)$ term in the denominator of (1.3) and (1.4), which is omnipresent in steady state queueing theoretic results for open networks. In fact, $L_{opt} = K/(1-\rho)$ for the balanced case, and the factor of K is bounded above by $\rho/2$ in the derivation to follow, leading to (1.3). Thus, it appears that scheduling can offer only a *linear* improvement in performance, and that scheduling is unable to prevent a queueing system from becoming unstable when $\rho \geq 1$.

The bound in (1.3) does not provide a rigorous justification of the existence of a scheduling policy that will reduce L by at least a factor of four. The model development in Harrison (1988) contains a persuasive verbal argument, but lacks a heavy traffic limit theorem claiming that a sequence of queueing network scheduling problems converges to the limiting Brownian control problem as the network approaches heavy traffic. Although an optimal solution to the two-station Brownian control problem is found in Wein (1989a), the interpretation of this solution in (1989b) is based on intuition gained from existing heavy traffic limit theorems, and no attempt is made to rigorously justify this interpretation with a weak convergence result. However, related weak convergence results have been obtained by Kushner and Ramachandran (1989), Martins and Kushner (1989), and Kushner and Martins (1989) in the context of queueing network control problems, where the service rates, arrival rates, and routing probabilities can be controlled. They prove the convergence of the controlled processes to a controlled reflected diffusion process, the convergence of the associated costs, and the convergence of the optimal value function of the queueing system to the optimal value function of the limit process.

We believe that bounds (1.3)-(1.4) offer a rough estimate for the impact that scheduling can have on two-station queueing networks satisfying the balanced heavy loading conditions (1.2). This belief is partially based on the fact that the slackness in the bound counteracts the possible inability of the proposed policy to achieve L_{opt} when the system is not in extremely heavy traffic. Although extensive numerical results have not been performed, a simulation study was undertaken in Wein (1989b) on a two-station network where $\rho_1 = \rho_2 = .9$, and the proposed scheduling policy achieved a mean sojourn time (with a 95% confidence interval) of $38.6(\pm 0.9)$ at a mean throughput rate of $\bar{\lambda} = .127(\pm .001)$. This network was not of product form under Poisson input and FCFS sequencing, because different customer classes possessed different exponential processing time distributions, and thus bound (1.3) does not necessarily apply. However, simulating Poisson input and FCFS sequencing for this problem yields a mean sojourn time of $159.0(\pm 7.0)$ at a mean throughput rate of $\bar{\lambda} = .127(\pm .000)$. Since, $159.0/38.6 = 4.12$, these simulation results certainly display the improvement in performance suggested by (1.3).

The next section reviews the result for L_{opt} in the more general network problem considered in Wein (1989a). In Section 3, inequalities (1.3) and (1.4) are derived under the exponential processing time assumption.

2. Optimal Performance in the Brownian Network

In this section, we state the results in Wein (1989a) for the optimal long run expected average number of customers in the network. Unlike the Poisson input, FCFS sequencing case in equation (1.1), L_{opt} depends on the detailed routing structure of the particluar network, and on the first and second moments of the various service time distributions. The queueing network is indexed by single-server stations $i = 1, 2$, customer types $j = 1, ..., J$, and customer classes $k = 1, ..., K$. Recall from Section 1 that the desired throughput rate of type j customers is λ_j, and $\bar{\lambda} = \sum_{j=1}^{J}$ is the desired total throughput rate. Since a different class of customer is defined for each combination of customer type and stage of completion, we have $K \geq J$. Furthermore, J of the K classes correspond to the first stage of some customer type's route. Let $q_k = \lambda_j/\bar{\lambda}$ if class k is the first stage of customer class j's route, for $j = 1, ..., J$, and let $q_k = 0$, otherwise. Then define $\lambda_k = q_k\bar{\lambda}$, so that λ_k represents the average number of class k customers that must flow through the network per unit of time in order to satisfy the throughput rate constraint.

Each customer class is served at a particular station $s(k)$, and has its own general service time distribution with mean m_k and variance s_k^2. Define the $2 \times K$ *resource consumption* matrix $A = (A_{ik})$ by

$$A_{ik} = \begin{cases} 1, & \text{if i = s(k),} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

A customer of class k, upon completion of service at station $s(k)$, turns into a class j customer with probability P_{kj}, and exits the network with probability $1 - \sum_{j=1}^{K} P_{kj}$, independent of all previous history. The $K \times K$ Markovian switching matrix $P = (P_{kj})$ has spectral radius less than one, so that all customers eventually exit the network. Notice that the matrix P consists of all zero and one entries, since we are assuming deterministic routes. The assumption of deterministic routes is only for expositional convenience; see Kelly (1979) and Harrison (1988) for the inclusion of probabilistic routing, for events such as rework or scrap. Define the $K \times K$ *input-output* matrix $R = (R_{kj})$ by

$$R_{kj} = m_j^{-1}(\delta_{jk} - P_{jk}), \tag{2.2}$$

where $\delta_{jk} = 1$ if $j = k$, and $\delta_{jk} = 0$ otherwise.

Since the routing matrix P is transient, the matrix R is invertible, and there exists a unique nonnegative $K-$vector $\beta = (\beta_k)$ satisfying the flow balance equations

$$\lambda = R\beta, \tag{2.3}$$

where $\lambda = (\lambda_1, ..., \lambda_K)$. We interpret β_k as the average fraction of time that server $s(k)$ must devote to serving class k customers in order to satisfy the throughput constraint with equality. Let the traffic intensities (ρ_1, ρ_2) be defined by

$$\rho_i = \sum_{k=1}^{K} A_{ik}\beta_k, \quad \text{for} \quad i = 1, 2, \tag{2.4}$$

so that ρ_i is the average utilization of server i if the throughput rate constraint holds with equality. Thus, the server utilizations in (2.4) are precisely the same values as appear in (1.1) for the Poisson, FCFS case.

Now define the $2 \times K$ workload profile matrix $M = (M_{ik})$ by

$$M = AR^{-1}. \tag{2.5}$$

The value of M_{ik} is interpreted as the expected remaining processing time at station i for a customer of class k until that customer exits the network. Without loss of generality, suppose the customer classes are indexed so that

$$\max_{1 \le k \le K} (\rho_2 M_{1k} - \rho_1 M_{2k}) = \rho_2 M_{11} - \rho_1 M_{21} > 0, \tag{2.6}$$

and

$$\min_{1 \le k \le K} (\rho_2 M_{1k} - \rho_1 M_{2k}) = \rho_2 M_{12} - \rho_1 M_{22} < 0. \tag{2.7}$$

Now define the positive coefficients h_1 and h_2 by

$$h_1 = \frac{1}{\rho_1 M_{22} - \rho_2 M_{12}}, \tag{2.8}$$

and

$$h_2 = \frac{1}{\rho_2 M_{11} - \rho_1 M_{21}}. \tag{2.9}$$

Finally, recalling that n is the system parameter defined in the balanced heavy loading conditions (1.2), define

$$\mu = \sqrt{n}(\rho_1 - \rho_2), \tag{2.10}$$

$$\xi = \sqrt{n}\rho_1(1 - \rho_1), \tag{2.11}$$

$$\varrho = \begin{bmatrix} \rho_2 \\ -\rho_1 \end{bmatrix}, \tag{2.12}$$

and the $K \times K$ matrix $\Sigma = (\Sigma_{jl})$, where

$$\Sigma_{jl} = \sum_{k=1}^{K} [\beta_k m_k^{-1} P_{kj}(\delta_{jl} - P_{kl}) + \beta_k m_k^{-1} s_k^2 R_{jk} R_{lk}]. \tag{2.13}$$

The matrix Σ is the covariance matrix of a K-dimensional Brownian motion process imbedded in the limiting Brownian control problem derived in Harrison (1988). We should note that Harrison presents two possible versions of the covariance matrix Σ, and (2.13) is the more "refined" version proposed in equation (11.6) of that paper. Finally, the $K-$dimensional Brownian motion is reduced to a one-dimensional Brownian motion that has drift μ in (2.10) and variance σ^2, where

$$\sigma^2 = \varrho^T M \Sigma M^T \varrho. \tag{2.14}$$

To repeat, the queueing network scheduling problem is to dynamically release customers (subject to a specified entering class mix) and sequence customers in a two-station queueing network to minimize the long run expected average number of customers in the network subject to achieving a long run expected average throughput rate of at least $\bar{\lambda}$. The optimal objective function value in the approximating Brownian control problem is denoted by $f(a^*)$ in Wein (1989a), and by (7.3), (7.19), and (7.20) of that paper,

$$\begin{aligned} f(a^*) = (\frac{\sigma^2}{2\mu(\rho_1 - \rho_2)}) \Big[& h_1\rho_2(1-\rho_1)\ln\left(\frac{(h_1+h_2)\rho_2(1-\rho_1)}{h_1\rho_2(1-\rho_1)+h_2\rho_1(1-\rho_2)}\right) \\ & + h_2\rho_1(1-\rho_2)\ln\left(\frac{(h_1+h_2)\rho_1(1-\rho_2)}{h_1\rho_2(1-\rho_1)+h_2\rho_1(1-\rho_2)}\right)\Big], \end{aligned} \tag{2.15}$$

if $\rho_1 \neq \rho_2$. By (9.14) of Wein (1989a), it follows that

$$f(a^*) = \frac{\sigma^2 h_1 h_2}{4\xi(h_1+h_2)} \quad \text{if } \rho_1 = \rho_2. \tag{2.16}$$

The Brownian approximation is based on a rescaling of the basic processes by the system parameter n. In particular, if $Q_k(t)$ is the number of class k customers in the original queueing network at time t, then $Z_k(t)$ is the number of class k customers in the Brownian network at time t, where

$$Z_k(t) = \frac{Q_k(nt)}{\sqrt{n}}, \quad \text{for } t \geq 0. \tag{2.17}$$

Since the objective function leading to $f(a^*)$ was in terms of a long run average criterion, it follows that

$$L_{opt} = \sqrt{n} f(a^*). \tag{2.18}$$

This is the long run expected average number of customers in the queueing system under an optimal release and sequencing policy, as predicted by the Brownian model.

3. Derivation of the Upper Bounds

In this section, we specialize the results in Section 2 to the case where the service time distributions are exponential. The results do not rely on the fact that each class at a given station has the same service time distribution. However, this assumption is needed in order to obtain a product form network, and hence (1.1). We start by simplifying the variance σ^2.

Lemma 1. *If all service time distributions are exponential, then*

$$\sigma^2 = 2\sum_{k=1}^{K}\beta_k(\rho_2 M_{1k} - \rho_1 M_{2k})(\rho_2 A_{1k} - \rho_1 A_{2k}) - \sum_{k=1}^{K}\lambda_k(\rho_1 M_{2k} - \rho_2 M_{1k})^2. \tag{3.1}$$

Proof. By (2.2) and (2.13), Σ can be written as

$$\Sigma_{jl} = \sum_{k=1}^{K}[\beta_k m_k^{-1} P_{kj}(\delta_{jl} - P_{kl}) + \beta_k m_k^{-1} s_k^2 m_k^{-1}(\delta_{kj} - P_{kj}) m_k^{-1}(\delta_{kl} - P_{kl})]. \tag{3.2}$$

Since the service times are exponential, $s_k^2 = m_k^2$ for $k = 1, ..., K$, and

$$\Sigma_{jl} = \sum_{k=1}^{K}[\beta_k m_k^{-1} P_{kj}(\delta_{jl} - P_{kl}) + \beta_k m_k^{-1}(\delta_{kj} - P_{kj})(\delta_{kl} - P_{kl})]. \tag{3.3}$$

Let $\Gamma = M\Sigma M^T$, so that for $i, j = 1, 2$,

$$\Gamma_{ij} = \sum_{s=1}^{K}\sum_{t=1}^{K} M_{is}\Sigma_{st}M_{jt}, \tag{3.4}$$

$$= \sum_{s=1}^{K}\sum_{t=1}^{K} M_{is}\sum_{k=1}^{K}\beta_k m_k^{-1}[P_{ks}(\delta_{st} - P_{kt}) + (\delta_{ks} - P_{ks})(\delta_{kt} - P_{kt})]M_{jt}, \tag{3.5}$$

$$= \sum_{k=1}^{K}\beta_k m_k^{-1}\sum_{s=1}^{K}\sum_{t=1}^{K} M_{is}P_{ks}(\delta_{st} - P_{kt})M_{jt}$$

$$+ \sum_{k=1}^{K}\beta_k m_k^{-1}\Big[\sum_{s=1}^{K} M_{is}(\delta_{ks} - P_{ks})\Big]\Big[\sum_{t=1}^{K} M_{jt}(\delta_{kt} - P_{kt})\Big]. \tag{3.6}$$

Observe that (2.5) implies

$$\sum_{s=1}^{K} M_{is}(\delta_{ks} - P_{ks}) = A_{ik} m_k \ \text{ for } i = 1,2, \ \text{ and } k = 1, ..., K. \tag{3.7}$$

Thus, the second term on the right side of (3.6) is

$$\sum_{k=1}^{K} \beta_k m_k^{-1} \Big[\sum_{s=1}^{K} M_{is}(\delta_{ks} - P_{ks})\Big] \Big[\sum_{t=1}^{K} M_{jt}(\delta_{kt} - P_{kt})\Big] = \sum_{k=1}^{K} \beta_k m_k A_{ik} A_{jk}, \tag{3.8}$$

which equals zero by (2.1). The first term on the right side of (3.6) can be expressed as

$$\sum_{k=1}^{K} \beta_k m_k^{-1} \sum_{s=1}^{K}\sum_{t=1}^{K} M_{is} P_{ks} (\delta_{st} - P_{kt}) M_{jt} \tag{3.9}$$

$$= \sum_{k=1}^{K} \beta_k m_k^{-1} \Big[\sum_{s=1}^{K} M_{is} P_{ks} M_{js} - \sum_{s=1}^{K}\sum_{t=1}^{K} M_{is} P_{ks} P_{kt} M_{jt}\Big], \tag{3.10}$$

$$= \sum_{k=1}^{K} \beta_k m_k^{-1} \Big[\sum_{s=1}^{K} M_{is} P_{ks} M_{js} - (M_{ik} - A_{ik} m_k)(M_{jk} - A_{jk} m_k)\Big], \ \text{ by (3.7)}, \tag{3.11}$$

$$= \sum_{k=1}^{K} \beta_k (M_{ik} A_{jk} + M_{jk} A_{ik}) + \sum_{k=1}^{K} M_{ik} \Big[\sum_{s=1}^{K} \beta_s m_s^{-1} P_{sk} - \beta_k m_k^{-1}\Big] M_{jk}$$
$$- \sum_{k=1}^{K} \beta_k m_k A_{ik} A_{jk}, \tag{3.12}$$

$$= \sum_{k=1}^{K} \beta_k (M_{ik} A_{jk} + M_{jk} A_{ik}) - \sum_{k=1}^{K} M_{ik} \lambda_k M_{jk}, \quad \text{by (2.1) and (2.3).} \tag{3.13}$$

Thus,

$$\Gamma_{ij} = \sum_{k=1}^{K} \beta_k (M_{ik} A_{jk} + M_{jk} A_{ik}) - \sum_{k=1}^{K} M_{ik} \lambda_k M_{jk}, \ \text{ for } i, j = 1, 2. \tag{3.14}$$

Finally, $\sigma^2 = \varrho^T \Gamma \varrho$, by (2.14) and the definition of Γ, so

$$\sigma^2 = 2\rho_2^2 \sum_{k=1}^{K} \beta_k A_{1k} M_{1k} - \rho_2^2 \sum_{k=1}^{K} \lambda_k M_{1k}^2 - \rho_1\rho_2 \sum_{k=1}^{K} \beta_k (M_{1k} A_{2k} + M_{2k} A_{1k})$$
$$+ \rho_1\rho_2 \sum_{k=1}^{K} \lambda_k M_{1k} M_{2k} + 2\rho_1^2 \sum_{k=1}^{K} \beta_k A_{2k} M_{2k} - \rho_1^2 \sum_{k=1}^{K} \lambda_k M_{2k}^2$$
$$- \rho_1\rho_2 \sum_{k=1}^{K} \beta_k (M_{1k} A_{2k} + M_{2k} A_{1k}) + \rho_1\rho_2 \sum_{k=1}^{K} \lambda_k M_{1k} M_{2k}. \tag{3.15}$$

Collecting terms yields (3.1) ∎

In the next lemma, the quantity σ^2 is bounded from above. First, let us define

$$\hat{M}_k = \rho_2 M_{1k} - \rho_1 M_{2k}, \quad \text{for} \quad k = 1, ..., K. \tag{3.16}$$

Lemma 2.

$$\sigma^2 \leq 2\rho_1\rho_2(\hat{M}_1 - \hat{M}_2). \tag{3.17}$$

Proof.

$$\sigma^2 = 2\sum_{k=1}^{K} \beta_k \hat{M}_k \rho_2 A_{1k} - 2\sum_{k=1}^{K} \beta_k \hat{M}_k \rho_1 A_{2k} - \sum_{k=1}^{K} \lambda_k \hat{M}_k^2, \text{ by (3.1) and (3.16)}, \tag{3.18}$$

$$\leq 2\rho_2 \sum_{k=1}^{K} \beta_k A_{1k} \hat{M}_k - 2\rho_1 \sum_{k=1}^{K} \beta_k A_{2k} \hat{M}_k, \tag{3.19}$$

$$\leq 2\rho_2 \hat{M}_1 \sum_{k=1}^{K} \beta_k A_{1k} - 2\rho_1 \hat{M}_2 \sum_{k=1}^{K} \beta_k A_{2k}, \quad \text{by} \quad (2.6) - (2.7), \tag{3.20}$$

$$\leq 2\rho_1\rho_2(\hat{M}_1 - \hat{M}_2), \quad \text{by} \quad (2.4). \quad \blacksquare \tag{3.21}$$

Now we are ready to prove inequalities (1.3) and (1.4).

Proposition 3.

$$L_{opt} \leq \frac{\rho_1}{2(1-\rho_1)}, \quad \text{if} \quad \rho_1 = \rho_2. \tag{3.22}$$

Proof. By (2.11), (2.16), and (2.18), we have

$$L_{opt} = \frac{\sigma^2 h_1 h_2}{4\rho_1(1-\rho_1)(h_1+h_2)}, \quad \text{if} \quad \rho_1 = \rho_2. \tag{3.23}$$

By (2.8)-(2.9),

$$\frac{h_1 h_2}{h_1 + h_2} = \frac{1}{\hat{M}_1 - \hat{M}_2}. \tag{3.24}$$

Thus, by Lemma 2 and (3.24),

$$L_{opt} \leq \frac{2\rho_1^2(\hat{M}_1 - \hat{M}_2)}{4\rho_1(1-\rho_1)(\hat{M}_1 - \hat{M}_2)}, \tag{3.25}$$

$$= \frac{\rho_1}{2(1-\rho_1)}. \quad \blacksquare \tag{3.26}$$

Proposition 4.

$$L_{opt} \le \max_{i=1,2}\Big(\frac{\rho_i}{1-\rho_i}\Big), \quad \text{if} \quad \rho_1 \ne \rho_2. \tag{3.27}$$

Proof. By (2.10), (2.15), and (2.18), it follows that

$$\begin{aligned} L_{opt} = \frac{\sigma^2}{2(\rho_1-\rho_2)^2}\Bigg[& h_1\rho_2(1-\rho_1)\ln\left(\frac{(h_1+h_2)\rho_2(1-\rho_1)}{h_1\rho_2(1-\rho_1)+h_2\rho_1(1-\rho_2)}\right) \\ & + h_2\rho_1(1-\rho_2)\ln\left(\frac{(h_1+h_2)\rho_1(1-\rho_2)}{h_1\rho_2(1-\rho_1)+h_2\rho_1(1-\rho_2)}\right)\Bigg]. \end{aligned} \tag{3.28}$$

By (2.8)-(2.9), this can be expressed as

$$\begin{aligned} L_{opt} = \frac{\sigma^2}{2(\rho_1-\rho_2)^2}\Bigg[& -\frac{\rho_2(1-\rho_1)}{\hat{M}_2}\ln\left(\frac{(\hat{M}_2-\hat{M}_1)\rho_2(1-\rho_1)}{\hat{M}_2\rho_1(1-\rho_2)-\hat{M}_1\rho_2(1-\rho_1)}\right) \\ & + \frac{\rho_1(1-\rho_2)}{\hat{M}_1}\ln\left(\frac{(\hat{M}_2-\hat{M}_1)\rho_1(1-\rho_2)}{\hat{M}_2\rho_1(1-\rho_2)-\hat{M}_1\rho_2(1-\rho_1)}\right)\Bigg]. \end{aligned} \tag{3.29}$$

Notice that the argument inside the $\ln(\cdot)$ terms in (3.20) are positive, by (2.6)-(2.7). Since $\ln(x) \le x-1$ for $x>0$, we have

$$\begin{aligned} L_{opt} &= \frac{\sigma^2}{2(\rho_1-\rho_2)^2}\Bigg[-\frac{\rho_2(1-\rho_1)}{\hat{M}_2}\left(\frac{\hat{M}_2(\rho_2-\rho_1)}{\hat{M}_2\rho_1(1-\rho_2)-\hat{M}_1\rho_2(1-\rho_1)}\right) \\ &\qquad + \frac{\rho_1(1-\rho_2)}{\hat{M}_1}\left(\frac{\hat{M}_1(\rho_2-\rho_1)}{\hat{M}_2\rho_1(1-\rho_2)-\hat{M}_1\rho_2(1-\rho_1)}\right)\Bigg], && (3.30) \\ &= \frac{\sigma^2}{2(\rho_1-\rho_2)^2}\left(\frac{(\rho_1-\rho_2)^2}{\hat{M}_1\rho_2(1-\rho_1)-\hat{M}_2\rho_1(1-\rho_2)}\right), && (3.31) \\ &= \frac{\sigma^2}{2[\hat{M}_1\rho_2(1-\rho_1)-\hat{M}_2\rho_1(1-\rho_2)]}, && (3.32) \\ &\le \frac{2\rho_1\rho_2(\hat{M}_1-\hat{M}_2)}{2[\hat{M}_1\rho_2(1-\rho_1)-\hat{M}_2\rho_1(1-\rho_2)]}, \quad \text{by Lemma 2.} && (3.33) \end{aligned}$$

Suppose $\rho_1 > \rho_2$. Then

$$\begin{aligned} L_{opt} &\le \frac{2\rho_1\rho_2(\hat{M}_1-\hat{M}_2)}{2[\hat{M}_1\rho_2(1-\rho_1)-\hat{M}_2\rho_2(1-\rho_1)]}, && (3.34) \\ &= \frac{\rho_1}{1-\rho_1}. && (3.35) \end{aligned}$$

Similarly, if $\rho_2 > \rho_1$, then

$$L_{opt} \leq \frac{2\rho_1\rho_2(\hat{M}_1 - \hat{M}_2)}{2[\hat{M}_1\rho_1(1-\rho_2) - \hat{M}_2\rho_1(1-\rho_2)]}, \tag{3.36}$$

$$= \frac{\rho_2}{1-\rho_2}. \quad \blacksquare \tag{3.37}$$

Acknowledgements

We are grateful to J. Michael Harrison for helpful discussions. This research is partially supported by a grant from the Leaders for Manufacturing Program at MIT, and by an IBM/University Manufacturing Systems Research Grant.

REFERENCES

Baskett, F., K. M. Chandy, R. R. Muntz, and F. G. Palacios. 1975. Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *J. Assoc. Comput. Mach.* **22**, 248-260.

Harrison, J. M. 1988. Brownian Models of Queueing Networks with Heterogeneous Customer Populations, in W. Fleming and P. L. Lions (eds.), *Stochastic Differential Systems, Stochastic Control Theory and Applications,* IMA Volume **10**, Springer-Verlag, New York, 147-186.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks*, John Wiley and Sons, New York.

Kushner, H. J., and L. F. Martins. 1989. Limit Theorems for Pathwise Average Cost Per Unit Time Problems for Queues in Heavy Traffic. Lefschetz Center for Dynamical Systems Report #89-18, Brown University.

Kushner, H. J., and K. M. Ramachandran. 1989. Optimal and Approximately Optimal Control Policies for Queues in Heavy Traffic. *SIAM J. Control and Optimization* **27**, 1293-1318.

Martins, L. F. and H. J. Kushner. 1989. Routing and Singular Control for Queueing Networks in Heavy Traffic. To appear in *SIAM J. Control and Optimization.*

Wein, L. M. 1989a. Optimal Control of a Two-Station Brownian Network. To appear in *Mathematics of Operations Research.*

Wein, L. M. 1989b. Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network With Controllable Inputs. To appear in *Operations Research.*