Access Control of Parallel Multiserver Loss Queues

Cheng-Yuan Ku

Department of Information Management

National Chung Cheng University, Chia-Yi County, Taiwan, ROC

Scott Jordan

Department of Electrical and Computer Engineering

University of California, Irvine, CA 92697-2625, USA

Abstract

This paper considers access control in a target multiserver loss queue fed by a set of upstream parallel multiserver loss queues and by a stream of new customers. The target queue faces a choice of how many servers to reserve for each stream. Revenue is gained by each station when it serves a customer, but the amount of revenue at the target queue depends on the source of the customer. We prove that the policy that maximizes total discounted revenue consists of a set of monotonically decreasing thresholds as functions of the occupancy of each queue. We prove monotonicity properties with respect to system parameters. We show that there exists an ordering of the thresholds based on the relative revenue paid at the target queue. Finally, we compare this system with a tandem queue model.

Keywords: Loss networks, Connection admission control, Dynamic programming

1. Introduction

This paper considers access control policies in a target multiserver station fed by a set of upstream parallel multiserver loss queues and by a stream of new customers. Each station has an arbitrary number of servers, but no queue. Service times of each customer are i.i.d., but the revenue generated at the target queue is differentiated based on the source of the customer, with new customers at the target queue paying less than upstream customers. Departures from upstream queues are Bernoulli routed either to the target queue or out of the network. We are interested in the access control policy for the target queue that maximizes total discounted revenue.

A network of queues can serve as a useful model for systems arising in many contexts, including production networks and communication networks. There is a long history of descriptive models being used to evaluate and predict the performance of existing and proposed systems, and thus improve the design of these systems. In particular, there is a rich literature using Markov decision processes (MDPs) to analyze control policies in such networks. See [1] for an excellent overview of the use of MDPs in communication networks.

We are in particular interested in the control of *loss networks*, i.e. networks of queues with no buffer beyond one space per server. Many communication network papers have considered allocation of capacity and/or control of arrivals to queues, see e.g. [2][4][11][18][22][26][27][37]. However, few communication networks papers have considered admission control of networks of more than two queues. Fewer yet have considered admission control of networks of loss queues.

The queueing literature has had a similar focus. Excellent surveys of research on access control policies in queueing networks can be found in [29][30]. Although there is a rich literature on control of single multiserver queues (see e.g. [14][24][35][36]), again there are very few results either on control of networks of more than two queues or on control of networks of loss queues.

We will start with a brief review of some of the literature on networks containing exactly two queues. Initial work approached control of such networks as a routing problem. Threshold policies and monotonicity properties governing routing of arrivals to one of two parallel queues have been proven to hold by Ephremides et al. [6]. Hajek [12] provided similar results for more general two queue systems that can be either arranged either in parallel or in series, and Beutler and Teneketzis [3] have demonstrated that monotonicity properties of optimal routing policies extend to partially observable queues. Admission control policies have also been considered in networks of two queues with infinite buffers. Davis [5] considered two exponential servers in parallel serving a renewal arrival process. An inductive proof based on value iteration shows that the optimal policy is admission monotonic and routing monotonic. Ghoneim and Stidham [10] studied two exponential servers in series. Using a similar approach, they establish that the optimal value function is concave in each argument and submodular. Hariharan et al. [13] extended the monotonicity properties derived by Davis and Hajek to control of admission and routing in two parallel queues, each with an infinite number of identical rate servers. Finally, Ku and Jordan [20] considered a system with two multiserver loss queues in series. Under appropriate conditions, the optimal admission policy is proven to be a monotonic threshold.

The literature on control of networks of more than two queues focuses almost entirely on routing approaches, see e.g. [7-9][15-17][23][25][31][33][34]. Typical routing results govern which queue to send a job to or which job to service next. There are very few papers on admission control in networks of more than two queues. Weber and Stidham [32] demonstrate

monotonicity properties in a network consisting of a series of queues, but they allow for admission control only to the first queue. As mentioned by Stidham and Weber [30], attempts to generalize structural results concerning admission control to more than two queues in parallel have met with little success. We have found no papers that provide results for admission control in loss networks of more than two queues.

In review, almost all of the current admission control literature has assumed infinite queues, with holding costs as the performance metric. In contrast, here we assume a loss network, and use loss as the performance metric.

In this paper, we prove that, under appropriate conditions, the optimal admission policy for the parallel multiserver loss system considered is given by a set of thresholds. Each threshold indicates whether to admit or block a customer, based on the source of that customer and on the occupancy of each queue. We prove that these thresholds are nonincreasing functions of the queue occupancies. We provide a set of results that characterize the variation of these thresholds with system parameters. These results, while similar to previous results for networks of two queues [20], are the first of their kind for networks of more than two queues.

Our second set of results is unique to the parallel network. We find that there exists an ordering of the thresholds based on the relative revenue paid at the target queue. We also prove that the optimal policy will reserve more capacity at the target queue if the customers are distributed more evenly at the upstream parallel queues. Furthermore, we compare this parallel queue network to a tandem network in which the first queue is given by the aggregation of the upstream queues in the parallel network. We show that the optimal policy in the parallel network reserves less capacity than the equivalent tandem network. Therefore, the tandem model, when used as a simpler approximation to the parallel model, generates a conservative

policy in terms of the reservation for internal customers. These results are stronger characterizations of the optimal policy than are typical in most admission control research.

Our primary technique, stochastic dynamic programming, has been used in much of the other research on control in queueing networks reviewed above. However, the structure of the network considered here is quite different than networks with infinite queues. This additional complexity is often presented by mutual interaction of multiple stations and presence of boundary states in the state space. In addition, using loss as the performance metric also changes the nature of the proof techniques.

The parallel multiserver loss queue model is presented in section 2. In section 3, we characterize the optimal admission policies and discuss the variation of these thresholds with system parameters. Numerical results are presented for a small system. In section 4, we compare this system to the simpler tandem queue system.

2. Model and problem formulation

Denote by n the number of upstream stations in the network, and denote the target queue by station $n+1$. Station i is a non-preemptive loss queue with m_i servers. Each station is presented with a Poisson stream of new customers, with arrival rate λ_i. Service times are i.i.d. and exponential at rate μ_i. Departures from station i $(1 \le i \le n)$ are routed out of the network with probability $\mathbf{P}_i$ or to station $n+1$ otherwise. This network is pictured in Fig. 1.

Revenue is paid by each customer at the start of service. If the customer entered the network at station i, it pays an amount r_i at station i and is denoted as type i. If the customer entered at an upstream station, it pays an amount $R_{i\to n+1}$ at station $n+1$. We assume that

$R_{1\to n+1} \geq R_{2\to n+1} \geq ... \geq R_{n\to n+1} > r_{n+1}$, so that upstream stations are ordered with priority and so that internal customers are preferable to new customers at station $n+1$, and that $1 > \mathbf{P}_i \geq 0$ for $1 \leq i \leq n$.

This system can be modeled as a $n+1$ dimensional continuous time Markov chain, with state $X = (x_1, x_2, ..., x_{n+1}) \in \Omega$ defined as the number of customers at stations 1, 2, ... , $n+1$ respectively. $\Omega \equiv \{0,1,...,m_1\}\text{x}\{0,1,...,m_2\}\text{x...x}\{0,1,...,m_{n+1}\}$ is the state space. It is convenient to define the following quantities:

- $\Omega_i \equiv \{0,1,...,m_1\}\text{x}\{0,1,...,m_2\}\text{x...x}\{0,1,...,m_i - 1\}\text{x...x}\{0,1,...,m_{n+1}\}$ for $1 \leq i \leq n+1$.
- $e_i \in \Omega$ is the vector with 1 in the i th entry and 0 elsewhere.

Uniformization results in an equivalent discrete time Markov chain by allowing fictitious transitions from a state to itself [21][24]. Choose any $Q \in \mathbf{R}$ s.t. $Q > \sum_{i=1}^{n+1} (\lambda_i + m_i \mu_i)$ and let $p_i = \frac{\lambda_i}{Q}$ and $q_i = \frac{\mu_i}{Q}$ for $1 \leq i \leq n+1$. The equivalent discrete time system has corresponding parameters $p_1, ..., p_{n+1}, q_1, ..., q_{n+1}$ and the appropriate discount factor $0 < \alpha < 1$.

We consider connection admission control policies that are capable of accepting or blocking arrivals of each customer type (internal or new) at each station. Blocked arrivals are lost. Our objective is to maximize the total discounted revenue over an infinite horizon. This criterion assumes an appropriate discount factor α, and, among all control policies π, attempts to maximize $V_\pi(X) = E_\pi\left[\sum_{n=0}^{\infty} R(X_n, X_{n+1})\alpha^n \middle| X_0 = X\right]$, where E_π represents the conditional expectation, given that control policy π is employed , $R(X_n, X_{n+1})$ is the revenue associated with a change of state from X_n to X_{n+1}, if any, and $\{X_n, n = 0,1,2,...\}$ is the sequence of states.

An admission control policy is defined by a $(n+1)\times(n+1)$ mapping matrix $A(X)=[a_{ij}(X)]$ with $a_{ij}(X):\Omega\to\{0,1\}$. $a_{ij}(X)=1$ (0) indicates the control action is to admit (reject) customer type i at station j when the system is in state X. For the convenience of notation, we designate that

- $a_{ij}(X)=0$ for $i\neq j$ and $j\neq n+1$
- $a_{ii}(X)=0$ for $X\in\Omega\setminus\Omega_i$
- $a_{ij}(X)=1$ for $x_i=0, i\neq n+1$ and $j=n+1$
- $a_{i(n+1)}(X)=0$ for $X\in\Omega\setminus\Omega_{n+1}$

Let $V(X)=\sup_{\pi}V_{\pi}(X)$. A control policy π^* is said to be α-optimal if $V_{\pi^*}(X)=V(X)$ for all $X\in\Omega$. According to Theorem 2.1 in Ross [28], this optimal control policy is chosen in each state to maximize the future expected discounted revenue as given by the following optimality equation. This dynamic programming equation, subject to the optimal control, consists of all possible transitions with corresponding transition probabilities.

$$V(X)=\max_{A(X)}\alpha\left\{\begin{array}{l}\sum_{i=1}^{n+1}p_i\{a_{ii}(X)[V(X+e_i)+r_i]+(1-a_{ii}(X))V(X)\}\\ +\sum_{i=1}^{n}(1-\mathbf{P}_i)x_iq_i\left\{\begin{array}{l}a_{i(n+1)}(X)[V(X-e_i+e_{n+1})+R_{i\to n+1}]\\ +(1-a_{i(n+1)}(X))V(X-e_i)\end{array}\right\}\\ +\sum_{i=1}^{n}\mathbf{P}_ix_iq_iV(X-e_i)+x_{n+1}q_{n+1}V(X-e_{n+1})\\ +(1-\sum_{i=1}^{n+1}(p_i+x_iq_i))V(X)\end{array}\right\} \quad (1)$$

If a tie occurs among multiple policies, we arbitrarily define the optimal policy to be the one that maximizes throughput. The optimal value function can be determined using successive approximation. Choose an arbitrary initial value function, V_0. Then define the step h value

function, $V_h(X)$, to be the expected discounted revenue starting in state $X = (x_1, x_2, ..., x_{n+1})$, by choosing actions that maximize future expected discounted revenue, assuming that transitions more than one time step in the future generate an average revenue according to the step $h-1$ value function. Then successive approximation gives:

$$V_h(X) = \max_{A(X)} \left\{ \alpha \sum_Y P_{X \to Y}(A) \{ R(X,Y) + V_{h-1}(Y) \} \right\} \tag{2}$$

where $P_{X \to Y}(A)$ is one-step transition probability. The optimal value function obeys:

$$V(X) = \lim_{h \to \infty} V_h(X) = \lim_{h \to \infty} V_{h-1}(X) \tag{3}$$

The optimal admission policy admits a customer if the immediate revenue generated by that customer exceeds the expected loss in future discounted revenue caused by future blocking due to this customer. We define difference functions $\Delta_h^i(X) = V_h(X) - V_h(X + e_i)$ and $\Delta^i(X) = V(X) - V(X + e_i)$ for $1 \le i \le n+1$ and $X \in \Omega_i$. Thus the optimal policy, in state X,

- For $1 \le i \le n+1$, admits a customer of type i at station i (i.e. $a_{ii}(X) = 1$) iff $\Delta^i(X) \le r_i$,
- For $1 \le i < n+1$, admits a customer of type i at station $n+1$ (i.e. $a_{i(n+1)}(X) = 1$) iff $\Delta^{n+1}(X - e_i) \le R_{i \to n+1}$.

3. Optimal access control

In this section, we present a sequence of theorems characterizing the form of the optimal admission control. Theorems 1 and 2 state that, under appropriate conditions, type 2,…, n and $n+1$ traffic may need to be controlled at station $n+1$. Theorems 3 and 4 state that optimal admission policies are given by a set of monotonically decreasing switching thresholds. Theorem 5 shows that there exists an ordering of the thresholds based on the relative revenue

paid at the target queue. Theorem 6 describes the variation of these thresholds with variations in system parameters.

The following lemma details the structure of the optimality value function, using value iteration. It will be used repeatedly to prove that the optimal policy, given by the corresponding limit, has similar desirable properties to those described here.

LEMMA 1

If:

(a) For $1 \le i \le n+1$, $\Delta_h^{n+1}(X)$ is monotonically increasing in x_i with the others fixed

(b) For $1 \le i \le n$ and $X \in \Omega_i$, $\Delta_h^i(X) < r_i$

(c) For $X \in \Omega_{n+1}$, $\Delta_h^{n+1}(X) < R_{1 \to n+1}$

Then:

(A) For $1 \le i \le n+1$, $\Delta_{h+1}^{n+1}(X)$ is monotonically increasing in x_i with the others fixed

(B) For $1 \le i \le n$ and $X \in \Omega_i$, $\Delta_{h+1}^i(X) < r_i$

(C) For $X \in \Omega_{n+1}$, $\Delta_{h+1}^{n+1}(X) < R_{1 \to n+1}$

<Outline of Proof>

In order to prove that $\Delta_{h+1}^{n+1}(X)$ is monotonically increasing in x_1 (property **A**), we consider two states X and $\overline{X}$ with $x_1 < \overline{x_1}$ and $x_i = \overline{x_i}$ for $2 \le i \le n+1$. We write $\Delta_{h+1}^{n+1}(X) = V_{h+1}(X) - V_{h+1}(X + e_{n+1})$ and substitute (2) into each term on the right hand side. The proof proceeds by collecting and comparing similar terms. The key is to demonstrate that terms generated by boundaries of the state space can be bounded by others. The resulting individual comparisons show that $\Delta_{h+1}^{n+1}(X) < \Delta_{h+1}^{n+1}(\overline{X})$. Property **(A)** is then used to prove properties **(B)** and **(C)**. The full proofs can be found in [19].

We first use this lemma to establish that customers should never be blocked at the upstream stations. This result is intuitive, as blocking upstream customers in stage 1 stations cannot increase revenue at the target stage 2 queue.

THEOREM 1

$\Delta^i(X) \le r_i$ for $1 \le i \le n$ and $X \in \Omega_i$. Consequently, it is optimal to always admit type i customers at station i, i.e., $a_{ii}(X) = 1$.

<Outline of Proof>

The statement follows directly from lemma 1 **(B)** using the limit given in (3). From the optimality equations, we know that the difference in the optimal value function between two neighboring states dictates whether the optimal policy will accept or deny the corresponding upcoming customers. Therefore, if this difference is always smaller than the corresponding revenue, then there is no need to control that type of customer at the indicated station.

We next use the lemma to establish that type 1 customers should never be blocked at the target queue. This result is also intuitive, as these customers generate the highest revenue at that station.

THEOREM 2

$\Delta^{n+1}(X) \le R_{1\to n+1}$ for $X \in \Omega_{n+1}$. Consequently, it is optimal to always admit type 1 customers at station $n+1$, i.e., $a_{1(n+1)}(X) = 1$.

<Outline of Proof>

The statement follows directly from lemma 1 **(C)** using the limit given in (3).

We next establish the key result that the optimal access control policies consist of a set of thresholds. An admission control policy for type i customers, $2 \le i \le n+1$, at station $n+1$ is called a *threshold policy* if there exists a $T^i(x_1,...,x_n)$ so that in state $X' = (x'_1, x'_2,..., x'_{n+1})$ type i customers are admitted (i.e. $a_{i(n+1)}(X') = 1$) iff $x'_{n+1} < T^i(x_1 = x'_1,..., x_n = x'_n)$.

THEOREM 3

$\Delta^{n+1}(X)$ is monotonically increasing in x_{n+1}. Consequently, the optimal access policy for the target queue consists of a set of thresholds.

<Outline of Proof>

The statement follows directly from lemma 1 **(A)** using the limit given in (3). The optimal control policy blocks the admission of a customer to a queue iff the difference function is less than the corresponding revenue. Therefore, if the difference function is monotonic in the state of the destination queue, then the optimal policy will be of the threshold type.

For such threshold policies, formally the thresholds are defined as follows. For $2 \le i \le n$

- $T^i(x_1,...,x_n) = \min(\, x_{n+1} \mid \Delta^{n+1}(X - e_i) > R_{i \to n+1}\,)$ with $T^i(x_1,...,x_n) = m_i$ if the minimum is undefined.

Similarly define

- $T^{n+1}(x_1,...,x_n) = \min(\, x_{n+1} \mid \Delta^{n+1}(X) > r_{n+1}\,)$ with $T^{n+1}(x_1,...,x_n) = m_{n+1}$ if the minimum is undefined.

We next establish the variation of these thresholds with the occupancies of each station.

THEOREM 4

$\Delta^{n+1}(X)$ is monotonically increasing in x_i for $1 \le i \le n$. Consequently, the optimal switching policies $T^i(x_1,...,x_n)$ are nonincreasing functions of x_1, x_2,…, and x_n.

<Outline of Proof>

The statement follows directly from lemma 1 **(A)** using the limit given in (3). If the difference function is monotonic in the state of the source queues, then the threshold will be also monotonic in those states.

It should be noted that optimal admission policies for more general networks do not necessarily obey the structure of monotonically decreasing switching thresholds. In particular, adding additional queues in series or adding cycles can destroy such properties due to interaction between policies at different stations [19].

Theorems 3 and 4 are illustrated with the use of a numerical example for the simplest such parallel multiserver loss system -- one with 2 upstream queues and 1 target queue. The system parameters are: $m_1 = 6$, $m_2 = 6$, $m_3 = 6$, $\alpha = 0.99995$, $Q = 100000$, $\lambda_1 = 2$, $\lambda_2 = 1.8$, $\lambda_3 = 0.8$, $\mu_1 = 4$, $\mu_2 = 3$, $\mu_3 = 3$, $\mathrm{P}_1 = 0.3$, $\mathrm{P}_2 = 0.1$, $r_1 = 5$, $r_2 = 5$, $r_3 = 5$, $R_{1\to 3} = 25$ and $R_{2\to 3} = 9$. Numerical results were obtained by the method of successive approximation. The optimal value function is found, and from that the optimal policy is inferred. Theorems 1 and 2 guarantee that no customers should be rejected at stations 1 and 2, and that type 1 customers should not be blocked at station 3. The numerical results confirm this. The access control policy therefore consists of policies for the admission of type 2 and 3 customers at station 3. These policies are verified to be threshold policies and are shown in Fig. 2 and Fig. 3.

Note that the thresholds are nonincreasing functions of x_1 and x_2 as guaranteed by theorem 4.

We now turn to a comparison of the different thresholds based on the relative revenue paid at the target queue.

THEOREM 5

(A) For $2 \le i \le n$, $T^i(x_1,...,x_n) \ge T^{n+1}(x_1,...,x_n)$. Consequently, if any upstream customer would be blocked at the target queue, then a new customer would also be blocked, when there are the same number of customers at each queue.

(B) $T^2(x_1, x_2+1, x_3,...,x_n) \ge T^3(x_1, x_2, x_3+1,...,x_n) \ge ... \ge T^n(x_1, x_2, x_3,...,x_n+1)$.

Consequently, if any upstream customer would be blocked at the target queue, then an upstream customer paying a lower revenue would also be blocked, when there are the same number of customers at each queue *excluding the transitioning upstream customer.*

<Proof>

(A) By hypothesis $R_{i\to n+1} > r_{n+1}$. Theorems 3 and 4 thus imply

$$T^i(x_1,...,x_n) = \min(\, x_{n+1} \,|\, \Delta^{n+1}(X - e_i) > R_{i\to n+1}\,)$$

$$\ge \min(\, x_{n+1} \,|\, \Delta^{n+1}(X) > R_{i\to n+1}\,)$$

$$\ge \min(\, x_{n+1} \,|\, \Delta^{n+1}(X) > r_{n+1}\,)$$

$$= T^{n+1}(x_1,...,x_n)$$

(B) By hypothesis, $R_{1\to n+1} \ge R_{2\to n+1} \ge ... \ge R_{n\to n+1} > r_{n+1}$. Theorems 3 and 4 thus imply

$$T^2(x_1, x_2+1, x_3,..,x_n) = \min(\, x_{n+1} \,|\, \Delta^{n+1}(x_1, x_2, x_3,..,x_n, x_{n+1}) > R_{2\to n+1}\,)$$

$$\ge \min(\, x_{n+1} \,|\, \Delta^{n+1}(x_1, x_2, x_3,..,x_n, x_{n+1}) > R_{3\to n+1}\,) = T^3(x_1, x_2, x_3+1,..,x_n)$$

$$\ge \ldots\ldots$$

$$\ge \min(x_{n+1} \,|\, \Delta^{n+1}(x_1, x_2, x_3,..,x_n, x_{n+1}) > R_{n\to n+1}\,) = T^n(x_1, x_2, x_3,..,x_n+1)$$

Theorem 5 demonstrates a form of priority at the downstream queue, based on the revenue paid at that queue. One may conjecture that $T^i(x_1, x_2, ..., x_n) \geq T^j(x_1, x_2, ..., x_n)$ for $2 \leq i < j \leq n$, namely that if any upstream customer would be blocked at the target queue, then any upstream customer with a lower priority would also be blocked *when there are the same number of customers at each queue*. This property, however, seems to depend on a "diagonal monotonicity" property, $\Delta^{n+1}(X - e_i) \leq \Delta^{n+1}(X - e_j)$ for $2 \leq i < j \leq n$, that we have been unable to prove [19].

Finally, we investigate the variation of each threshold with system parameters.

THEOREM 6

The admission control threshold on type i customers, $T^i(x_1, x_2, ..., x_n)$, is for $2 \leq i \leq n+1$ and $R_{1 \to n+1} \geq R_{2 \to n+1} \geq ... \geq R_{n \to n+1} > r_{n+1}$:

(1) monotonically decreasing in α, λ_j $(1 \leq j \leq n+1)$, $R_{j \to n+1}$ $(1 \leq j \leq n)$ and m_j $(1 \leq j \leq n)$

(2) monotonically increasing in μ_{n+1}, r_{n+1}, and $\mathbf{P}_j$ $(1 \leq j \leq n)$

(3) insensitive to r_j $(1 \leq j \leq n)$

Furthermore,

(4) $m_{n+1} - T^i(x_1, ..., x_n)$ ($2 \leq i \leq n+1$) are monotonically decreasing in m_{n+1}, while $T^i(x_1, ..., x_n) > 0$.

<Outline of Proof>

The proofs for sensitivity with respect to arrival rates, service rates, first stage revenues, and capacities build on lemma 1 and theorems 1-4. Using appropriately linked initial values, the monotonicity of the thresholds demonstrated in theorem 4 is used to establish an ordering on

the step h difference functions. The resulting limit proves the desired property. The proofs for sensitivity with respect to second stage revenues use a revenue scaling to which the optimal policy is invariant. Then the monotonicity with respect to first stage revenues is invoked to prove the desired property. The full proofs can be found in [19].

This theorem is valuable for the design of networking systems, since it demonstrates how the optimal admission policies vary with the capacities of the network and with the weighted priorities.

4. Optimal access control for identical first-stage queues

In this section, we consider the special case in which all multiserver loss queues at the first stage are identical with respect to arrival rates and capacities. We demonstrate two comparisons of efficiency. First, we show that the same or less capacity is reserved at the second stage queue as the disparity in occupancy of first stage queues increases.

Second, we compare this parallel queue network with identical first stage queues to a tandem network in which the first queue is given by the aggregation of the upstream queues in the parallel network. Such a tandem model could be used as a simpler model, and computation of optimal policies would be much quicker. The key question is how do the optimal policies for the two systems compare? We demonstrate here that the tandem model reserves no less capacity in the second stage queue than the parallel model.

We start with a parallel multiserver loss model in the special case in which all multiserver loss queues at the first stage are identical with respect to arrival rates and capacities. We also assume that customers leaving these upstream queues will go to station $n+1$ with the same probability and pay the same revenue. It follows from theorem 2 that the optimal policy admits

all customers of types 1,…, n at the first stage queue. Lemma 2 investigates how the control policy for new customers at the second stage queue is affected by the distribution of customers at the first stage.

LEMMA 2

For $X \in \Omega_{n+1}$, if statements **(a)**, **(b)** and **(c)** in Lemma 1 hold and if:

(d) $\Delta_h^{n+1}(X + ke_i - ke_j) > \Delta_h^{n+1}(X + (k+1)e_i - (k+1)e_j)$ for $k \in n$ and $x_i = x_j > 0$ s.t. $x_i + (k+1) \le m_i$, $x_j - (k+1) \ge 0$

(e) $\Delta_h^{n+1}(X + ke_i - ke_j) > \Delta_h^{n+1}(X + (k+1)e_i - (k+1)e_j)$ for $k \in n$ and $x_i = x_j + 1 > 1$ s.t. $x_i + (k+1) \le m_i$, $x_j - (k+1) \ge 0$

Then:

(D) $\Delta_{h+1}^{n+1}(X + ke_i - ke_j) > \Delta_{h+1}^{n+1}(X + (k+1)e_i - (k+1)e_j)$ for $k \in n$ and $x_i = x_j > 0$ s.t. $x_i + (k+1) \le m_i$, $x_j - (k+1) \ge 0$

(E) $\Delta_{h+1}^{n+1}(X + ke_i - ke_j) > \Delta_{h+1}^{n+1}(X + (k+1)e_i - (k+1)e_j)$ for $k \in n$ and $x_i = x_j + 1 > 1$ s.t. $x_i + (k+1) \le m_i$, $x_j - (k+1) \ge 0$

<Outline of Proof>

In order to prove property (**D**), we consider two type $n+1$ differences $\Delta_{h+1}^{n+1}(X + ce_i - ce_j)$ and $\Delta_{h+1}^{n+1}(X + de_i - de_j)$ with $0 \le c < d \le \min(x_i, m_i - x_i)$. We write $\Delta_{h+1}^{n+1}(X + ce_i - ce_j) = V_{h+1}(X + ce_i - ce_j) - V_{h+1}(X + ce_i - ce_j + e_{n+1})$ and substitute (2) into each term on the right hand side. As with lemma 1, the proof proceeds by collecting and comparing similar terms. The key is again to demonstrate that terms generated by boundaries of the state space can be bounded by others. Unlike in the proof of lemma 1 however, this time we know that the optimal policies are thresholds. This information can be used in portions of

state space. The resulting individual comparisons show that $\Delta_{h+1}^{n+1}(X + ce_i - ce_j) > \Delta_{h+1}^{n+1}(X + de_i - de_j)$. The proof of property **(E)** is similar. The full proofs can be found in [19].

We can now establish our desired result.

THEOREM 7

For $1 \le i, j \le n$, $X \in \Omega_{n+1}$, and $x_i \ge x_j$, $\Delta^{n+1}(X) \ge \Delta^{n+1}(X + e_i - e_j)$. Consequently, the threshold is monotonically increasing with greater disparity in first stage queue occupancy.

<Outline of Proof>

The statement follows directly from lemma 2 using the limit given in (3). Since type $n+1$ customers are admitted iff $\Delta^{n+1}(X) \le r_{n+1}$, the optimal policy will reserve the same or more spaces when customers are more equally distributed at the first stage.

This result is intuitive. A more equal distribution will decrease the short-term blocking probability at the first stage. This will have a secondary effect of increasing the flow from first stage queues to the second stage. Finally, this increased flow may increase the desired reserved capacity at the second stage queue.

We now turn to a comparison of this parallel queue network with identical first stage queues to a tandem network in which the first queue is given by the aggregation of the upstream queues in the parallel network.

Designate the original parallel system as system 1, and the tandem system (as pictured in Fig. 4) as system 2.

In system 2, the first queue therefore has an arrival rate and a service capacity equal to *n* times the individual arrival rates and service capacities of the first stage queues in system 1. It also follows from theorem 2 that the optimal policy admits all internal customers. So we would like to compare the optimal policies for new customers at the second stage for the two systems. A fair comparison equates system 1 in state X to system 2 in state $\left(\sum_{i=1}^{n} x_i, x_{n+1}\right)$.

We define the following value differences for system 2:

- $\overline{\Delta_h^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) = V_h(\sum_{i=1}^{n} x_i, x_{n+1}) - V_h(\sum_{i=1}^{n} x_i, x_{n+1}+1)$
- $\overline{\Delta^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) = V(\sum_{i=1}^{n} x_i, x_{n+1}) - V(\sum_{i=1}^{n} x_i, x_{n+1}+1)$.

Lemma 3 will be used to prove the main result.

LEMMA 3

Suppose statements **(a)**, **(b)** and **(c)** in lemma 1 hold for $\Delta_h^{n+1}(X)$ for system 1 and for $\overline{\Delta_h^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1})$ for system 2. Suppose also that:

(f) $\overline{\Delta_h^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) > \Delta_h^{n+1}(X)$ for $X \in \Omega_{n+1}$

Then :

(F) $\overline{\Delta_{h+1}^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) > \Delta_{h+1}^{n+1}(X)$ for $X \in \Omega_{n+1}$

<Outline of Proof>

Denote $\overline{X} = \left(\sum_{i=1}^{n} x_i, x_{n+1}\right)$. We write $\Delta_{h+1}^{n+1}(X) = V_{h+1}(X) - V_{h+1}(X + e_{n+1})$ and substitute (2) into each term on the right hand side. We similarly expand $\overline{\Delta_{h+1}^{n+1}}(\overline{X})$. As with previous lemmas,

the proof proceeds by collecting and comparing similar terms. As with lemma 2, knowledge that the optimal policy is a threshold is helpful. The full proofs can be found in [19].

We can now establish our desired result.

THEOREM 8

$\overline{\Delta^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) \geq \Delta^{n+1}(X)$ for $X \in \Omega_{n+1}$. Consequently, the threshold in the tandem system is no higher than the equivalent threshold in the parallel system.

<Outline of Proof>

The statement follows directly from lemma 3 using the limit given in (3). Since type $n+1$ customers are admitted in system 1 iff $\Delta^{n+1}(X) \leq r_{n+1}$ and in system 2 iff $\overline{\Delta^{n+1}}(\sum_{i=1}^{n} x_i, x_{n+1}) \leq r_{n+1}$, the optimal policies will reserve the same or more spaces for system 2 given an equal number of customers in the first stage.

This result is intuitive. The first stage in system 2 is more efficient than the first stage in system 1, since customers in system 2 are not limited to obtaining service from a particular server. Therefore, the blocking probability at the first stage in system 2 is less than in system 1. Correspondingly, first stage throughput in system 2 is higher. Finally, this results in an equal or larger reservation at the second stage for system 2.

This result can justify using the tandem model as a simpler approximation to the parallel model, in the sense that the resulting optimal policy for system 2 is a conservative policy for system 1 in terms of the reservation for internal customers.

5. Conclusion

We have considered access control in a target multiserver loss queue fed by a set of upstream parallel multiserver loss queues and by a stream of new customers. Such models arise in computer and telecommunication networks, in which continued service to internal customers is preferable to admission of new customers. We proved that the policy that maximizes total discounted revenue consists of a set of monotonically decreasing thresholds as functions of the occupancy of each upstream queue. We proved monotonicity properties with respect to system parameters. We showed that there exists an ordering of the thresholds based on the relative revenue paid at the target queue. Finally, we compared this system with a tandem queue model.

References

[1] E. Altman, Application of Markov Decision Processes in Communication Networks: A Survey, INRIA Research Report RR-3984, 2000.

[2] E. Altman, S. Stidham, Optimality of monotonic policies for two-action Markovian decision processes with applications to control of queues with delayed information, Queueing Systems, 21 (1995) 267-291.

[3] F. J. Beutler, D. Teneketzis, Routing in queueing networks under imperfect information, Stochastics and Stochastics Reports (1989) 81-100.

[4] P. -J. Courtois, G. Scheys, Minimization of the total loss rate for two finite queues in series, IEEE Transactions on Communications, 39 (1991) 1651-1661.

[5] E. Davis, Optimal control of arrivals to a two-server queueing system with separate queues, Ph.D. dissertation, Program of Operation Researches, North Carolina State University, 1977.

[6] A. Ephremides, P. Varaiya, J. Walrand, A simple dynamic routing problem, IEEE Transactions on Automatic Control, 25 (1980) 690-693.

[7] T. M. Farrar, Resource allocation in systems of queues, Ph.D. dissertation, University of Cambridge (1992).

[8] G. Foschini, On heavy traffic analysis and dynamic routing in packet-switched networks, Computer Performance, (North-Holland, 1977) 499-513.

[9] G. Foschini, J. Salz, A basic dynamic routing problem and diffusion, IEEE Trans. on Communications, 26 (1978) 320-327.

[10] H. Ghoneim, S. Stidham, Optimal control of arrivals to two queues in series, European Journal of Operational Research, 21 (1985) 399-409.

[11] J. M. Hah, P. L. Tien, M. C. Yuang, Neural-network-based call admission control in ATM networks with heterogeneous arrivals, Computer Communications, 20 (1997) 732-740.

[12] B. Hajek, Optimal control of two interacting service stations, IEEE Transactions on Automatic Control, 29 (1984) 491-499.

[13] R. Hariharan, V. G. Kulkarni, S. Stidham, Optimal control of admission and routing to two parallel infinite-server queues, Proc. 29th IEEE Conference on Decision and Control (1990).

[14] W. E. Helm, K.-H. Waldmann, "Optimal Control of Arrivals to Multiserver Queues in A Random Environment, J. Appl. Prob. 21 (1984) 602-615.

[15] A. Hordijk, G. Koole, Note on the optimality of the generalized shortest queue policy, Prob. Eng. Inf. Sci. 6 (1992).

[16] D. J. Houck, Comparison of policies for routing customers of parallel queueing systems, Operations Research, 35 (1987) 306-310.

[17] P. K. Johri, Optimality of the shortest line discipline with state-dependent service times, European Journal of Operational Research, 41 (1990) 157-161.

[18] G. Karlsson, Capacity reservation in ATM networks, Computer Communications, 19 (1996) 180-193.

[19] C.-Y. Ku, Access Control for Loss Network, Ph.D. dissertation, Department of EECS, Northwestern University, 1995.

[20] C.-Y. Ku, S. Jordan, Access control to two multiserver loss queues in series, IEEE Transactions on Automatic Control, 42 (1997) 1017-1023.

[21] P. R. Kumar, P. P. Varaiya, Stochastic Systems, Prentice-Hall, 1986.

[22] T. -H. Lee, K. -C. Lai, S. -T. Duann, Design of a real-time call admission controller for ATM networks, IEEE/ACM Transactions on Networking, 4 (1996) 758-765.

[23] T. Lehtonen, On the optimality of the shortest-line discipline, Ph.D. dissertation, Helsinki School of Economics (1981).

[24] S. A. Lippman, "Applying a new device in the optimization of exponential queueing systems", Operations Research, 23 (1975) 687-710.

[25] R. Menich, R. Serfozo, Optimality of shortest-queue routing for dependent service stations, Queueing Systems, 9 (1991) 403-418.

[26] P. Mohge, I. Rubin, Reserving for future clients in a multipoint application-why and how, IEEE Journal on Selected Areas in Communications, 15 (1997) 531-544.

[27] J. S. Park, S. H. Lee, S. C. Kim, J. Y. Lee, S. B. Lee, A conferencing system for real-time, multiparty, multimedia services, IEEE Transactions on Consumer Electronics, 44 (1998) 857-865.

[28] S. M. Ross, Introduction to Stochastic Dynamic Programming, Academic Press, 1983.

[29] S. Stidham, Optimal control of admission to a queueing system, IEEE Transactions on Automatic Control, 30 (1985) 705-713.

[30] S. Stidham, R. Weber, A survey of Markov decision models for control of networks of queues, Queueing Systems, 13 (1993) 291-314.

[31] R. Weber, On the optimal assignment of customers to parallel servers, Journal of Applied Probability, 15 (1978) 406-413.

[32] R. Weber, S. Stidham, Control of service rates in networks of queues, Advanced Applied Probability, 24 (1987) 202-218.

[33] W. Whitt, Deciding which queue to join; some counterexamples, Operations Research. 34 (1986) 55-62.

[34] W. Winston, Optimality of the shortest-processing-time discipline, Journal of Applied Probability, 14 (1977) 181-189.

[35] U. Yechiali, On optimal balking rules and toll charges in a GI/M/1 queueing process, Operations Research, 19 (1971) 349-370.

[36] U. Yechiali, Customers optimal joining rules for GI/M/S queue, Management Science, 18 (1972) 434-443.

[37] R. Zhang, Y. A. Phillis, Fuzzy control of arrivals to tandem queues with two stations, IEEE Transactions on Fuzzy Systems, 7 (1999) 361-367.

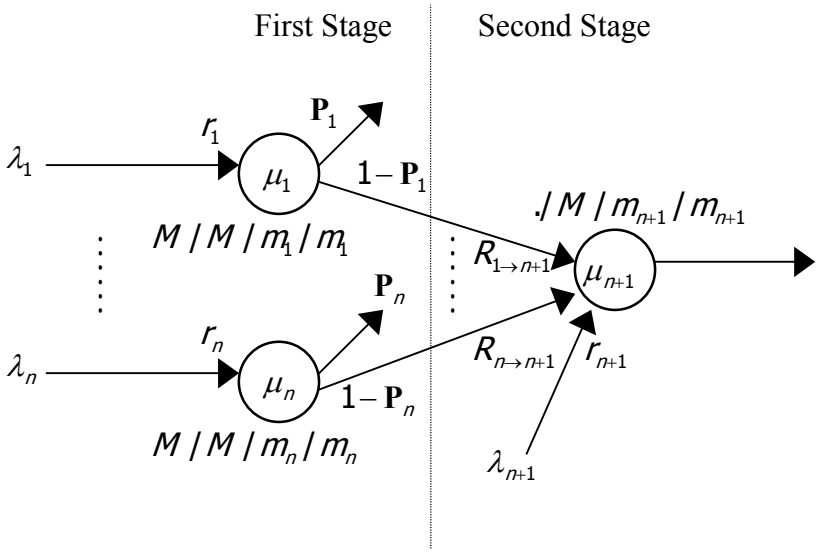

Fig. 1. Parallel multiserver loss network

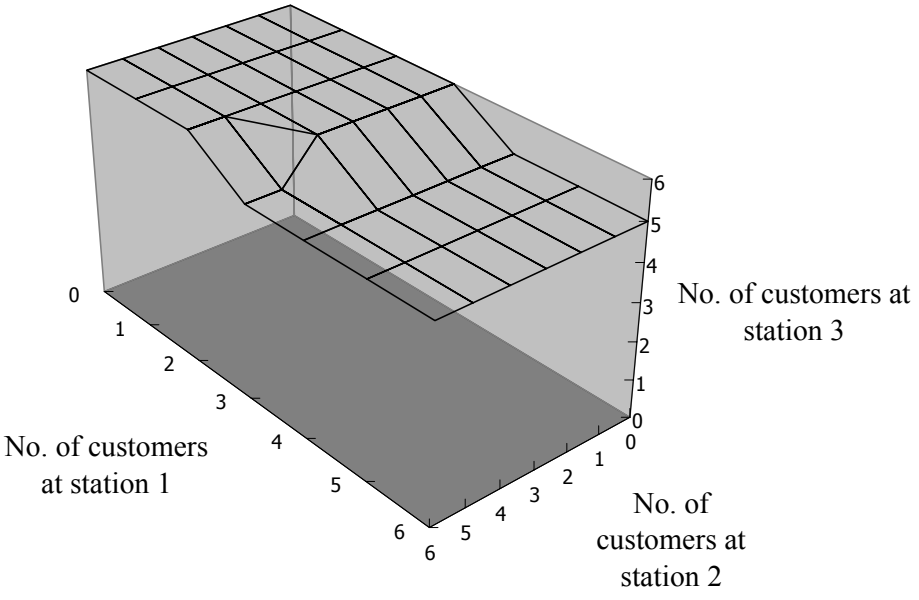

Fig. 2. The optimal threshold $T^2(x_1, x_2)$ on type 2 customers at station 3

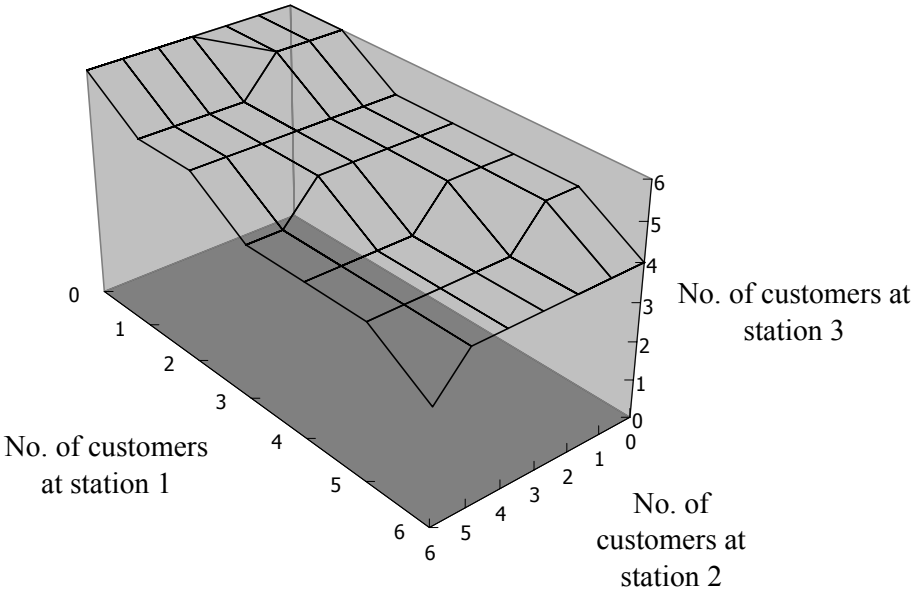

Fig. 3. The optimal threshold $T^3(x_1, x_2)$ on type 3 customers at station 3

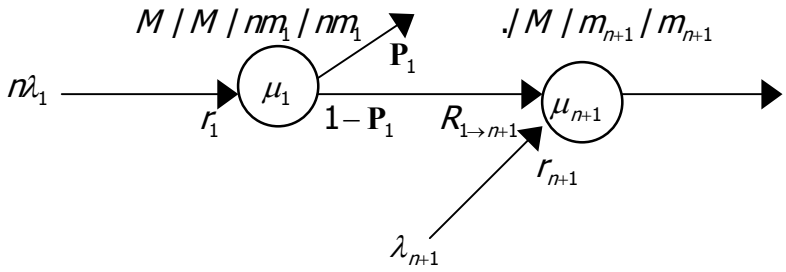

Fig. 4. Two loss queues in tandem

Figure captions

[1] Fig. 1. Parallel multiserver loss network

[2] Fig. 2. The optimal threshold $T^2(x_1, x_2)$ on type 2 customers at station 3

[3] Fig. 3. The optimal threshold $T^3(x_1, x_2)$ on type 3 customers at station 3

[4] Fig. 4. Two loss queues in tandem