December 1992 LIDS-P-2148

On the optimal admission schedule of a finite population to a queue [1]

Daniel C. Lee
John N. Tsitsiklis

Laboratory for Information and Decision Systems, M.I.T.
Cambridge, MA 02139

ABSTRACT

We study an optimal schedule for admitting a fixed number of customers from an auxiliary buffer with low holding cost to a main queueing system with high holding cost. We prove that the overall admission rate under the optimal schedule converges to the service rate of the main system as the number of customers grows. We also show asymptotic properties regarding admission rate and interadmission times in early part of the optimal schedule.

Key words: optimal admission schedule, admission rate, queueing delay

[1] Research supported by the NSF under grant ECS-8552419

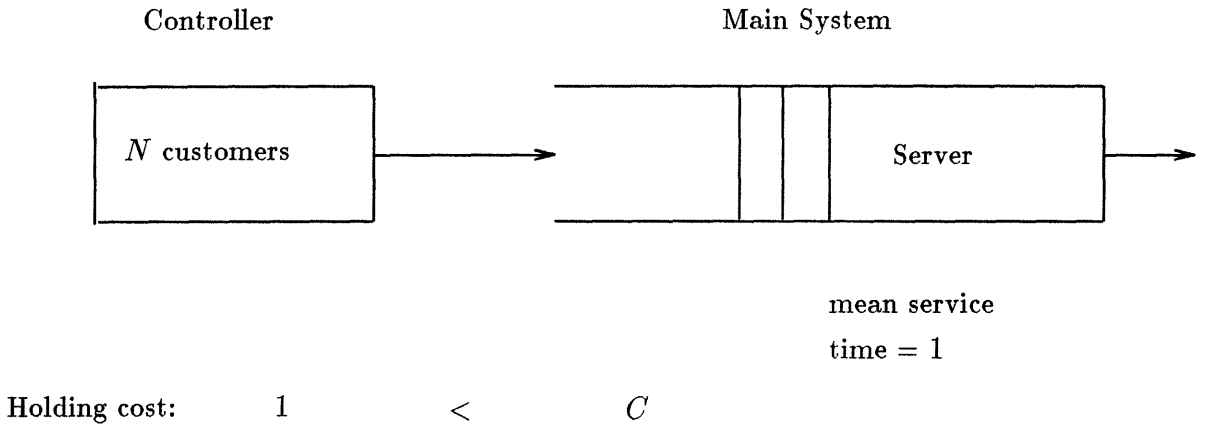

Figure 1: A system consisting of a controller followed by a single server queue

1 Introduction

The problem of optimal admission control in a queueing system has received significant attention over a number of years [1, 3, 4, 5, 6, 7, 9, 10, 15, 16]. In most of this literature, the assumption is made that the controller that regulates admissions has access to full information on the state of the system. In this paper, we consider the opposite situation in which the controller has no information on the state of the queueing system other than the knowledge of its own past actions. Our motivation comes primarily from the context of high–speed communication networks. The time scales in such networks are so fast that very limited real–time feedback is possible; in particular, many of the flow control actions have to be made essentially open loop [2]. Past literature on the subject of admission control under imperfect information is limited [3] [11] [13] [14]. These references deal with optimal on–line admission control strategies to cope with a steady arrival stream of customers.

In this paper, we study optimal off-line scheduling of admission times of a fixed number of customers. We concentrate our studies on the system depicted in Figure 1. It consists of a controller and a main system. Initially, the controller has a fixed number of customers, and the main system is empty. The main system is assumed to have an exponential server with service rate 1. In our model we allow any work–conserving service discipline that does not need knowledge of service times other than their statistics. There are two important additional components of our model:

- The controller's queue is "less expensive." In particular, we assume that the ith customer incurs a cost $W_i^1 + CW_i^2$, where W_i^1 is the time that the customer spends in the controller's queue, W_i^2 is the time that it spends in the main system's queue, and C is a constant larger than 1.

- The controller cannot observe the state of the main system's queue.

Motivation for the form of the cost function we have introduced comes from the context of flow control in communication networks. While network level flow control (e.g., the controller's actions in our model) cannot reduce the total delay experienced by a typical data packet, it attempts to shift delay from a network layer (e.g. the main system in our model) to higher layers (e.g. the controller in our model) in order to avoid wasteful congestion in the network layer [2]. We penalize the delay in the main system more than the delay in the controller in order to capture the essential idea that the congestion in the network layer is more harmful because of the resulting buffer overflows and retransmissions.

Given the above framework, we wish to study admission times that minimize the sum of the expected costs incurred by the customers. In relation to the admission control of steady arrival streams, the study in this paper can add insights to the case of rare

arrivals in big bulk. This paper addresses the question: how should a large bulk of customers be fed into when the arrival stream has extremely low intensity?

It is intuitive that customers should be fed into the service facility at a rate approximately equal to the service rate, and we provide some results corroborating this intuition. In section 2, we show that a periodic admission schedule becomes asymptotically optimal if the interadmission time is properly adjusted for the number of customers. In section 3, we show that the average admission rate under the optimal schedule converges to the service rate as the number of customers increases. In section 4, we show that the early interadmission times under the optimal schedule can be arbitrarily small due to an edge effect. In section 5, we provide asymptotic properties of the optimal schedule regarding admission rate averaged over partial time intervals.

2 Formulation and the performance of periodic admission

In the system described in Figure 1 there are N customers in the controller, and the main system is initially empty. The controller has the knowledge that the main system is empty initially but cannot observe the main system's queue status thereafter. The decisions to be made by the controller can be summarized by a finite nondecreasing sequence $\pi^N = (t_1^N, \ldots, t_N^N)$ of nonnegative real numbers, where t_i^N represents the time that the i-th customer is admitted. Given such a sequence π^N of admission times, the cost incurred at the controller is given by

$$\sum_{i=1}^{N} t_i^N. \tag{1}$$

In addition, if T_i is the (random) time at which the service of the ith customer at the main system is completed, the cost incurred at the main system is given by

$$C \sum_{i=1}^{N} E[T_i - t_i^N]. \tag{2}$$

The objective is to find, for any given N, a sequence π^N that minimizes the sum of the expressions (1) and (2). Note that t_1^N is obviously 0 for a minimal π^N. For any fixed N, this is a deterministic nonlinear programming problem in the variables $t_1^N, \ldots, t_N^N$. Even though it can be verified that the cost function is convex, no closed form solution is in sight. For this reason, we will limit ourselves to the case where N is very large and we will derive a schedule whose cost exceeds that of the optimal by a vanishingly small percentage, as N increases. In later sections we will derive, based on the intuition gained in this section, some properties of an exactly optimal schedule when N is large.

Let G_N be the cost of an optimal schedule for the problem we are considering. Let us consider the schedule π_*^N in which $t_i^N = (i-1)(1+\epsilon_N)$ for each i. That is, customers are admitted periodically, once every $(1+\epsilon_N)$ time units, where ϵ_N is a positive real number whose properties will be specified later. Let F_N be the cost of this particular schedule. The result that follows states that, asymptotically, F_N is very close to G_N.

Theorem 1 *If* $\lim_{N\to\infty} \epsilon_N = 0$ *and* $\lim_{N\to\infty} N\epsilon_N = \infty$, *then*

$$\lim_{N\to\infty} \frac{F_N}{G_N} = 1$$

Proof: It is evident that $G_N \leq F_N$ for all N; it remains to prove that the reverse inequality also holds, asymptotically. We start by deriving a lower bound on the optimal cost G_N.

Let Q_N be the value of the optimal cost if the controller had full and instantaneous information on the state of the main system. It is clear that, under perfect information, and under the assumption $C > 1$, the optimal schedule for the controller is to admit a new customer whenever the main system becomes idle. Then, the total expected cost incurred at the main system is CN. Furthermore, the time of admission of the i-th customer is equal to the sum of the service times of the first $(i-1)$ customers and its expectation is equal to $(i-1)$. We conclude that

$$G_N \geq Q_N = \sum_{i=1}^{N}(i-1) + CN = \frac{N(N-1)}{2} + CN \tag{3}$$

We now continue with an upper bound on F_N. The cost in the controller is easily found to be $(1+\epsilon_N)N(N-1)/2$. Note that the main system behaves as a $D/M/1$ queue. Given that the main system is initially empty, a standard stochastic dominance argument implies that the expected response time of each customer (in the main system) is bounded above by the expected response time in a $D/M/1$ queue in steady-state, which is $1/(1-\sigma_N)$ [8], where σ_N is the unique solution to

$$\sigma = \exp\{-(1+\epsilon_N)(1-\sigma)\}, \quad 0 < \sigma < 1 \tag{4}$$

Putting everything together, it follows that

$$F_N \leq (1+\epsilon_N)\frac{N(N-1)}{2} + \frac{CN}{1-\sigma_N}$$

Consider a quadratic function $f(\sigma) = 1-(1+\epsilon)(1-\sigma)+B(1-\sigma)^2$ and the equation

$$\sigma = f(\sigma), \quad 0 < \sigma < 1 \tag{5}$$

For sufficiently small ϵ and sufficiently large B, $f(\sigma)$ is no less than the right-hand side of equation (4) for each σ between 0 and 1, and $f(1) = 1$. Therefore, $1-\epsilon_N/B$, which is the unique solution to equation (5), is no less than σ_N for sufficiently large N. Therefore, $1/(1-\sigma_N) \leq B/\epsilon_N$, and we have

$$F_N \leq (1+\epsilon_N)\frac{N(N-1)}{2} + \frac{CBN}{\epsilon_N} \tag{6}$$

Using the assumption $N\epsilon_N \to \infty$, the term CBN/ϵ_N is negligible compared to the $O(N^2)$ term. We then get, using equation (3) and (6)

$$\lim_{N\to\infty} \frac{F_N}{G_N} \leq \lim_{N\to\infty} (1+\epsilon_N) = 1.$$

Q.E.D.

We note that the proof of Theorem 1 indicates that the periodic admission schedule we described is not only very close to being optimal asymptotically, but also comes very close to the cost of an optimal closed–loop, on–line admission policy.

3 Overall admission rate of an optimal schedule

The intuition provided by Theorem 1 leads to the conjecture that the admission rate of an optimal schedule must converge to the service rate. In this section we rigorously state and prove this conjecture. We now fix our notation. Throughout this section, t_i^N will stand for the admission time of the ith customer under an optimal schedule for the N–customer problem.

Theorem 2

$$\lim_{N\to\infty} \frac{N}{t_N^N} = 1$$

Proof: Consider the following two lemmas.

Lemma 3 $\liminf_{N\to\infty} t_N^N/N \geq 1$.

Proof: Suppose that this lemma is false; then, there exists some $\gamma > 0$ and an increasing sequence $\{N_k\}$ such that $t_{N_k}^{N_k} < (1-\gamma)N_k$ for all k. We will argue that the last customer is very likely to find the main system busy and that the cost of the schedule can be reduced by delaying the admission time of the last customer, thus contradicting optimality.

Let us fix some $\delta > 0$. Since the service time distribution is exponential, the service order does not affect the cost. Assume that the first–come–first–serve discipline is used in the main system. Let α_k be the probability that the service of the first $N_k - 1$ customers is finished by $t_{N_k}^{N_k} + \delta$. This can happen only if the sum of the service times of the first $N_k - 1$ customers is less than $(1-\gamma)N_k + \delta$. The weak law of large numbers implies that this event has vanishingly small probability, and $\lim_{k\to\infty} \alpha_k = 0$.

Let us now consider a modified admission schedule for the N_k–customer problem in which the admission of the last customer is delayed by δ. This modification increases the cost incurred in the controller by δ. To compare the costs at the main system, we use a coupling argument: we assume that the service times of all customers are the same under both policies. Then, the costs incurred at the main system are the same under both policies, with the possible exception of the last customer. If under the

original schedule, the last customer finds the main system idle, the same will be true under the modified schedule and the cost incurred under either schedule at the main system will be the same. If under the original schedule, the service of first $N_k - 1$ customers is not finished by the time $t_{N_k}^{N_k} + \delta$, then the modified schedule results to a cost saving of $C\delta$; this is an event that happens with probability $(1 - \alpha_k)$. Finally, if under the original schedule the last customer finds the main system busy at time $t_{N_k}^{N_k}$ and begins to be served before time $t_{N_k}^{N_k} + \delta$, then the modified schedule still results to some cost savings in the main system. Putting everything together, the cost savings of the modified schedule is at least $C\delta(1 - \alpha_k) - \delta$. Recall now that $\alpha_k \to 0$ and that $C > 1$. This implies that the cost savings will be positive when k is large enough, contradicting optimality of the original schedule. **Q.E.D.**

Let us consider a sequence of time intervals $[0, q_N)$ indexed by N, the number of customers. Denote by $A(q_N)$ the number of admissions in the interval $[0, q_N)$ under the optimal schedule. The following lemma concerns the average admission rate in interval $[0, q_N)$.

Lemma 4 *If $q_N \leq t_N^N$ for each N, and $q_N \to \infty$ as $N \to \infty$, then*

$$\liminf_{N\to\infty} \frac{A(q_N)}{q_N} \geq 1$$

Proof: We will show that if the number of admissions in $[0, q_N)$ is too small, then the main system will have long idle periods that can be exploited to reduce the costs, thus contradicting optimality. In order to describe sparsity of admissions in the schedule π^N we define a deterministic function $U_{\pi^N}(t)$ describing the unfinished work at the main system at each time t under this particular schedule with the server replaced by a deterministic one. Thus, $U_{\pi^N}(t)$ decreases at unit rate whenever it is positive and has upward jumps of size 1 each time that a new customer is admitted into the main system. Suppose now that $\limsup_{N\to\infty} A(q_N)/q_N < 1$. Then, there exists some $\epsilon > 0$ and an increasing sequence $\{N_k\}$ such that $A(q_{N_k}) < (1-\epsilon)q_{N_k}$. Consider some value of N for which $A(q_N) < (1-\epsilon)q_N$. It is clear that the set $I = \{t \in [0, q_N) \mid U_{\pi^N}(t) = 0\}$ has measure at least ϵq_N. Let us split the interval $[0, q_N)$ into $q_N^{1/4}$ intervals of equal length and let L_i be the ith such interval. (For simplicity, we assume that $q_N^{1/4}$ is integer.) Let $I_i = \{t \in L_i \mid U_{\pi^N}(t) = 0\}$. Since the sum of the measures of the sets I_i is at least ϵq_N, it follows that there exists some i^* such that the measure of the set I_{i^*} is at least $\epsilon q_N^{3/4}/2$. We also claim that there exists

$$i^* \leq q_N^{1/4} - \lceil C \rceil - 2 \tag{7}$$

such that the measure of I_{i^*} is at least $\epsilon q_N^{3/4}/2$. The reason is that if not, then the sum of such measures in the first $q_N^{1/4} - \lceil C \rceil - 2$ intervals would be less than $\epsilon q_N/2$ and the sum of the measures of the last $\lceil C \rceil + 2$ intervals would be bounded above by $(C+3)q_N^{3/4}$. But for large N, $\epsilon q_N/2 + (C+3)q_N^{3/4} < \epsilon q_N$, which contradicts our assumption on the measure of the set I.

We now consider the following modification of the assumed optimal schedule. Let us define i^* as described above and consider the first customer admitted (under the optimal schedule π^N) after the end of interval $L_{i^*+\lceil C\rceil}$. Such a customer exists because we have already argued in (7) that L_{i^*} is not one of the last $\lceil C\rceil + 2$ intervals. We will refer to this customer as the "special customer". Under the modified schedule, this customer is to be admitted at the beginning of the interval L_{i^*}.

We now compare the costs under the two policies. Regarding the controller, under the modified schedule a customer is admitted earlier, thus resulting in some cost savings. Since each interval L_i has length at least $q_N^{3/4}$, and the customer is admitted earlier by at least $\lceil C\rceil + 1$ whole intervals. Therefore, the special customer is admitted earlier by $(\lceil C\rceil + 1)q_N^{3/4} + h_N$ for some $h_N \geq 0$. The cost savings in the controller is $(\lceil C\rceil + 1)q_N^{3/4} + h_N$. Let us now consider the cost change in the main system. To do the comparison, we use a coupling argument, by considering the sample paths resulting from the application of the two different policies while keeping the service time of each job the same. Due to the exponential service times, the cost at the main system is the same for any work–conserving queueing discipline. We therefore can, and will, assume that under both schedules, the special customer has the lowest service priority, and other customers have preemptive priority over this special customer. Due to the priority discipline we have assumed, it is clear that all customers except for the special one incur the same cost at the main system. We can therefore focus on the special customer. If under the modified schedule the service of the special customer is completed during the interval L_{i^*}, its cost at the main system is bounded above by $Cq_N^{3/4}$; denote the probability of this event by β_N. If the special customer is not finished during the interval L_i^*, it will nevertheless be served no later than the time it would be served under the optimal schedule. Thus, an upper bound on the excess cost of the modified schedule is C times the amount of time by which the admission of this customer has been hastened, which is $(\lceil C\rceil + 1)q_N^{3/4} + h_N$. To summarize, the cost savings resulting from the considered schedule modification is at least

$$(\lceil C\rceil + 1)q_N^{3/4} + h_N - \beta_N C q_N^{3/4} - (1-\beta_N)C\{(\lceil C\rceil + 1)q_N^{3/4} + h_N\}.$$

Recall that the measure of I_{i^*} is at least $\epsilon q_N^{3/4}/2$. We claim that the probability that the main system remains idle for at least $\epsilon q_N^{3/4}/6$ time units in L_{i^*} under the optimal schedule converges to 1 as $N \to \infty$. (This claim is essential for establishing this lemma and proved in Appendix A.) Therefore, for N large enough, β_N is sufficiently close to 1, so cost savings are positive, thus contradicting optimality of the original schedule. **Q.E.D.**

If we use $q_N = t_N^N$ in Lemma 4, $A(t_N^N) = N$ and from Lemma 3 $t_N^N \to \infty$ as $N \to \infty$, so Lemma 4 implies that $\liminf_{N\to\infty} \frac{N}{t_N^N} \geq 1$. From Lemma 3 we also have $\liminf_{N\to\infty} \frac{N}{t_N^N} \geq 1$. Therefore,

$$\lim_{N\to\infty} \frac{N}{t_N^N} = 1$$

Q.E.D.

4 First interadmission time

In this section we show that periodic admission is not an optimal schedule due to an edge effect. (Note that the main system is initially empty.) We will show that under an optimal scheudle, the first interadmission time becomes arbitrarily small as the number of customers increases. If a large number of customers are waiting in the controller, allowing idle time in the main system can cause big losses, so one would be more eager to admit a customer, in spite of the risk of congesting the main system. We can figuratively say that the large number of customers in the controller buffer pressures the controller to admit customers quickly. From now on, we denote the N customers that are initially in the controller by $e_1, e_2, \cdots, e_N$, and their admission times under an optimal schedule by $t_1^N \leq t_2^N \leq \cdots \leq t_N^N$. Obviously t_1^N is 0. The main result of this section is $\lim_N \ t_2^N = 0$.

In establishing this result, we will use interesting properties of an optimal schedule regarding the probability that the main system becomes idle in an early time interval. We denote by a left continuous function $X_N(t)$ the population of the main system at time t under the optimal schedule π^N. Let $[0, q_N)$ be a sequence of time intervals indexed by N. We will consider the following probability.

$$c_N \ \equiv \ P(\ X_N(t) > 0, \quad \forall t \in (t_3^N, q_N) \mid \ X_N(t_3^N) = 1\ ) \tag{8}$$

Lemma 5 *If for some* $\epsilon \in (0, 1/C)$

$$q_N \leq \left(\frac{1}{C} - \epsilon\right) N \qquad \forall N, \tag{9}$$

then $\{c_N\}$ *is bounded below by a positive number.*

Proof: In proving this we will often compare the optimal schedule π^N with the modified schedule $\tilde{\pi}^N$ for which e_N is admitted at time $t = 0$ instead of t_N^N. We will show that if there exists an increasing sequence $\{N_k\}$ such that $c_{N_k} \to 0$ as $k \to \infty$, the modified schedule incurs less cost for some N_k, thus contradicting optimality.

As a result of the modification, the cost incurred in the controller is reduced by t_N^N. Now we consider the increase of the cost in the main system. Note that due to the exponential service time the expected cost does not depend on service disciplines as long as they are work–conserving. For both schedules, we assume that customer e_N has the lowest priority, and other customers can preempt e_N. Then, the total increase of cost in the main system is C times the increase of e_N's response time. Denote by a left continuous function $\tilde{X}_N(t)$ the population of the main system at time t under this modified schedule $\tilde{\pi}^N$. If the main system becomes idle before time q_N under the modified schedule $\tilde{\pi}^N$, the response time of e_N is less than q_N under this schedule $\tilde{\pi}^N$, thus the cost increase is at most q_N. If the main system stays busy until time

q_N under $\tilde{\pi}^N$, the expected response time of e_N conditioned on this event is no more than $q_N + N$. The reason is that service times are memoryless, and there are only N customers. Combining all these, we see that the total increase of expected response time in the main system as a result of the schedule modification is at most

$$\begin{aligned} & P(\ \tilde{X_N}(t) > 0,\ \forall t < q_N\ )\ (q_N + N)\ + \Big\{\ 1 - P(\tilde{X_N}(t) > 0,\ \forall t < q_N\ )\ \Big\}\ q_N \\ = \ & q_N\ +\ P(\ \tilde{X_N}(t) > 0,\ \forall t < q_N\ )N \end{aligned}$$

Therefore, the increase of cost due to the schedule modification is

$$-t_N^N + CNP(\ \tilde{X_N}(t) > 0,\ t < q_N\ )\ + Cq_N\ ,$$

which is bounded above by

$$-t_N^N + CNP(\ \tilde{X_N}(t) > 0,\ t < q_N\ )\ + C(1/C - \epsilon)N$$

due to hypothesis (9). From Theorem 2, for any $\gamma > 0$, we have $t_N^N > (1-\gamma)N$ for sufficiently large N. Therefore, for any $\gamma > 0$, if N is sufficiently large, the increase of cost is bounded above by

$$\Big[\gamma + CP(\ \tilde{X_N}(t) > 0,\ \forall t < q_N\ ) - C\epsilon\Big]\ N \tag{10}$$

Define the probability that the main system remains busy in time interval (t_3^N, q_N) under $\tilde{\pi}^N$, given that there is no service completion in $[0, t_3^N]$ (recall that under $\tilde{\pi}^N$, customers e_1,e_N, e_2 are admitted in $[0, t_3^N)$):

$$\tilde{c}_N\ \equiv\ P(\ \tilde{X}_N(t) > 0,\ \forall t \in (t_3^N, q_N\ )\ |\ \tilde{X_N}(t_3^N) = 3\ )$$

Then, $P(\ \tilde{X_N}(t) > 0,\ \forall t < q_N\ ) \leq \tilde{c}_N$ for each N. Suppose that there exists a sequence $\{N_k\}$ such that $c_{N_k} \to 0$ as $k \to \infty$. We argue that this implies $\tilde{c}_{N_k} \to 0$ as $k \to \infty$ (proof in appendix B), and thus $P(\ \tilde{X_N}(t) > 0,\ \forall t < q_{N_k}\ ) \to 0$. For each k, compare $\tilde{\pi}^{N_k}$ with π^{N_k}. In expression (10), if we pick $\gamma < C\epsilon$, the cost of schedule $\tilde{\pi}^{N_k}$ is less than that of π^{N_k} for sufficiently large k. This contradicts optimality of the original schedule. Therefore, $\{c_N\}$ is bounded below by a positive number. **Q.E.D.**

Lemma 6

$$t_2^N\ \leq\ \ln C \quad \textit{and} \quad t_3^N \leq t^* \qquad \textit{for each } N,$$

where t^* is the unique solution to equation

$$(C + t - \ln C)\exp(-t)\ =\ 1/C \tag{11}$$

Proof: Assume the first–come–first–serve discipline. Suppose that $t_2^N > \ln C$. Consider hastening the admission time of customers $e_2, e_3, \cdots, e_N$ by some $\delta < t_2^N - \ln C$. Then, the cost in the controller is reduced by $\delta(N-1)$. The cost incurred in the main system increases only if the service time of e_1 is longer than $t_2^N - \delta$, of which the probability is $\exp\{-(t_2^N - \delta)\} < 1/C$. Therefore, the expected cost increase in

the main system is less than $(1/C)C\delta(N-1)$. Putting two terms together, the total cost is reduced, and this contradicts optimality. Therefore, $t_2^N \leq \ln C$.

Consider the event, $\mathcal{A}$, that the system main system is busy at time $t > \ln C$ serving e_1 or e_2. The probability of this event under π^N is bounded above by the one under another schedule $\bar{\pi}^N$ that admits e_1 at time 0 and e_2 at time $\ln C$. That is,

$$\begin{aligned} & P(\mathcal{A};\pi^N) \\ \leq\ & P(\mathcal{A};\bar{\pi}^N) \\ =\ & \exp(-t) + \{1-\exp(-\ln C)\}\exp(\ln C - t) + \int_{\ln C}^{t} \exp(-\tau)\exp(\tau - t)d\tau \\ =\ & (C + t - \ln C)\exp(-t) \end{aligned}$$

The function $(C + t - \ln C)\exp(-t)$ is monotonically decreasing in t for $t > 0$, and assumes at $t = 0$ the value $C - \ln C$, which is greater than $1/C$, and decays to 0 as $t \to \infty$. Therefore, there is a unique solution, say t^*, to equation (11), and for any $t > t^*$, $P(\mathcal{A};\pi^N) < 1/C$. Suppose $t_3^N > t^*$. Then, by hastening the admissions of customers $e_3, e_4, \cdots, e_N$ by a sufficiently small amount we can decrease the total cost because the expected increase of the cost in the main system is less than the cost savings at the controller. This contradicts optimality. **Q.E.D.**

Now we can show that the first interadmission time becomes arbitrarily small for a large N.

Theorem 7

$$\lim_{N\to\infty} \quad t_2^N = 0$$

Proof: Suppose not. Then, there is an increasing sequence $\{N_k\}$ such that $\{t_2^{N_k} | k = 1, 2, 3, \cdots\}$ is bounded below by a positive number. We will show that by hastening admission of e_2 we can reduce the cost when N_k is large, thus contradicting optimality.

Compare the optimal schedule π^N with the modified schedule $\hat{\pi}^N$ that admits e_2 at time 0 and keeps the admission times of other customers the same as in π^N. By changing the admission time of e_2, we decrease the cost in the controller by

$$t_2^N \tag{12}$$

Let us consider the change of cost incurred in the main system. Since the service time is memoryless, the cost does not depend upon the service discipline as long as it is work–conserving. We assume that e_2 has the lowest priority, and all other customers can preempt e_2. Then, the expected change of the cost is C times the expected change of e_2's response time. If the service time of e_1, say τ, is longer than t_2^N, e_2's response time increases by

$$t_2^N \tag{13}$$

Now we consider the expected response time of e_2 for the case, $\tau < t_2^N$. We define the expected queue depletion time under the optimal schedule π^N given that there is one customer in the main system immediately prior to t_3^N:

$$R_N = E\left[\ \inf\{t \geq t_3^N | X_N(t) = 0\}\ |\ X_N(t_3^N) = 1\ \right] - t_3^N$$

Under π^N, the expected response time of e_2 is more than $\exp(t_2^N - t_3^N)R_N$. Under $\hat{\pi}^N$, the expected response time of e_2 is bounded above by $t_3^N + \exp(\tau - t_3^N)R_N$. Therefore, the difference of the expected response times between two schedules is at most

$$t_3^N + \{\exp(\tau - t_3^N) - \exp(t_2^N - t_3^N)\}\ R_N$$

which equals

$$t_3^N + \exp(-t_3^N)\{\exp(\tau) - \exp(t_2^N)\}R_N \tag{14}$$

Taking weighted average of terms (12)(13)(14) by their probabilities, we have the following upper bound on the expected increase of the total cost:

$$\begin{aligned} &-t_2^N + \ Ct_2^N \exp(-t_2^N) + \\ &\quad C\int_0^{t_2^N} \exp(-\tau)[\ t_3^N + R_N \exp(-t_3^N)\{\exp(\tau) - \exp(t_2^N)\}\ ]d\tau\ , \end{aligned}$$

which is bounded above by

$$Ct_2^N \exp(-t_2^N) + \ Ct_2^N t_3^N \ \ - \ \ CR_N \exp(-t_3^N)\int_0^{t_2^N} \{\ \exp(t_2^N - \tau) - 1\ \}d\tau \tag{15}$$

The first two terms of expression (15) are bounded above over all N from Lemma 6. For the last term, we have

$$R_N \ \geq \ P(\ \ X_N(t) > 0,\ t_3^N \leq t \leq \frac{N}{2C}\ \ |\ X_N(t_3^N) = 1\ \ )\ \ \left(\frac{N}{2C} - t_3^N\right)$$

It follows from Lemmas 5 and 6 that for some $\delta > 0$, we have $R_N \geq \delta N$ for sufficiently large N. From Lemma 6, t_3^N is bounded above, so $R_N \exp(-t_3^N)$ grows unbounded with N. Suppose that $\{t_2^{N_k} | k = 1, 2, \cdots\}$ is bounded below by a positive number. Then,

$$\int_0^{t_2^{N_k}} [\ \exp(t_2^{N_k} - \tau) - 1\ ]d\tau$$

is also bounded below by a positive number. Thus, the last term of expression (15) blows to $-\infty$ as N_k grows. Therefore, the change of cost becomes negative for a large N_k, contradicting optimality. Therefore, $\lim_{N\to\infty} t_2^N = 0$. **Q.E.D.**

5 Intermediate admission rate

We have discussed the overall admission rate in section 3. In this section we discuss the partial admission rate averaged over early time intervals under an optimal schedule.

In section 3, Lemma 4 was used as a stepping stone to establish the limit of the admission rate averaged over entire admission history. We point now that this lemma also provides results on admission rate averaged over partial time intervals. As long as the intervals $[0, q_N]$ in the sequence are short enough relative to N (namely, less than t_N^N), and q_N grows to ∞ as N grows, then the asymptotic admission rate averaged over this sequence of intervals is at least the service rate. In this section we explore the flip side of this lemma. Namely, we show that admission rate averaged over a sufficiently long portion of the early admission history is is bounded above the service rate in the asymptote.

Theorem 8 *If* $(\ln N)/q_N \to 0$ *as* $N \to \infty$, *then,*

$$\limsup_{N\to\infty} \frac{A(q_N)}{q_N} \leq 1$$

Proof: For each N, consider a modified schedule for which the admission of $e_{A(q_N)}$ is delayed by 1, and other customers' admission times are unchanged. The cost in the controller increases by 1 as a result of the modification. Now we examine the cost change in the main system. For both the optimal and the modified schedule, we assume that $e_{A(q_N)}$ has the lowest priority, and other customers can preempt it. For other customers, the service discipline is the first-come-first-serve. The cost change in the main system is C times the change of this special customer's response time.

Consider the event, denoted by $\mathcal{B}$, that under the optimal schedule π, the main system is kept busy in the time interval $[t^N_{A(q_N)}, t^N_{A(q_N)} + 1]$ with customers that have been admitted before $e_{A(q_N)}$. Note that $P(\mathcal{B}) \geq P(\sum_{i=1}^{A(q_N)-1} \xi_i > q_N + 1)$, where ξ_i is the service time of e_i. If this event $\mathcal{B}$ happens, the response time of $e_{A(q_N)}$ decreases by 1 as a result of schedule modification. Even in the case that $\mathcal{B}$ does not happen, the conditional expectation of the increase of $e_{A(q_N)}$'s response time cannot be more than $N-1$ because there are only $N-1$ other customers. Therefore, the change of cost Δ_N satisfies

$$\begin{aligned}
\Delta_N &\leq 1 - CP(\mathcal{B}) + C(N-1)\{1 - P(\mathcal{B})\} \\
&= 1 - C + CN\{1 - P(\mathcal{B})\} \\
&\leq 1 - C + CNP(\sum_{i=1}^{A(q_N)-1} \xi_i < q_N + 1) \qquad (16)
\end{aligned}$$

Due to optimality of the original schedule, $\Delta_N \geq 0$, so expression (16) must be nonnegative, and we have

$$\frac{C-1}{CN} \leq P(\sum_{i=1}^{A(q_N)-1} \xi_i < q_N + 1)$$

Using the Chernoff bound [12], we have

$$\frac{C-1}{CN} \leq P(\sum_{i=1}^{A(q_N)-1} \xi_i < q_N + 1) \leq \exp(s(q_N+1)) \left(\frac{1}{s+1}\right)^{A(q_N)-1} \qquad \forall s > 0$$

This implies that

$$\ln\left(\frac{C-1}{C}\right) - \ln(N) \;\leq\; s(q_N+1) - (A(q_N)-1)\ln(s+1), \qquad \text{and thus}$$

$$\frac{A(q_N)-1}{q_N} \leq \frac{s}{\ln(s+1)}\left(1+\frac{1}{q_N}\right) + \frac{\ln(N) - \ln(\frac{C-1}{C})}{q_N \ln(s+1)} \quad , \qquad \forall s > 0$$

Since $q_N \to \infty$ and $(\ln N)/q_N \to 0$ as $N \to \infty$,

$$\limsup_{N\to\infty} \frac{A(q_N)}{q_N} \leq \frac{s}{\ln(s+1)} \;, \quad \forall s > 0$$

By taking the limit as $s \to 0$, we have

$$\limsup_{N\to\infty} \frac{A(q_N)}{q_N} \leq 1$$

Q.E.D.

Combining Lemma 4 and Theorem 8, we can establish a convergence result for the admission rate averaged over an initial time interval $[0, q_N]$ for some range of q_N. If the sequence $\{q_N\}$ grows faster than the order of $\ln N$, and $q_N \leq N$, then we have

$$\lim_{N\to\infty} \frac{A(q_N)}{q_N} = 1$$

In words, the admission rate of the optimal schedule asymptotically becomes the service rate. For a sequence of intervals whose length increases slower than $\ln N$, we have only shown (Lemma 4) that the admission rate is no less than the service rate, asymptotically.

6 Discussion

In this section, we briefly discuss outstanding open problems. In Theorem 1, we showed that periodic admission becomes asymptotically close to the optimal schedule in performance as long as interadmission times are properly adjusted for growing number of customers. In the proof of this theorem, we saw that the dominant term of the total cost is the term $N(N-1)/2$ that comes from waiting time at the controller. We derived an asymptotically optimal policy whose cost was $N(N-1)/2$ plus lower order terms. It would be interesting, although quite hard, to derive a policy that matches both the quadratic and the linear part of the optimal cost function.

The proof of Theorem 7 may be extended to show that $\lim_{N\to\infty} t_k^N = 0$ for any fixed k. This implies that the number of customers admitted during the time interval $[0, 1]$

converges to infinity as N increases. Based on this observation, we are led to consider policies of the following type: admit K_N customers at time 0; from then on, admit a customer once every r_N time units. We conjecture that once the values of K_N and r_N are properly adjusted, such policies will lead to a better linear term, compared to the policy of Subsection 3.1 in which we had $K_N = 1$.

A Appendix to Lemma 4

Recall that we assumed $A(q_N) < (1-\epsilon)q_N$ in order to prove the contraposition. Also recall that the measure of I_{i^*} is at least $\epsilon q_N^{3/4}/2$.

Let $v_N = \inf\{\tau \in L_{i^*} \mid U_{\pi^N}(\tau) = 0\}$ and $z_N = \sup\{\tau \in L_{i^*} \mid U_{\pi^N}(\tau) = 0\}$. Let k_N be the number of admissions during the interval $[0, v_N]$ and let ℓ_N be the number of admissions during the interval $[v_N, z_N]$. Clearly, the subset of L_{i^*} on which $U_{\pi^N}(t) = 0$ is contained in $[v_N, z_N]$. Therefore,

$$\ell_N \leq z_N - v_N - \epsilon q_N^{3/4}/2. \tag{17}$$

Furthermore,

$$k_N \leq v_N \tag{18}$$

Let T_i^N be the time at which the service of the ith customer ends. Let S_N be the sum of the service times of the customers admitted during $[v_N, z_N]$. Suppose that the following two events occur (a) $T_{k_N}^N \leq v_N + \epsilon q_N^{3/4}/6$ and (b) $S_N \leq z_N - v_N - 2\epsilon q_N^{3/4}/6$. Then the interval $[v_N, z_N]$ consists of at most $\epsilon q_N^{3/4}/6$ time units spent to serve customers admitted before time v_N and at most $z_N - v_N - 2\epsilon q_N^{3/4}/6$ time units spent to serve customers admitted during $[v_N, z_N]$. It follows that there are at least $\epsilon q_N^{3/4}/6$ units of idle time during that interval. Now we only need to show that the probability that both events (a) and (b) happen converges to 1. Note that S_N is the sum of ℓ_N exponential random variables. Its mean is at most $z_N - v_N - \epsilon q_N^{3/4}/2$ (because of inequality (17)) and its standard deviation is at most $Bq_N^{1/2}$, where B is a constant independent of N. (This is because $\ell_N \leq A(q_N) < (1-\epsilon)q_N$.) For event (b) not to occur, S_N must be at least $\epsilon q_N^{1/4}/(6B)$ standard deviations above its mean and the probability of this happening goes to zero, by the Chebychev inequality. Therefore, the probability of event (b) goes to 1. For event (a) not to occur, the sum of the service times of the first k_N customers must exceed $v_N + \epsilon q_N^{3/4}/6$. On the other hand, the mean of this sum of these service times is bounded above by v_N [cf. inequality (18)], and it easily follows (as in the case of event (b)) that the probability of event (a) also converges to 1. Since these two events are independent, the probability that both events happen converges to 1. **Q.E.D.**

B Appendix to Lemma 5

Recall

$$c_N \equiv P(\, X_N(t) > 0,\ \forall t \in (t_3^N, q_N\,) \mid\ X_N(t_3^N) = 1\,)$$

$$\tilde{c}_N \equiv P(\ \tilde{X}_N(t) > 0,\ \forall t \in (t_3^N, q_N\ )\ |\ \tilde{X}_N(t_3^N) = 3\ )$$

Let $\{c_{N_k}\}$ be an arbitrary subsequence of $\{c_N\}$. We will show that if $\lim_{k\to\infty} c_{N_k} = 0$, then $\lim_{k\to\infty} \tilde{c}_{N_k} = 0$.

Proof: Note that c_N and $\tilde{c}_N$ do not depend on the service priority due to the memoryless service time distribution. In this proof we assume the first-come-first-serve priority for both schedules π^N and $\tilde{\pi}^N$. Define a random variable A_3 to be the time e_3 is ready to be served in the optimal schedule π^N. Note that under schedule π^N, e_1 and e_2 must be served before e_3 is ready to be served. Define $\tilde{A}_3$ to be the time the service of e_3 is ready to be served in the schedule $\tilde{\pi}^N$. Note that under schedule $\tilde{\pi}^N$, e_1,e_N, and e_2 must be served before e_3 is ready to be served. Also, define

$$v_N(\tau) \equiv P(X(t) > 0,\ \forall t \in (t_3^N, q_N)\ |\ X(t_3^N) = 1,\ A_3 = \tau + t_3^N\ ) \quad (19)$$

$$\hat{v}_N(\tau) \equiv P(\tilde{X}(t) > 0,\ \forall t \in (t_3^N, q_N)\ |\ \tilde{X}(t_3^N) = 3,\ \tilde{A}_3 = \tau + t_3^N\ ) \quad (20)$$

Since $\tilde{\pi}^N$ has one less customer to admit than π^N has, after t_3^N, we obtain $\hat{v}_N(\tau) \leq v_N(\tau)$, $\forall \tau > 0$. Now,

$$c_N = \int_0^\infty \exp(-\tau) v_N(\tau) d\tau, \text{ and} \quad (21)$$

$$\tilde{c}_N = \frac{1}{2!}\int_0^\infty \tau^2 \exp(-\tau)\hat{v}_N(\tau)d\tau \leq \frac{1}{2!}\int_0^\infty \tau^2 \exp(-\tau) v_N(\tau) d\tau \quad (22)$$

Suppose $c_{N_k} \to 0$ as $k \to \infty$. Define $r_{N_k} \equiv 1/\sqrt{c_{N_k}}$; then, $r_{N_k} \to \infty$ and $r_{N_k} c_{N_k} \to 0$ as $k \to \infty$. Let $T_k \equiv \sqrt{r_{N_k}}$; then,

$$\tilde{c}_{N_k} \leq \frac{1}{2!}\int_0^{T_k} \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau + \frac{1}{2!}\int_{T_k}^\infty \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau$$

The right hand side of this inequality has two terms. We now claim that both terms converge to 0 as k increases. Consider the first term.

$$\begin{aligned}
\frac{1}{2!}\int_0^{T_k} \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau &\leq \frac{1}{2!} T_k^2 \int_0^{T_k} \exp(-\tau) v_{N_k}(\tau) d\tau \\
&= \frac{1}{2!} r_{N_k} \int_0^{T_k} \exp(-\tau) v_{N_k}(\tau) d\tau \\
&\leq \frac{1}{2!} r_{N_k} \int_0^{\infty} \exp(-\tau) v_{N_k}(\tau) d\tau \\
&= \frac{1}{2!} r_{N_k} c_{N_k}
\end{aligned}$$

Since $r_{N_k} c_{N_k} \to 0$ as k increases, the first term converges to 0. As for the second term

$$\frac{1}{2!}\int_{T_k}^\infty \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau \leq \frac{1}{2!}\int_{T_k}^\infty \tau^2 \exp(-\tau) d\tau$$

Since T_k grows infinitely as k increases, the second term converges to 0 as k increases. Therefore, $\tilde{c}_{N_k} \to 0$ as $k \to \infty$. **Q.E.D.**

References

[1] J. S. Baras, D.-J. Ma, and A. M. Makowsky. K competing queues with linear costs and geometric service requirements: the μc–rule is always optimal. *Syst. Control Letters*, 6:173–180, 1985.

[2] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, Englewood Cliffs, NJ, second edition, 1991.

[3] F. Beutler and D. Teneketzis. Routing in queueing networks under imperfect information: stochastic dominance and thresholds. *Stochastics and Stochastics Reports*, 26:81–100, 1989.

[4] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$–rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[5] A. Ephremides, P. Varaiya, and J. C. Walrand. A simple dynamic routing problem. *IEEE Transactions on Automatic Control*, 25(8):690–693, August 1980.

[6] B. Hajek. Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control*, 29(6):491–499, June 1984.

[7] M. Hsiao and A. A. Lazar. Optimal flow control of multiclass queueing networks with decentralized information. Technical report, Columbia University, Department of Electrical Engineering, New York, 1986.

[8] L. Kleinrock. *Queueing Systems.* John Wiley & Sons, New York, London, Sydney, Toronto, 1975.

[9] P. R. Kumar and W. Lin. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*, 29(8):696–703, August 1984.

[10] A. A. Lazar. Optimal flow control of a class of queueing networks in equilibrium. *IEEE Transactions on Automatic Control*, 28(11):1001–1007, November 1983.

[11] D. C. Lee. *On open–loop admission control into a queueing system.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1992. Dept. of Electrical Engineering and Computer Science.

[12] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, New York, second edition, 1984.

[13] P. D. Sparaggis, D. Towsley, and C. G. Cassandras. Optimality of static routing policies in queueing systems with blocking. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 809–814, Brighton, England, December 1991.

[14] G. D. Stamoulis and J. N. Tsitsiklis. Optimal distributed policies for choosing among multiple servers. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 815–820, Brighton, England, December 1991.

[15] S. Stidham, Jr. Optimal control of admissions to a queueing system. *IEEE Transactions on Automatic Control*, 30(8):705–713, August 1985.

[16] J. Walrand. *An Introduction to Queueing Networks.* Prentice–Hall, Englewood Cliffs, New Jersey, 1988.