Optimally Scheduling Admission of N Customers from Low–Price Buffer to High-Price Queue, Part II: Edge Effect

Daniel C. Lee
Department of Electrical Engineering, University of Southern California
3740 McClintock Ave., Los Angeles, CA 90089, USA

Abstract

This paper continues to study the problem of optimally admitting a finite number customers from an auxiliary buffer with a low holding cost into a single–server queue with a high holding cost without observing the status of the single-server queue, which was introduced in Part I. This paper first proves a few asymptotic properties of the sequence of optimal schedules indexed by the number of customers. In particular, it is proven that the first inter-admission time converges to 0 as the number of customers increases. Then, a number of optimal schedules were numerically computed for different cost ratio and the number of customers. On the basis of the numerical results and queueing theoretic intuition, it is conjectured that in each optimal schedule the inter-admission times are monotonically non-decreasing if arranged in chronological order.

Key words: optimal queue admission, inter-admission time, queueing delay, stochastic scheduling

1 Introduction

This paper continues to examine the the problem of optimally admitting a finite number of customers from an auxiliary buffer with a low holding cost into a single–server queue with a high holding cost without observing the status of the single-server queue, which was introduced in Part I [3]. The optimization problem is described as

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^{N} \left[\ \tilde{t}_i^N + E(\tilde{R}_i)\ \right] \\ \text{subject to} & \tilde{t}_i^N \geq 0, i = 1, 2, \cdots, N \end{array}$$

or equivalently

$$\begin{array}{ll} \text{minimize} & g(x_2, x_3, \cdots, x_N) \\ \text{subject to} & x_i \geq 0, i = 2, 3, \cdots, N \end{array} \tag{1}$$

where

$$\begin{aligned} & g(x_2, x_3, \cdots, x_N) \\ \equiv \ & \sum_{i=2}^{N} (N - i + 1) x_i + C \sum_{i=1}^{N} E(\tilde{R}_i) \end{aligned} \tag{2}$$

Note that $E(\tilde{R}_i)$ can be viewed either as a function of $(\tilde{t}_1^N, \tilde{t}_2^N, \cdots, \tilde{t}_N^N)$ or $(x_2, x_2, \cdots, x_N)$. The simplicity of the problem easily leads to some questions and speculations on the optimal schedules. For example, if the controller were able to observe the queue length of main system, the optimal feedback admission control would be obviously that the controller admits a customer whenever the main system becomes empty. Thus, the inter-admission times would depend on the random service times of customers, and the expected value of these inter-admission times would be all identical to the expected value of the service time, 1. Compared with such a closed-loop control problem, the controller cannot observe the main system in the scheduling problem addressed by this paper, and the admission control is static (off–line) in nature. The auxiliary buffer initially holds the full load of customers, so the cost at the auxiliary buffer accrues at a high rate. As the auxiliary buffer unloads these customers to the main system, the cost burden at the auxiliary buffer lessens. At the same time, the main system, which is initially empty, begins to have positive probability of being crowded as the auxiliary buffer unloads the customers to the main system. Therefore, inter-admission times at the early stage of the optimal schedule are speculated to be shorter than those at the later stage. This paper explores such an edge effect. Constructing mathematically valid statements is found to be rather difficult. This paper can be viewed as an effort to produce and prove mathematically concrete statements regarding the edge effect. Part I [3] has established that in optimal scheduling for a sufficiently large number of customers the time to complete all the admissions should be at least the total expected service time of all the customers. Although Part I [3] shows that sufficiently long time is taken to admit the whole set of customers optimally, Section 2 in the present paper proves that in optimal scheduling for a sufficiently large number of customers the first inter-admission time ($t_2^N - t_1^N = t_2^N \equiv x_2^*$) is arbitrarily close to 0. Such a proposition in this paper were partially motivated by the question of whether there exists an atomic time unit based on which one can discretize the problem of optimal scheduling and still schedule various numbers of admissions optimally. The results in this paper show that there is no positive lower bound on the inter-admission times of the optimal schedules; thus, there is no such an atomic time unit. In section 3, the method of numerical computation of the optimal schedule and the results of numerical evaluations are discussed.

2 First Inter-admission Time

In this section we show an edge effect displayed in optimal schedules. (Recall that the main system is initially empty.) Even if it takes the total expected service time to complete the admission for the case of a large number of customers under an optimal schedule (Theorem 1 of Part I), the first two adjacent admission will be proven to be very close together in an optimal schedule. We will show that the first inter-admission time becomes arbitrarily small as the number of customers increases. Recall that $t_1^N = 0 \le t_2^N \le t_3^N \le \cdots, \le t_N^N$ are defined to be the admission times of an optimal schedule. The main result of this section is $\lim_{N\to\infty} \quad t_2^N = 0$.

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. Recall that we denote by a left–continuous function $X_{\pi^N}(t)$ the population of the main system at time t under the optimal schedule π^N. Consider a sequence of positive numbers, $\{q_N\}$ indexed by N. We will consider the conditional probability that the main system stays busy from time t_3^N to q_N given that e_3 sees only one customer upon entry:

$$\begin{aligned} & c_N \\ \equiv \quad & P[\ X_{\pi^N}(t) > 0, \forall t \in (t_3^N, q_N) | X_{\pi^N}(t_3^N) = 1] \end{aligned} \quad (3)$$

Lemma 1 *If for some $\epsilon > 0$,*

$$q_N \le \left(\frac{1}{C} - \epsilon\right) N \qquad \forall N, \quad (4)$$

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 $\hat{\pi}^N$ for which e_N is admitted at time $t = 0$ instead of t_N^N;

$$\begin{aligned} \pi^N &= (\ 0, t_2^N, t_3^N, \cdots, t_{N-1}^N, t_N^N\ ) \\ \hat{\pi}^N &= (\ 0, t_2^N, t_3^N, \cdots, t_{N-1}^N, 0\ ) \end{aligned}$$

Note that there are at least two admissions at time 0 under the modified schedule $\hat{\pi}^N$. As a result of the modification, the cost incurred in the controller is reduced by t_N^N;

$$f_a(\hat{\pi}^N) - f_a(\pi^N) \quad = \quad -t_N^N \quad (5)$$

Now we compare the costs associated with the two schedules in the main system. Recall that the cost does not depend on service disciplines as long as they are work–conserving and do not discriminate against the service times. For both schedules, we assume that customer e_N has the lowest priority, and other customers can preempt e_N. For the cost comparison in the main system, we use the coupling argument; We use an identical sample path of $S_1, S_2, \cdots$ for both schedules in our analysis. We denote by $R_i(\pi^N)$ and $R_i(\hat{\pi}^N)$ the response times of customer e_i for schedules π^N and $\hat{\pi}^N$, respectively, both under this service discipline. Because other customers can preempt customer e_N, customer e_N does not impede the service of other customers. Also, the admission time of each customer except e_N remains unchanged in the schedule modification. As a result, we have

$$R_i(\pi^N) = R_i(\hat{\pi}^N), \quad \forall i \ne N$$

for each realization of the set of random service variables $S_1, S_2, \cdots, S_N$. Then, total increase of cost in the main system is $CE\{R_N(\hat{\pi}^N) - R_N(\pi^N)\}$. We denote by a left–continuous function $X_{\hat{\pi}^N}(t)$ the population of the main system at time t under the modified schedule $\hat{\pi}^N$. Consider the event, which is determined by random variables $S_1, S_2, \cdots, S_N$, that the main system becomes idle before time q_N under the modified schedule $\hat{\pi}^N$; denote this event by $\mathcal{E}_N$. In this event $\mathcal{E}_N$, $R_N(\hat{\pi}^N)$ is less than q_N because e_N has to depart from the system before time q_N for this event to occur. (Recall that e_N is admitted at time 0 in schedule $\hat{\pi}^N$.) Therefore, we have $E[R_N(\hat{\pi}^N)\ |\mathcal{E}_N\ ] \le q_N$, and thus

$$E[\ R_N(\hat{\pi}^N) - R_N(\pi^N)\ \ |\ \ \mathcal{E}_N\ ]\ \le\ q_N$$

Consider the complementary event $\mathcal{E}_N^c$ that the main system stays busy until time q_N under $\hat{\pi}^N$. In this event, e_N is still in the main system at time q_N, and the response time $R_N(\hat{\pi}^N)$ is q_N plus the time from q_N till the system becomes idle. (Recall that e_N has the lowest preemptive priority.) Due to memoryless property of the exponential distribution, the the expected time conditioned on event $\mathcal{E}_N^c$ between adjacent departures after q_N is still 1 while $X_{\hat{\pi}^N}(t)$ is positive. Also, note that there are no more than N customers. Therefore, we have $E[R_N(\hat{\pi}^N) \mid \mathcal{E}_N^c] \le q_N + N$, and thus

$$E[\ R_N(\hat{\pi}^N) - R_N(\pi^N) \ \mid\ \mathcal{E}_N^c\ ] \ \le\ q_N + N$$

Combining all these, we see that the increase of the cost in the main system as a result of the schedule modification is

$$\begin{aligned} & f_m(\hat{\pi}^N) - f_m(\pi^N) \\ \le\ & P(\mathcal{E}_N)Cq_N + P(\mathcal{E}_N^c)C(q_N + N) \\ =\ & Cq_N \ +\ CP(\ \mathcal{E}_N^c\ )N \end{aligned} \tag{6}$$

From (5) and (6), the change of cost due to the schedule modification is

$$\begin{aligned} & f_a(\hat{\pi}^N) - f_a(\pi^N) + f_m(\hat{\pi}^N) - f_m(\pi^N) \\ \le\ & -t_N^N + Cq_N \ +\ CP(\ \mathcal{E}_N^c\ )N \end{aligned}$$

Due to hypothesis (4), we have

$$\begin{aligned} & f_a(\hat{\pi}^N) - f_a(\pi^N) + f_m(\hat{\pi}^N) - f_m(\pi^N) \\ \le\ & -t_N^N + C(1/C - \epsilon)N + CP(\ \mathcal{E}_N^c\ )N \end{aligned}$$

From Theorem 1 in Part I [3], 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, we have

$$\begin{aligned} & f_a(\hat{\pi}^N) - f_a(\pi^N) + f_m(\hat{\pi}^N) - f_m(\pi^N) \\ \le\ & -(1-\gamma)N + C(1/C - \epsilon)N + CP(\ \mathcal{E}_N^c\ )N \\ =\ & [\gamma + CP(\ \mathcal{E}_N^c\ ) - C\epsilon]\, N \end{aligned} \tag{7}$$

Now we define conditional probability

$$\hat{c}_N \ \equiv\ P[\ X_{\hat{\pi}^N}(t) > 0,\ \forall t \in (t_3^N, q_N\ ) \mid X_{\hat{\pi}^N}(t_3^N) = 3\ ]$$

(Probability that the main system remains busy in time interval (t_3^N, q_N) under $\hat{\pi}^N$, conditioned on the event that there is no departure in $[0, t_3^N)$) Recalling again that e_N has the lowest preemptive priority, we have

$$\begin{aligned} & P(\ \mathcal{E}_N^c\ ) \\ \equiv\ & P[\ X_{\hat{\pi}^N}(t) > 0,\ \forall t \in (0, q_N)\ ] \end{aligned}$$

$$\begin{aligned} =\ & P[X_{\hat{\pi}^N}(t_3^N) = 3\ ]\hat{c}_N + P[X_{\hat{\pi}^N}(t_3^N) = 2\ ] \times \\ & P[\ X_{\hat{\pi}^N}(t) > 0,\ \forall t \in (t_3^N, q_N\ ) \mid X_{\hat{\pi}^N}(t_3^N) = 2\ ] + \\ & P[X_{\hat{\pi}^N}(t_3^N) = 1,\ S_1 + S_N \ge t_2^N\ ] \ \times \\ & P[\ X_{\hat{\pi}^N}(t) > 0,\ \forall t \in (t_3^N, q_N\ ) \mid \\ & X_{\hat{\pi}^N}(t_3^N) = 1,\ S_1 + S_N \ge t_2^N\ ] \\ \le\ & \{\ P[X_{\hat{\pi}^N}(t_3^N) = 3\ ] + P[X_{\hat{\pi}^N}(t_3^N) = 2\ ] + \\ & P[X_{\hat{\pi}^N}(t_3^N) = 1,\ S_1 + S_N \ge t_2^N]\ \}\hat{c}_N \\ \le\ & \hat{c}_N \end{aligned} \tag{8}$$

Suppose that there exists a sequence $\{N_k\}$ such that $c_{N_k} \to 0$ as $k \to \infty$. We argue that this implies $\hat{c}_{N_k} \to 0$ as $k \to \infty$ (proof in Appendix A), and thus $P(\ \mathcal{E}_{N_k}^c\ ) \to 0$ from inequality (8). For each k, compare $\hat{\pi}^{N_k}$ with π^{N_k}. From expression (7), if we pick $\gamma < C\epsilon$, the cost of schedule $\hat{\pi}^{N_k}$ is less than that of π^{N_k} for sufficiently large k. This contradicts optimality of schedule π^{N_k}. Therefore, $\{c_N\}$ is bounded below by a positive number. **Q.E.D.**

We now examine the following properties of the optimal schedule.

Lemma 2

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

where t^* is the unique solution to equation

$$(C + t - \ln C)\exp(-t) \ =\ 1/C \tag{9}$$

Proof: We use the first–come–first–serve (FCFS) discipline for our analysis for this proof. Suppose that $t_2^N > \ln C$. Then, we can consider modifying the optimal schedule π^N into another schedule $\breve{\pi}^N$, in which we hasten the admission times of customers $e_2, e_3, \cdots, e_N$ by some $\delta < t_2^N - \ln C$;

$$\begin{aligned} \pi^N \ &=\ (\ 0, t_2^N\ ,\ \ t_3^N\ ,\ \ \cdots,\ \ t_{N-1}^N\ ,\ \ t_N^N\ ) \\ \breve{\pi}^N \ &=\ (\ 0, t_2^N - \delta, t_3^N - \delta, \cdots, t_{N-1}^N - \delta, t_N^N - \delta\ ) \end{aligned}$$

Then, the cost in the controller is reduced by $\delta(N-1)$;

$$f_a(\breve{\pi}^N) - f_a(\pi^N) \ =\ -\delta(N-1) \tag{10}$$

For the cost comparison in the main system, we again use a coupling argument. (We use an identical sample path of $S_1, S_2, \cdots$ for both schedules in our analysis.) In the event $S_1 \le t_2^N - \delta$, the response times in the main system associated with π^N and $\breve{\pi}^N$ are the same. In the event $S_1 > t_2^N - \delta$, the total response time in the main system increases as a result of the schedule modification, but not more than by $C\delta(N-1)$. The random variable S_1 has an exponential distribution, so we have $P(S_1 > t_2^N - \delta) = \exp\{-(t_2^N - \delta)\}$ and

$$\begin{aligned} & f_m(\breve{\pi}^N) - f_m(\pi^N) \\ \leq\ & C\delta(N-1)\exp\{-(t_2^N - \delta)\} \end{aligned} \tag{11}$$

Combining (10) and (11), we have

$$\begin{aligned} & f_a(\breve{\pi}^N) - f_a(\pi^N) + f_m(\breve{\pi}^N) - f_m(\pi^N) \\ \leq\ & -\delta(N-1) + C\delta(N-1)\exp\{-(t_2^N - \delta)\} \\ =\ & \left[\ C\exp\{-(t_2^N - \delta)\}\ -\ 1\ \right]\ \delta(N-1) \end{aligned}$$

Because $t_2^N - \delta > \ln C$, we have

$$\begin{aligned} P(S_1 > t_2^N - \delta) &= \exp\{-(t_2^N - \delta)\} \\ &< \exp\{-(\ln C)\} = 1/C \end{aligned}$$

so the change of the total expected cost is negative. This contradicts the optimality of π^N. Hence, we proved $t_2^N \leq \ln C$.

Function $g(t) \equiv (C + t - \ln C)\exp(-t)$ is strictly monotone decreasing in t for $t > 0$, and we have $g(0) = C - \ln C > 1/C$ and $\lim_{t\to\infty} g(t) = 0$. Therefore, there is a unique solution, t^*, to equation (9). Also, $g(\ln C) = 1 > 1/C$, so $t^* > \ln C$. Suppose $t_3^N > t^*$. Then, pick a sufficiently small number δ with property $t_3^N - \delta > t^*$. Compare the optimal schedule π^N with another schedule π_1^N;

$$\begin{aligned} \pi^N &= (\ 0, t_2^N, t_3^N, \quad \cdots, t_{N-1}^N, \quad t_N^N\ ) \\ \pi_1^N &= (\ 0, t_2^N, t_3^N - \delta, \cdots, t_{N-1}^N - \delta, t_N^N - \delta\ ) \end{aligned}$$

Obviously,

$$f_a(\pi_1^N) - f_a(\pi^N) \quad = \quad -(N-2)\delta \tag{12}$$

In the event that the main system is empty under the optimal schedule π^N at time $t_3^N - \delta$, (i.e. $X_{\pi^N}(t_3^N - \delta) = 0$), the total response times in the main system do not change as a result of the schedule modification. In the event $X_{\pi^N}(t_3^N - \delta) > 0$, the increase of the total response time is no more than $(N-2)\delta$. Therefore,

$$\begin{aligned} & f_m(\pi_1^N) - f_m(\pi^N) \\ \leq\ & C\ P\left[\ X_{\pi^N}(t_3^N - \delta) > 0\ \right]\ (N-2)\delta \end{aligned} \tag{13}$$

From (12) and (13), the total change of the cost is

$$\begin{aligned} & f_a(\pi_1^N) - f_a(\pi^N) + f_m(\pi_1^N) - f_m(\pi^N) \\ \leq\ & \{\ C\ P\left[\ X_{\pi^N}(t_3^N - \delta) > 0\ \right] - 1\ \}\ \times \\ & \qquad (N-2)\delta \end{aligned} \tag{14}$$

We now consider another schedule, which delays the admission of e_2 to $\ln C$ and keeps the admission of other customers unchanged from π^N;

$$\pi_2^N = (0, \ln C, t_3^N, t_4^N, \cdots, t_N^N)$$

Then, for each $t \in (\ln C, t_3^N)$,

$$\begin{aligned} & P[X_{\pi^N}(t) > 0] \\ \leq\ & P[X_{\pi_2^N}(t) > 0] \\ =\ & P(S_1 \leq \ln C, S_2 > t - \ln C) + \\ & \qquad P(S_1 > \ln C, S_1 + S_2 > t) \\ =\ & (C + t - \ln C)\exp(-t) \equiv g(t)\exp(-t) \end{aligned}$$

Because $t_3^N - \delta > t^*$ and $g(t)$ is strictly decreasing in t

$$\begin{aligned} & P[X_{\pi^N}(t_3^N - \delta) > 0] \\ \leq\ & g(t_3^N - \delta)\exp[-(t_3^N - \delta)] \\ <\ & g(t^*)\exp(-t^*) = 1/C \end{aligned}$$

Therefore, the cost change in (14) is negative;

$$f_a(\pi_1^N) - f_a(\pi^N) + f_m(\pi_1^N) - f_m(\pi^N) < 0$$

This contradicts the optimality of π^N. Therefore, $t_3^N \leq t^*$. **Q.E.D.**

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

Theorem 1

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

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 then 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 $\bar{\pi}^N$ that admits e_2 at time 0 and keeps the admission times of other customers the same as in π^N;

$$\begin{aligned} \pi^N &= (0, t_2^N \quad , \quad t_3^N, t_4^N, \cdots, t_N^N) \\ \bar{\pi}^N &= (0,\ 0 \quad , \quad t_3^N, t_4^N, \cdots, t_N^N) \end{aligned}$$

By changing the admission time of e_2, we decrease the cost in the controller by t_2^N;

$$f_a(\bar{\pi}^N) - f_a(\pi^N) \quad = \quad -t_2^N \tag{15}$$

Now, we consider the change of cost incurred in the main system. Regarding the service discipline, we assume without affecting the cost incurred in the main system that e_2 has the lowest priority, and all other customers can preempt e_2. All customers other than e_2 are served according the the FCFS discipline. Then, we have

$$R_i(\bar{\pi}^N) = R_i(\pi^N), \quad \forall i \neq 2,$$

so the expected change of the cost is

$$f_m(\bar{\pi}^N) - f_m(\pi^N) = CE[R_2(\bar{\pi}^N) - R_2(\pi^N)]$$

In the event $S_1 \geq t_2^N$, under the modified schedule $\bar{\pi}$ e_2 is not served in $[0, t_2^N]$ but waits in the main system, so e_2's response time increases by t_2^N as a result of the schedule modification;

$$E\left[\ R_2(\bar{\pi}^N) - R_2(\pi^N) \mid S_1 \geq t_2^N\ \right] \quad = \quad t_2^N \tag{16}$$

Now,

$$\begin{aligned}
& E\left[\ R_2(\pi^N) \mid S_1 < t_2^N\ \right] \\
= \ & P\left(S_2 \leq t_3^N - t_2^N \mid S_1 \leq t_2^N\right) \times \\
& \quad E\left[R_2(\pi^N) \mid S_2 \leq t_3^N - t_2^N, S_1 \leq t_2^N\right] + \\
& P\left(S_2 > t_3^N - t_2^N \mid S_1 \leq t_2^N\right) \times \\
& \quad E\left[R_2(\pi^N) \mid S_2 > t_3^N - t_2^N, S_1 \leq t_2^N\right] \\
\geq \ & P\left(S_2 > t_3^N - t_2^N \mid S_1 \leq t_2^N\right) \times \\
& \quad E\left[R_2(\pi^N) \mid S_2 > t_3^N - t_2^N, S_1 \leq t_2^N\right] \\
= \ & P\left(S_2 > t_3^N - t_2^N\ \right) \times \\
& \quad E\left[R_2(\pi^N) \mid S_2 > t_3^N - t_2^N, S_1 \leq t_2^N\right]
\end{aligned} \tag{17}$$

If we condition on $S_2 > t_3^N - t_2^N$ in addition to $S_1 < t_2^N$, the response time of e_2 is $t_3^N - t_2^N$ plus the time from t_3^N until the main system becomes empty because e_2 has the lowest preemptive priority. We now define such conditional expected value of the queue depletion time. (Recall that X_{π^N} is left–continuous.)

$$\begin{aligned}
d_N \quad \equiv \quad & E[\inf\{t \geq t_3^N | X_{\pi^N}(t) = 0\} \mid \\
& \qquad X_{\pi^N}(t_3^N) = 1 \quad] \quad - t_3^N
\end{aligned}$$

To continue from (17), we have

$$\begin{aligned}
& E\left[R_2(\pi^N) \mid S_1 < t_2^N\right] \\
\geq \ & P\left(S_2 > t_3^N - t_2^N\right)\ (d_N + t_3^N - t_2^N) \\
\geq \ & P\left(S_2 > t_3^N - t_2^N\right)\ d_N
\end{aligned} \tag{18}$$

Now we consider the response time of e_2 for the modified schedule $\bar{\pi}^N$ under the condition $S_1 < t_2^N$. Recall that e_2 enters the main system at time 0 together with e_1 in this schedule. By a reasoning similar to the case of the optimal schedule, we have

$$\begin{aligned}
& E\left[R_2(\bar{\pi}^N) \mid S_1 < t_2^N\right] \\
= \ & P\left(S_2 \leq t_3^N - S_1 \mid S_1 \leq t_2^N\right) \times \\
& \quad E\left(S_1 + S_2 \mid S_2 \leq t_3^N - S_1, S_1 \leq t_2^N\right) + \\
& P\left(S_2 > t_3^N - S_1 \mid S_1 \leq t_2^N\right)(t_3^N + d_N) \\
\leq \ & P\left(S_2 \leq t_3^N - S_1 \mid S_1 \leq t_2^N\right) t_3^N + \\
& \qquad P\left(S_2 > t_3^N - S_1 \mid S_1 \leq t_2^N\right)(t_3^N + d_N) \\
= \ & t_3^N \ + \ P\left(S_2 > t_3^N - S_1 \mid S_1 \leq t_2^N\right)\ d_N
\end{aligned} \tag{19}$$

From (18) and (19), we have

$$\begin{aligned}
& E\left[R_2(\bar{\pi}^N) - R_2(\pi^N) \mid S_1 \leq t_2^N\ \right] \\
\leq \ & t_3^N \ + \ P\left(S_2 > t_3^N - S_1 \mid S_1 \leq t_2^N\right)\ d_N \ \ - \\
& \qquad P\left(S_2 > t_3^N - t_2^N\right)\ d_N
\end{aligned} \tag{20}$$

Combining (16) and (20), we have

$$\begin{aligned}
& E\left[R_2(\bar{\pi}^N) - R_2(\pi^N)\right] \\
\leq \ & P\left(S_1 > t_2^N\right)\ t_2^N \ + \ P\left(S_1 \leq t_2^N\right)\ t_3^N \ + \\
& \quad P\left(S_2 > t_3^N - S_1, S_1 \leq t_2^N\right)\ d_N \ - \\
& \quad\ P\left(S_1 \leq t_2^N\right)\ P\left(S_2 > t_3^N - t_2^N\right)\ d_N \\
\leq \ & t_2^N + t_3^N + \ d_N\ t_2^N\ \exp(-t_3^N) - \\
& d_N\ \{\exp(t_2^N) - 1\}\ \exp(-t_3^N) \\
\leq \ & \ln C + t^* + d_N\{t_2^N + 1 - \exp(t_2^N)\}\ \exp(-t_3^N) \\
& \text{(from Lemma 2)}
\end{aligned}$$

Therefore, the change of cost as a result of modification is

$$\begin{aligned}
& f_a(\bar{\pi}^N) - f_a(\pi^N) + f_m(\bar{\pi}^N) - f_m(\pi^N) \\
\leq \ & C \ln C + Ct^* \ + \\
& \ Cd_N\{t_2^N + 1 - \exp(t_2^N)\}\exp(-t_3^N)
\end{aligned} \tag{21}$$

For the last term, we have

$$\begin{aligned}
& d_N \\
\geq \ & P\left[X_{\pi^N}(t) > 0, \forall t \in (t_3^N, \frac{N}{2C}) | X_{\pi^N}(t_3^N) = 1\right] \times \\
& \quad \left(\frac{N}{2C} - t_3^N\right)
\end{aligned}$$

Invoking Lemma 1, with $q_N = \frac{1}{2C}N$, we can assure that

$$P\left[\ \ X_{\pi^N}(t) > 0, \ \forall t \in (t_3^N, \frac{N}{2C}) \ \ | \ X_{\pi^N}(t_3^N) = 1 \ \ \right]$$

is bounded below by a positive number. Therefore, there exists some $\delta > 0$ such that $d_N \geq \delta N$ for sufficiently large N. From Lemma 2, t_3^N is bounded above, so $d_N \exp(-t_3^N)$ grows unbounded with N. Suppose that $\{t_2^{N_k} | k = 1, 2, \cdots\}$ is bounded below by a positive number. Then, factor $\{t_2^{N_k} + 1 - \exp(t_2^{N_k})\}$ in (21) is bounded above by a negative number. Thus, the last term of expression (21) blows to $-\infty$ as N_k grows. Therefore, the change of cost becomes negative for a large N_k, contradicting optimality of π^{N_k}. Therefore, $\lim_{N\to\infty} t_2^N = 0$. **Q.E.D.**

To provide some idea on the convergence rate of t_2^N, Fig. 1 shows how t_2^N varies as N increases for the cases

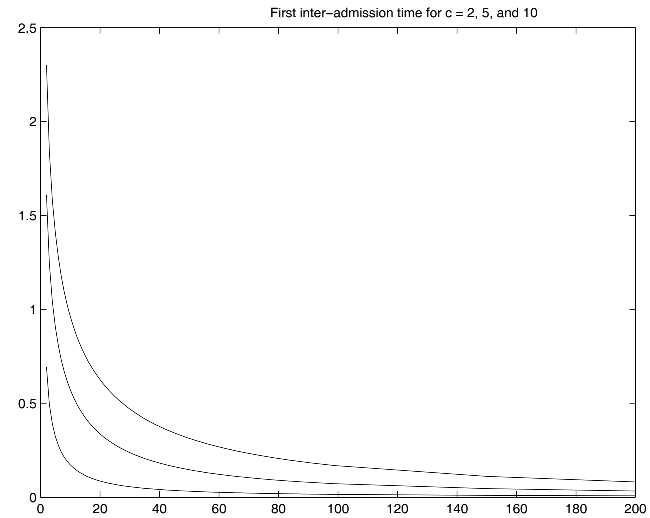

Figure 1: First inter-admission time of optimal schedule, t_2^N vs N. The top curve is for $C = 10$; the bottom curve is for $C = 2$.

$C = 2, 5, 10$. The top curve is for the case $C = 10$, and the bottom curve for $C = 2$. The curves all show that t_2^N decreases monotonically with N. These curves also show that for a range of small N's the first inter-admission time decreases rapidly. (Note that for the simple case $N = 2$, we can easily derive $t_2^2 = \ln C$.) For a range of large N's, this decreasing rate becomes fairly slow.

3 Numerical Results

This section presents numerically computed optimal schedules. We need to note that for a large N the numerical computation for finding an optimal schedule is prohibitively complex. In this section, we present numerical results for manageable values of N. We found optimal schedules for different values of C and N. Fig. 1 plots the first inter-admission times in optimal schedules for different values of N, the number of customers. Figure 2 shows the numerically obtained optimal schedules for the cases of $C = 2, 5, 10$ and $N = 20, 100, 200$. In Fig. 2, we notice that in each optimal schedule the inter-admission times are monotonically non-decreasing. In other words, for the customers admitted later, the inter-admission times are longer. This property was observed in all our numerical results. Therefore, this property is conjectured. Letting $(N-1)$ dimensional vector $(x_2^*, x_3^*, \cdots, x_N^*)$ denote the inter-admission times in an optimal schedule of admitting N customers, we conjecture: $x_i^* \leq x_{i+1}^*$ for $i = 2, 3, \cdots, N-1$. Also, the shape of the curves in Fig. 2 is interesting. In the sequence of admissions according to each optimal schedule, the inter-admission times in the early admissions are rapidly increasing. Then, they reach a plateau, and finally the last several inter-admission times again show a rapidly increasing trend. The numerical results and the theorems indicate the following nature of optimal scheduling for a large N: At the beginning stage of the schedule, customers are admitted rapidly. For the most part of the scheduling period in the middle, the admission is more or less periodic. In the late stage of the schedule, the inter-admission times are long. Although the inter-admission times vary like this, the overall average inter-admission time is close to the average service time.

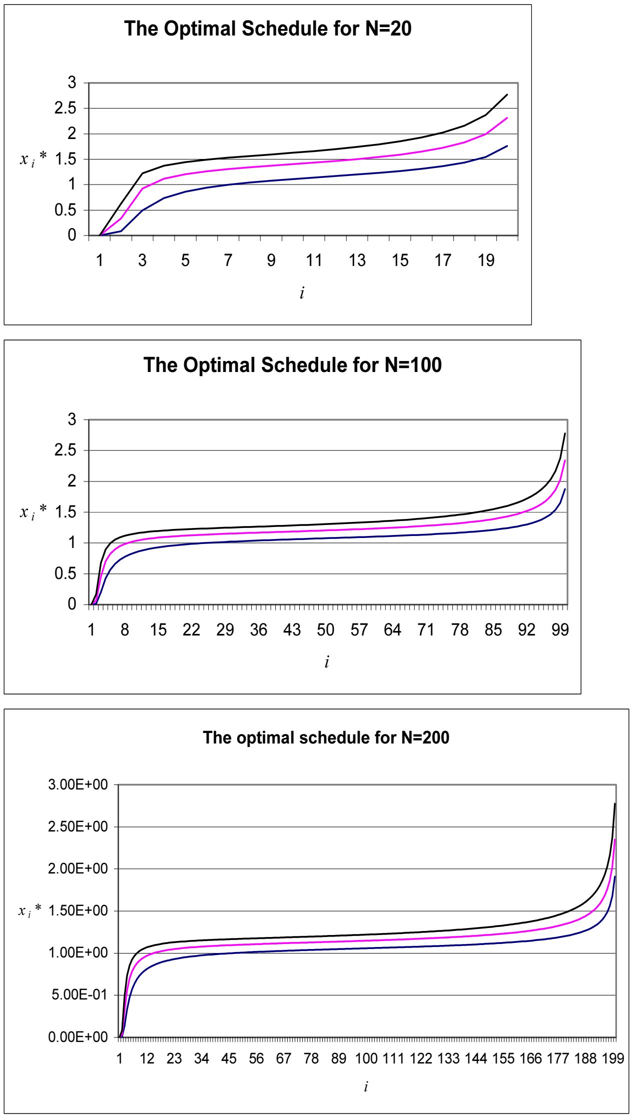

Figure 2: Optimal schedules for cases $C = 2, 5, 10$. In each box the top curve is for $C = 10$; the bottom curve is for $C = 2$.

4 Discussion

The present paper and [3] discussed optimal offline schedules of admitting finite numbers of customers from an auxiliary buffer with a low holding cost to a main queueing system with a higher holding cost. The main queueing system was assumed to be initially empty. The analytic description of the optimal schedules seems difficult. The main theorems of the present paper and Part I [3] provided an interesting perspective on the optimal schedules. While the first inter-admission time in optimal schedules converges to 0 as the number of customers increases (Theorem 1), the overall rate of the admission converges to the service rate (Theorem 1 of Part I [3]). In addition, in all optimal schedules obtained in different cases of holding cost ratios and numbers of customers, the inter-admission times arranged in chronological order were monotonically increasing. This property is left as a conjecture to prove in the future.

Theorem 1 proved that $\lim_{N\to\infty} t_2^N = 0$. This proposition was partially motivated by the question of whether there exists an atomic time unit based on which one can discretize the problem of optimal scheduling and still schedule various numbers of admissions optimally. Theorem 1 indicates that there is no positive lower bound on the inter-admission times of the optimal schedules; therefore, it also indicates that there is no such an atomic time unit. For future study, it is conjectured that $\lim_{N\to\infty} t_i^N = 0$, for each fixed i. Further, more characterization of the optimal schedule can can follow in the future. Reference [3] concerns the asymptotic transmission rate averaged over the entire of schedule. The rate averaged over different time intervals (different stages of the bulk transmission) will be interesting.

Although the optimal scheduling problem discussed in this paper was motivated by data transports through communication networks, the main interest of the paper and [3] was in the queueing theoretic intuition for the new scheduling problem. Another dimension to be considered for future continuation of this study may be that of modeling the network transport. The service rate of the single–server queue can be interpreted as the "average" bandwidth provided by the network service in accordance with the service provision agreement, of which the transport mechanism is aware. (As the present paper assumes random service time of the queue, the model addresses a broad class of network services that only provide "average" bandwidth with fluctuation of available bandwidth, not a special service such as the virtual–leased–line service associated with mechanisms like Expedited Forwarding [2].) The scheduling can be interpreted as the strategy of the feedback-free transport mechanism.

There are several directions that can be taken in order to create more elaborate models. First, the simplest extension would be to relax the assumption of exponential service time. In the simple tandem–queue model of this paper, the random delay in the entire network connection (or network service in more general interpretation) was represented by the queueing delay of a single-server queue. In this setting, the statistical characteristics of the network delay is specified simply by the service time distribution of the single-server queue. Naturally, the service time distribution that fits with the delay statistics of different network connections widely varies. The analysis in the present study assumed the exponential distribution. In order to broaden the applicability of the tandem-queues model, the service-time distribution needs to be generalized. It is conjectured that the asymptotic transmission rate of optimal schedules still converges to the service rate for the general distribution of the service time. Second, empirical validation of the single–server–queue model for the network delay is necessary. In the real network service, data units may go through several switches (or routers). Empirical data containing the admission times of the data units and their network sojourn times (time from entry to exit) need to be collected. These sojourn times then need to be compared with the queueing delays resulting from the single–server queue. In this case the probability distribution of the random service time can be constructed from the data. Third, we can consider the refinement of the main system model from the single–server queue to a network of queues. This approach, however, will add the analytic complexity enormously.

A Appendix to Lemma 1

Recall

$$\begin{aligned} c_N &\equiv P\left[X_{\pi^N}(t) > 0, \forall t \in (t_3^N, q_N) | X_{\pi^N}(t_3^N) = 1\right] \\ \hat{c}_N &\equiv P\left[X_{\hat{\pi}^N}(t) > 0, \forall t \in (t_3^N, q_N) | X_{\hat{\pi}^N}(t_3^N) = 3\right] \end{aligned}$$

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 $\lim_{k\to\infty} \hat{c}_{N_k} = 0$.

Proof: For this proof we employ the first-come-first-serve discipline for both schedules π^N and $\hat{\pi}^N$. 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 $\hat{A}_3$ to be the time the service of e_3 is ready to be served in the schedule $\hat{\pi}^N$.

Note that under schedule $\hat{\pi}^N$, e_1,e_N, and e_2 must be served before e_3 is ready to be served. Also, define

$$\begin{aligned} v_N(\tau) &\equiv P[X_{\pi^N}(t) > 0, \ \forall t \in (t_3^N, q_N) \ \mid \\ &\qquad X_{\pi^N}(t_3^N) = 1, \ A_3 = \tau + t_3^N \ ] \\ \hat{v}_N(\tau) &\equiv P[X_{\hat{\pi}^N}(t) > 0, \ \forall t \in (t_3^N, q_N) \ \mid \\ &\qquad X_{\hat{\pi}^N}(t_3^N) = 3, \ \hat{A}_3 = \tau + t_3^N \ ] \end{aligned}$$

Since $\hat{\pi}^N$ 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,

$$\begin{aligned} c_N &= \int_0^\infty \exp(-\tau) v_N(\tau) d\tau, \quad \text{and} \\ \hat{c}_N &= \frac{1}{2!} \int_0^\infty \tau^2 \exp(-\tau) \hat{v}_N(\tau) d\tau \\ &\leq \frac{1}{2!} \int_0^\infty \tau^2 \exp(-\tau) v_N(\tau) d\tau \end{aligned}$$

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,

$$\begin{aligned} \hat{c}_{N_k} \ &\leq \ \frac{1}{2!} \int_0^{T_k} \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau + \\ &\qquad \frac{1}{2!} \int_{T_k}^\infty \tau^2 \exp(-\tau) v_{N_k}(\tau) d\tau \end{aligned}$$

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

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

Because $\lim_{k \to \infty} T_k = \infty$, the second term converges to 0 as k increases. Therefore, $\hat{c}_{N_k} \to 0$ as $k \to \infty$. **Q.E.D.**

References

[1] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, Englewood Cliffs, NJ, second edition, 1992.

[2] V. Jacobson, K. Nichols, and K. Poduri. *An expedited forwarding PHB.* Internet Society, June 1999. RFC 2598.

[3] D. C. Lee. Optimally scheduling admission of n customers from low–price buffer to high-price queue, part I: Asymptotic rate. *Proc. 6th WSEAS International Conference on Telecommunications and Informatics*, May 2004, Cancun, Mexico.