Optimally Scheduling Admission of N Customers from Low–Price Buffer to High-Price Queue, Part I: Asymptotic Rate

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

Abstract

This paper introduces 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. This paper formulates a static (off-line) optimization of the admission schedule for a finite number of customers, assuming that the single–server queue is initially empty. It is proven that the overall admission rate under the optimal schedule converges to the service rate of the main system as the number of customers increases. Part II (the sequel) of this paper will present numerical solutions and prove more asymptotic properties of the sequence of optimal schedules indexed by the number of customers.

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

1 Introduction

This paper introduces a problem of optimally scheduling the transfer of N customers from one queueing facility with a lower holding cost to another queueing facility with a higher holding cost in the tandem queuing system. The system to be studied is depicted in Figure 1. The system basically comprises of two queues in tandem; we refer to the first queue as the controller and the second queue as the main system. The controller is an auxiliary buffer that controls the admission of customers to the main system. The main system is a single–server queue. Initially, the controller has a finite number of customers, and the main system is empty. In the main system, the service time of each customer has the exponential distribution with expected value 1^1, and service times of customers are statistically independent. Regarding the service discipline of the queue, we allow any work–conserving service discipline that does not discriminate customers on the basis of their remaining service times [7, p113]. There are two important additional components of our model:

- The controller cannot observe the state of the main system (single–server queue).
- The controller's queue is "less expensive." In particular, we assume that the customer indexed by i incurs a cost $\tilde{t}_i + CE(R_i)$, where $\tilde{t}_i$ is the time that the customer spends in the controller's queue, R_i is the time that it spends in the main queueing system, and C is a constant larger than 1. ($E(R_i)$ denotes the expected value of the random variable R_i.)

These assumptions are primarily motivated for macroscopically modeling various types of data transports through communications networks. The customers in the queueing model can be considered as data units (packets, messages, files, etc. [2]). The controller represents the premise of the data source, and the main system represents the network service delivering data to the final destination. Thus, for example, the random delay of the protocol data units (PDUs) in the network connection is modeled by the queueing delay in the main system. (We do not specify the layer of the PDU. We can view the transfer of data from source

[1] The results in this paper are not limited by the assumption that the expected service time is 1. For example, if the average service time is $1/\mu$, then the right-hand side of the inequality of Lemma 1 becomes $1/\mu$ instead of 1, the right-hand side of the equality in Theorem 1 becomes μ, etc. This assumption can be viewed as a convenient choice of time unit, which makes the presentation simple.

premise to the network domain at different levels of data granularity.) In this example, the randomness of the PDU's delay in a network from entry to exit is due to a number of mechanisms. One is the statistical multiplexing of data traffic with that of other connections at the intermediate switching nodes constituting the network connection. If the network contains a multiaccess link, the signal interference and collision resolution mechanism also contributes to the randomness [12]. Such randomness is specified by the randomness of the service time in the single-server queue in the system model depicted in Fig. 1. (In this example, we need to note that this random service time does not exclusively represent the transmission time of the PDU in switches. Rather, the random queuing delay in the main system resulting from the random service time represents the random end–to–end latency of the PDU in the network caused by aforementioned mechanisms. The service rate of the single–server queue can be also viewed as the average bandwidth provided by the network service.)

There are initially a fixed number, N, of customers in the controller, and the main system initially is empty. The controller has the responsibility of scheduling the admission (transmission) of these N customers into the main system. This paper concerns the admission schedule of these N customers. The goal of the scheduling is to minimize the weighted sum of the total queueing delay in the controller and the total expected queueing delay in the main system. In that weighted sum, the expected delays in the main system is given more weight. This uneven weight penalizes the queueing delay in the main system more than the delay in the controller. This captures the essential idea that the congestion in the network is more harmful because of the wasted network resources resulting buffer overflows and retransmissions. While network level flow (congestion) control (e.g., the controller's actions in our model) cannot reduce the total delay experienced by a typical data unit, it is supposed to shift delay from a network (e.g. the main system in our model) to the buffer of a source (e.g. the controller in our model) in order to avoid wasteful congestion in the network [2].

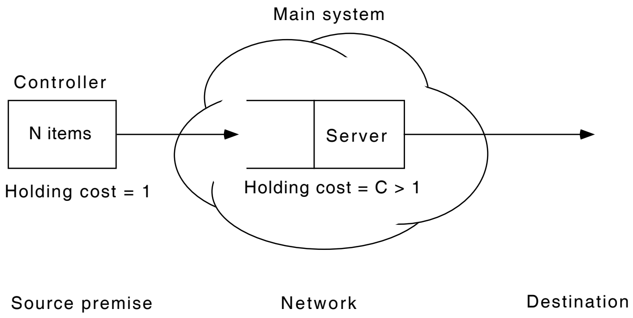

Figure 1: A system consisting of a controller and a single–server queue (main system)

This model addresses the admission (transmission) schedule of the finite amount of data. 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 information be fed into when the stream of bulk arrivals has extremely low intensity? The problem is motivated by data collection and dissemination systems that must deal with rare but critical moments at which the data volume abruptly increases. For example, in a network constituting a surveillance system, sensors are connected to the command and control center. In such a network, when an alarming situation arises, suddenly a large volume of data needs to be delivered to the command and control center. Finite number of PDUs and the assumption of initially empty main system are fit to model such a situation. While the tandem-queue model seems be an oversimplified view of the network connection, it has the advantage of providing simple intuition that can be applied to a wide variety of network topologies. All the uncertainties in the delays experienced by data units within the network are characterized by the randomness of the service times.

Note that this paper concerns deterministic off-line scheduling of admissions. This idea of off-line scheduling problem follows from the assumption that the controller cannot observe the state of the main system's queue, and that it has to schedule the admission times of all the customers based on the probability distribution of the service time. Communication networks are often characterized by service stations that act individually, each processing only a local knowledge of its immediate environment. Even when information can be exchanged among stations, there are propagation delays and processing delays that render such information partially obsolete [3]. This effect is especially evident in high–speed, wide–area 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 such as the rate control schemes [2]. Even at low–speed transmission, the need to schedule/control the flow of data with imperfect or no state information arises as we extend the network deployment to the space (e.g., interplanetary internet [1, 4]) due to the extremely long propagation delays.

Although long interpretation of the tandem-queue model has just been presented for motivation, the primary interest of the present paper is the queue scheduling problem itself and general intuitions obtained from its study. In order to apply the results to specific networking situations, more detailed refinement of the model would be necessary. In section 2 the mathematical formulation of the deterministic scheduling problem will be presented. In section 3, the related problems found in the literature and their relation to the present paper's problem are explored. As the closed-form solution is not in sight, the present paper focuses on the asymptotic admission rate of optimal schedules. Section 4 will show that the average admission rate under the optimal schedule approaches the service rate of the queue server as the number of customers N increases. In the sequel [8] to the present paper, numerical evaluation of optimal schedules and the edge effect due to the finite number of customers will be shown. In particular, it will be proven that the first inter-admission time converges to 0, as N increases. It is interesting to see the edge effects in Part II and the convergence of the average rates in the present paper.

2 Formulation

In the system described in Figure 1, there are initially (at time 0 by convention) a fixed number, N, of customers in the controller, and the main system is initially empty. The controller has the responsibility of scheduling the admission of these N customers into the main system. This paper concerns the admission schedule of these N customers. The controller has the knowledge that the main system is empty initially but cannot observe the main system's queue status thereafter (no feedback information). The scheduling decisions to be made by the controller can be represented by a finite sequence, $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$, of N nonnegative real numbers, where each number represents the time when a customer is admitted to the main system. The goal of scheduling is to minimize a cost function. We now define the cost function. For schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$, denote by $\tilde{R}_i, i = 1, 2, \cdots, N$ the flow time (synonymously, response time) of the customer indexed by i in the main system. (The flow time is defined as the time from when the customer is admitted into the main queueing system until the customer exits; it is service time plus waiting time, if any.) Then, the end-to-end delay of the customer indexed by i is $\tilde{t}_i^N + \tilde{R}_i$. We include the sum of expected end-to-end queueing delays, $\sum_{i=1}^N \left[\tilde{t}_i^N + E(\tilde{R}_i) \right]$, as an additive part of the cost function in order to reflect the undesirable nature of long end-to-end delay. The additional penalty for congestion in the main system is represented by term $\sum_{i=1}^N C^* E(\tilde{R}_i)$ with $C^* > 0$. Thus, the cost function is defined as $\sum_{i=1}^N \left[\tilde{t}_i^N + E(\tilde{R}_i) \right] + \sum_{i=1}^N C^* E(\tilde{R}_i)$. Using notation $C = 1 + C^*$, we can rewrite the cost function as

$$\sum_{i=1}^N \tilde{t}_i^N + \sum_{i=1}^N C E(\tilde{R}_i) \tag{1}$$

Note that terms $\sum_{i=1}^N C E(\tilde{R}_i)$ is a function of the schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$, and the scheduling problem is to find the schedule that minimizes the cost.

2.1 On the Cost Function to Minimize

2.1.1 Cost as a function of $\tilde{t}_1^N, \tilde{t}_2^N, \ldots, \tilde{t}_N^N$

We first clarify that in our notation of a schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$ we do not impose monotonicity $\tilde{t}_i^N \leq \tilde{t}_{i+1}^N$; that is, the customer indexed by i can be admitted later than the customer indexed by $i + 1$. (We define $\tilde{\pi}^N$ this way in order to facilitate proving lemmas and theorems to be presented in subsequent sections.) Therefore, schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$ and any of its permutations yield an identical value of cost (1).

The cost function (1) can be also viewed as the cost incurred in the controller

$$f_a(\tilde{\pi}^N) \equiv \sum_{i=1}^N \tilde{t}_i^N \tag{2}$$

plus the cost incurred in the main system

$$f_m(\tilde{\pi}^N) \equiv \sum_{i=1}^N C E(\tilde{R}_i) \tag{3}$$

Denote by $\tilde{S}_i, i = 1, 2, \cdots, N$ the service time of the customer that is admitted at the time $\tilde{t}_i$ (the ith element in the sequence $\tilde{\pi}^N$). We assume that service times of customers are statistically independent and all have the exponential distribution with mean 1. As illustration, for a First-Come-First-Serve discipline and schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$ with property $\tilde{t}_1^N \leq \tilde{t}_2^N \leq \ldots \leq \tilde{t}_N^N$, we have

$$\begin{aligned} \tilde{R}_1 &= \tilde{S}_1 \\ \tilde{R}_i &= \max[\ 0,\ \tilde{R}_{i-1} - (\tilde{t}_i^N - \tilde{t}_{i-1}^N)\ ] + \tilde{S}_i \ \ , \ \ 2 \leq i \leq N \end{aligned}$$

Beyond this illustration, we note that the cost in the main system (3) is determined by $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$ and invariant of service disciplines, even if they may be preemptive, as long as the customer to be served at each moment is decided independently of the remaining service times of the customers. We briefly justify this statement. We denote by $X_{\tilde{\pi}^N}(t)$ the random function (or a stochastic process) that represents the number of customers in the main system at time t under schedule $\tilde{\pi}^N$. For clarity we define this function (each sample path) to be left–continuous. It is well known (e.g. as in the proof of Little's law in [6]) that

$$\sum_{i=1}^{N} E[\ \tilde{R}_i\ ] = E\left[\ \int_0^\infty\ X_{\tilde{\pi}^N}(t)dt\ \right]$$

Function $X_{\tilde{\pi}^N}(t)$ jumps up by 1 at times $\tilde{t}_i^N, i = 1, 2, \cdots, N$. While $X_{\tilde{\pi}^N}(t) > 0$, i.e., during the server's busy period, downward jumps of $X_{\tilde{\pi}^N}(t)$ follow the Poisson process [10] without regard to the service discipline, due to the memoryless property [5] of the exponentially distributed service times. Therefore, the statistical nature of stochastic process $X_{\tilde{\pi}^N}(t)$ is invariant of the service disciplines, and thus so is $\sum_{i=1}^N E[\ \tilde{R}_i\ ]$. The objective is to find, for any given N, a sequence $\tilde{\pi}^N$ that minimizes the total cost

$$f_a(\tilde{\pi}^N) + f_m(\tilde{\pi}^N) \equiv f(\tilde{\pi}^N)$$

For any fixed N, this can be viewed as a deterministic nonlinear programming problem with the variables $\tilde{t}_1^N, \ldots, \tilde{t}_N^N$ with non-negativity constraint. In formulating this optimization problem, monotonicity $\tilde{t}_1^N \le \tilde{t}_2^N \le \ldots \le \tilde{t}_N^N$ is not imposed. In words, the customer indexed by i can be admitted later than the customer indexed by $i+1$. Note that the cost associated with $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$, $f(\tilde{\pi}^N)$, is obviously identical to the cost associated with any permutation of $\tilde{\pi}^N$.

2.1.2 Cost as a convex function of inter-admission times

We can express the optimization problem through another set of variables. For any vector indicating the admission times (schedule), for the purpose of evaluating the cost we can take the permutation whose elements are monotonically increasing; i.e., for the purpose of evaluating the cost only consider a schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \ldots, \tilde{t}_N^N)$ such that $\tilde{t}_1^N \le \tilde{t}_2^N \le \ldots \le \tilde{t}_N^N$. Then, we define new variables $x_i \equiv \tilde{t}_i^N - \tilde{t}_{i-1}^N, \quad i = 2, 3, \cdots, N$. Thus, we have

$$\tilde{R}_1 \quad = \quad \tilde{S}_1 \tag{4}$$

$$\tilde{R}_i \quad = \quad \max[\ 0,\ \tilde{R}_{i-1} - x_i\ ] + \tilde{S}_i \quad , \quad 2 \le i \le N \tag{5}$$

Note that for the optimal schedule $\tilde{t}_1^N$ is obviously 0, so we can exclude schedules that does not have $\tilde{t}_1^N = 0$. Thus, The cost function can be expressed as

$$\begin{aligned} f(\tilde{\pi}^N) &= \sum_{i=2}^{N} \tilde{t}_i^N + \sum_{i=1}^{N} CE(\tilde{R}_i) \\ &= \sum_{i=2}^{N}\sum_{l=2}^{i} x_l + \sum_{i=1}^{N} CE(\tilde{R}_i) \\ &= \sum_{j=2}^{N}(N-j+1)x_j + \sum_{i=1}^{N} CE(\tilde{R}_i) \\ &\equiv g(\underline{x}) \end{aligned} \tag{6}$$

where $\underline{x} \equiv (x_2, x_3, \cdots, x_N)$. Now, we show that this cost function is convex. The convexity of the cost function ensures that a local minimum is a global minimum as well. This property is useful for numerical evaluation of the minimum, which will be presented in the sequel [?] to the present paper. In the cost function (6), part $\sum_{j=2}^N (N-j+1)x_j$ is obviously convex of $(x_2, x_3, \cdots, x_N)$. Also, from Eqns. (4) (5) we can prove that $\tilde{R}_i$ is a convex function of $(x_2, x_3, \cdots, x_N)$ for any realization of $\tilde{S}_1, \tilde{S}_2, \cdots, \tilde{S}_N$. (Note that the point-wise maximum of two convex functions is convex.) Therefore, $\sum_{i=1}^N E[\ \tilde{R}_i\ ]$ is again a convex function of $(x_2, x_3, \cdots, x_N)$. Therefore, the cost (6) is a convex function of $(x_2, x_3, \cdots, x_N)$.

2.2 On the Minima

Now we establish the existence of a minimum for each N. Function f_a is continuous of variables $\tilde{t}_1^N$,…,$\tilde{t}_N^N$. For each realization of random variables $S_1, S_2, \cdots, S_N$, $\tilde{R}_i$ is continuous of variables $\tilde{t}_1^N$,…,$\tilde{t}_N^N$; therefore, so is $E[\tilde{R}_i]$. Thus, the cost function is continuous of variables $\tilde{t}_1^N$,…,$\tilde{t}_N^N$. The cost function increases without bound as a variable $\tilde{t}_i^N$ increases, so for the search of the minimum we can limit our attention to a compact set of variables $\{(\tilde{t}_1^N, \ldots, \tilde{t}_N^N) | 0 \le \tilde{t}_i^N \le B\ \}$ for a sufficiently large B. Therefore, the continuous cost function on this compact set has a minimum value in the set [11, p89].

Let us recall that the value of the cost function is invariant of permutation operation of the schedule vector $\tilde{\pi}^N$. We denote by $\pi^N = (t_1^N, t_2^N, \cdots, t_N^N)$ a minimizer of the cost function that has the monotonicity property

$$t_1^N \quad \le \quad t_2^N \le \cdots \le t_N^N \tag{7}$$

Note that obviously $t_1^N = 0$. We denote by e_i the i-th customer admitted to the main system under the optimal schedule. Thus, we can refer to the N customers by $e_1, e_2, \cdots, e_N$, and their admission times under the optimal schedule are $t_1^N \leq t_2^N \leq \cdots \leq t_N^N$, respectively. The closed form solution $\pi^N = (t_1^N, t_2^N, \cdots, t_N^N)$ is deemed difficult to obtain. We will derive some asymptotic properties of optimal schedules corresponding to the increase of N. Note that for each N, the minimizer of the cost function having the aforementioned monotonicity property may not be unique. This paper observes the asymptotic property of an arbitrary sequence of optimal schedules $\{\pi^N\}$ indexed by N, which is constructed by taking an optimal schedule for each N. We denote by S_i the service time of e_i, for $i = 1, 2, \cdots, N$. We denote by $R_i(\tilde{\pi}^N)$ the response time of e_i (the customer that is admitted to the main system ith in order under the optimal schedule) in the case of a general schedule $\tilde{\pi}^N$. Thus, the response time of e_i in the case of the optimal schedule is denoted by $R_i(\pi^N)$.

3 Related Works

The problem of optimally scheduling a finite number of arrivals to a queueing facility produced interesting studies [14, 13, 9], historically independent of the present paper. A similar problem of scheduling the arrival times of a finite number of customers to a single–server queueing system is found in references [14, 9]. The scheduling problem in these papers is formulated for naval operations or manufacturing job shop scheduling, and the scheduling objective is different from the present paper; their scheduling objective is to minimize the weighted sum of the total expected queueing delay in the main system and the expected system completion time (the time at which the last customer leaves the main system queue, which is the length of the "server availability" [9]). Thus, the cost function is different from the one formulated in the present paper.

We now clarify this difference of cost functions. The objective of [14, 9] if translated into the context and notation of the present paper, would be to minimize the cost function,

$$c_s \left[\sum_{i=2}^{N} x_i + E(\tilde{R}_N) \right] + c_w \sum_{i=1}^{N} E(\tilde{R}_i)$$

where $c_s, c_w > 0$. The minimization of this cost can be equivalently stated as minimizing:

$$\begin{aligned} & f_p(x_2, x_3, \cdots, x_N) \\ \equiv \; & \left[\sum_{i=2}^{N} x_i + E(\tilde{R}_N) \right] + c_w \sum_{i=1}^{N} E(\tilde{R}_i) \end{aligned} \tag{8}$$

(Note that c_w can be any positive number; it is not limited to the case $c_w > 1$.) The part of this cost function (8), $\left[\sum_{i=2}^{N} x_i + E(\tilde{R}_N) \right]$, explicitly shows its difference from cost function to be minimized in the present paper.

In addition to the difference of the cost function, the discussions in the present paper and its sequel are quite different in that they discuss much about necessary asymptotic properties of optimal schedules. For example, an edge effect on the optimal schedule (namely, $\lim_{N\to\infty} t_2^N = 0$) will be proven in the sequel to the present paper [**?**], along with numerical evaluation of minima. Section 4 of the present paper will use elaborate mathematics to show that despite the edge effect the average admission rate under the optimal schedule approaches the service rate of the queue server as the number of customers increases.

4 Overall Customer Admission Rate

In this section we rigorously state and prove that the rate of admitting customers into the main system according to an optimal schedule converges to the service rate (the reciprocal of the expected service time) as the number of customers increases.

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

Proof: Suppose $\liminf_{N\to\infty} \frac{t_N^N}{N} < 1$; then, there exist some $\gamma > 0$ and an increasing sequence $\{N_k\}$ indexed by k such that $t_{N_k}^{N_k} < (1-\gamma)N_k$ for all k. We will argue that the cost can then be reduced by delaying the admission time of the last customer e_N, thus contradicting optimality. Let us pick some $\delta > 0$. Compare the costs associated with the following two schedules

$$\begin{aligned} \pi^{N_k} &= (\; t_1^{N_k}, \cdots, t_{N_k-1}^{N_k}, t_{N_k}^{N_k} \;) \\ \breve{\pi}^{N_k} &= (\; t_1^{N_k}, \cdots, t_{N_k-1}^{N_k}, \; t_{N_k}^{N_k} + \delta \;) \end{aligned}$$

where π^{N_k} is the optimal schedule. Obviously, the cost difference in the controller is

$$f_a(\breve{\pi}^{N_k}) - f_a(\pi^{N_k}) \;\; = \;\; \delta \tag{9}$$

Now we compare the cost difference in the main queueing system. Assume that the first–come–first–serve discipline is used in the main system. Recall that we denote by $S_1, S_2, \cdots, S_{N_k}$ the service requirements of the customers $e_1, e_2, \cdots, e_{N_k}$, respectively. Let α_k be the probability that the service of the first $N_k - 1$ customers is finished by $t^{N_k}_{N_k} + \delta$. A necessary condition for this event to happen is $\sum_{i=1}^{N_k-1} S_i \leq t^{N_k}_{N_k} + \delta$. Therefore,

$$
\begin{aligned}
\alpha_k &\leq P(\sum_{i=1}^{N_k-1} S_i \leq t^{N_k}_{N_k} + \delta) \\
&\leq P[\sum_{i=1}^{N_k-1} S_i \leq (1-\gamma)N_k + \delta] \\
&\qquad \text{(from the supposition)}
\end{aligned}
$$

Because $E[S_i] = 1$ for each i, the weak law of large numbers implies that the right hand side of the inequality approaches 0 as N_k increases; thus $\lim_{k \to \infty} \alpha_k = 0$.

To compare the costs at the main system, we use a coupling argument; we consider an identical realization of random variables $S_1, S_2, \cdots$ (an identical sample path) for the analysis of both policies. Then, regarding the response times of customers $e_1, e_2, \cdots, e_{N_k - 1}$, we have $R_i(\check{\pi}^{N_k}) = R_i(\pi^{N_k})$, $i = 1, 2, \cdots, N_k - 1$. If under the optimal schedule π^{N_k} the last customer e_{N_k} finds the main system idle upon entry, then e_{N_k} will also find the main system idle under the modified schedule $\check{\pi}^{N_k}$. In this event, we have $R_{N_k}(\pi^{N_k}) = R_{N_k}(\check{\pi}^{N_k})$. 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 we have $R_{N_k}(\check{\pi}^{N_k}) = R_{N_k}(\pi^{N_k}) - \delta$; the probability of this event is $(1 - \alpha_k)$. Finally, if e_{N_k} finds the main system busy at time $t^{N_k}_{N_k}$ and begins to be served before time $t^{N_k}_{N_k} + \delta$ under the optimal schedule π^{N_k}, we have $R_{N_k}(\check{\pi}^{N_k}) < R_{N_k}(\pi^{N_k})$. Putting these together, the difference of expected cost in the main system is

$$
\begin{aligned}
& f_m(\check{\pi}^{N_k}) - f_m(\pi^{N_k}) \\
= \; & CE\left[R_{N_k}(\check{\pi}^{N_k}) - R_{N_k}(\pi^{N_k})\ \right] \\
\leq \; & -C\delta(1 - \alpha_k) \qquad (10)
\end{aligned}
$$

Combining (9) (10), the the difference of the total cost is

$$
f_a(\check{\pi}^{N_k}) - f_a(\pi^{N_k}) + f_m(\check{\pi}^{N_k}) - f_m(\pi^{N_k}) \leq \delta - C\delta(1 - \alpha_k)
$$

Recall now that $\alpha_k \to 0$ and that $C > 1$. This implies that the cost change will be negative when k is large enough, contradicting optimality of the original schedule. **Q.E.D.**

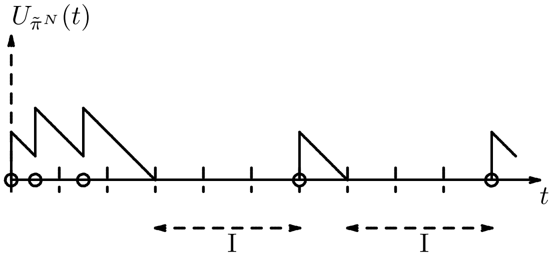

Figure 2: $U_{\tilde{\pi}^N}(t)$; symbol ◯ illustrates entry of a customer into the main system.

Let us consider a sequence of time intervals $[0, q_1), [0, q_2), [0, q_3), \cdots, [0, q_N), [0, q_{N+1}), \cdots$ such that $q_N \longrightarrow \infty$ as $N \longrightarrow \infty$, where the index of the sequence N is the number of customers in the bulk information. In Lemma 4, we will discuss admission rates in such time intervals for the case of the optimal schedule. In order to establish Lemma 4, in the next two lemmas we observe implications of relatively low admission rates in such time intervals for a general admission schedule. Denote by $A_{\tilde{\pi}^N}(q_N)$ the number of admissions in the interval $[0, q_N)$ under schedule $\tilde{\pi}^N$.

In order to describe sparsity of admissions in the schedule $\tilde{\pi}^N$, we introduce a deterministic function $U_{\tilde{\pi}^N}(t)$. For each schedule of admissions we can imagine a fictitious case where the service time of each customer in the main system is deterministically 1. By $U_{\tilde{\pi}^N}(t)$, we denote the unfinished work at the main system at each time t under particular schedule $\tilde{\pi}^N$ in the fictitious case of the deterministic service time. Thus, $U_{\tilde{\pi}^N}(t)$ 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. (See Figure 2.) We define

$$
m(q_N) \equiv \lceil q_N^{1/4} \rceil \equiv \min\{m |\ m \text{ is integer, } m \geq q_N^{1/4}\ \}
$$

$$
d(q_N) \equiv \frac{q_N}{m(q_N)} \leq q_N^{3/4}
$$

We now split the interval $[0, q_N)$ into $m(q_N)$ intervals of equal length and let L_i be the ith such interval;

$$
L_i = [\ (i-1)d(q_N)\ ,\ id(q_N)\ ]
$$

Denote $I_i \equiv \{t \in L_i \mid U_{\tilde{\pi}^N}(t) = 0\}$. Also, we denote the measure of set I_i by $\mu(I_i)$.

Lemma 2 *For sufficiently large N, if $A_{\tilde{\pi}^N}(q_N) < (1 - \epsilon)q_N$, then there exists some integer i^* such that*

1) measure $\mu(I_{i^*})$ is at least $\epsilon q_N^{3/4}/2$, and
2) $i^* \le q_N^{1/4} - \lceil C \rceil - 2$

Proof: If $A_{\tilde{\pi}^N}(q_N) < (1-\epsilon)q_N$, it is clear that the measure of the set $I \equiv \{t \in [0, q_N) \mid U_{\tilde{\pi}^N}(t) = 0\}$ is at least ϵq_N; i.e.

$$\mu(I) \quad \ge \quad \epsilon q_N \tag{11}$$

Recalling that each interval L_i has length $d(q_N)$, we establish

$$\begin{aligned} & \mu(I_{m(q_N)-\lceil C\rceil-2}) + \mu(I_{m(q_N)-\lceil C\rceil-1}) \\ & \quad +\mu(I_{m(q_N)-\lceil C\rceil}) + \cdots \mu(I_{m(q_N)}) \\ \le \; & (\lceil C\rceil + 3)d(q_N) \\ \le \; & (C+4)q_N^{3/4} \end{aligned}$$

Suppose there is no such i^* as is described above. Then, we would have

$$\begin{aligned} & \mu(I_1) + \mu(I_2) + \cdots + \mu(I_{m(q_N)-1-\lceil C\rceil-2}) \\ \le \; & \frac{\epsilon q_N^{3/4}}{2}(m(q_N) - 1 - \lceil C\rceil - 2) \\ \le \; & \frac{\epsilon q_N^{3/4}}{2}(q_N^{1/4} - \lceil C\rceil - 2) \le \frac{\epsilon q_N}{2} \end{aligned}$$

Therefore, we would have $\mu(I) = \sum_{i=1}^{m(q_N)} \mu(I_i) \le \frac{\epsilon q_N}{2} + (C+4)q_N^{3/4}$. However, for sufficiently large N, $\epsilon q_N/2 + (C+4)q_N^{3/4} < \epsilon q_N$, which contradicts inequality (11). **Q.E.D.**

The next lemma is regarding the measure of the server's idle time in the interval L_{i^*} mentioned in Lemma 2, $\mu(\{t \in L_{i^*} \mid X_{\tilde{\pi}^N}(t) = 0\})$. Note that this measure is a random variable due to the randomness of the customers' service times. Also, note that due to the memoryless property of the exponential probability density function, the distribution of random variable $\mu(\{t \in L_{i^*} \mid X_{\tilde{\pi}^N}(t) = 0\})$ is invariant to the service discipline as long as the service discipline is work-conserving.

Lemma 3 *For sufficiently large N, if $A_{\tilde{\pi}^N}(q_N) < (1-\epsilon)q_N$, then $P[\ \mu(\{t \in L_{i^*} \mid X_{\tilde{\pi}^N}(t) = 0\}) > \epsilon q_N^{3/4}/6\ ]$ is arbitrarily close to 1.*

Proof: Let $v_N = \inf\{\tau \in L_{i^*} \mid U_{\tilde{\pi}^N}(\tau) = 0\}$ and $z_N = \sup\{\tau \in L_{i^*} \mid U_{\tilde{\pi}^N}(\tau) = 0\}$. 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 $[v_N, z_N)$. (See Figure 3.) Clearly, we then have

$$k_N \quad \le \quad v_N \tag{12}$$

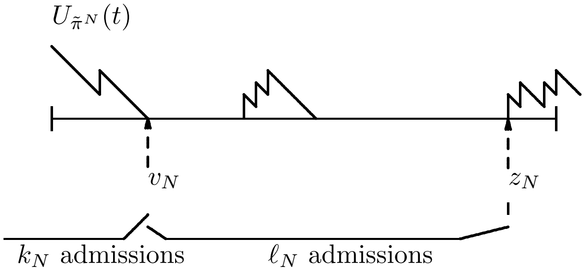

Figure 3: Interval L_{i^*} of length $q_N^{3/4}$

Clearly, the subset I_{i^*} of L_{i^*}, on which $U_{\tilde{\pi}^N}(t) = 0$, is contained in $[v_N, z_N]$, and by the definition of i^* we have $\mu(I_{i^*}) \ge \epsilon q_N^{3/4}/2$. Therefore, the subset of $[v_N, z_N]$ on which $U_{\tilde{\pi}^N}(t) > 0$ has measure at most $z_N - v_N - \epsilon q_N^{3/4}/2$. Therefore, we have

$$\ell_N \quad \le \quad z_N - v_N - \epsilon q_N^{3/4}/2. \tag{13}$$

We assume without loss of generality that the service discipline is first-come-first-serve, and $\tilde{t}_1^N \le \tilde{t}_2^N \le \cdots \le \tilde{t}_N^N$ in the schedule $\tilde{\pi}^N = (\tilde{t}_1^N, \tilde{t}_2^N, \cdots, \tilde{t}_N^N)$. Define T_i^N be the time at which the service of the ith customer ends. Also consider the sum of the service times of the customers admitted during $[v_N, z_N)$, $\sum_{i=k_N+1}^{k_N+\ell_N} \tilde{S}_i$. Consider the following two events; (a) $T_{k_N}^N \le v_N + \epsilon q_N^{3/4}/6$ and (b) $\sum_{i=k_N+1}^{k_N+\ell_N} \tilde{S}_i \le z_N - v_N - 2\epsilon q_N^{3/4}/6$. Note that these events are statistically independent because event (a) is determined by the service times of the first k_N customers, and event (b) by the service times of the next ℓ_N customers. If these two events occur, 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 $[v_N, z_N)$. Therefore, if events (a) and (b) occur, the measure of idle time in L_{i^*} is at least $\epsilon q_N^{3/4}/6$. Thus, we now only need to show that the probability that both events (a) and (b) happen is arbitrarily close to 1 for sufficiently large N. Note that

$$\begin{aligned} & E\Big(\sum_{i=k_N+1}^{k_N+\ell_N} \tilde{S}_i\ \Big) \\ = \; & \ell_N \\ \le \; & z_N - v_N - \frac{\epsilon q_N^{3/4}}{2} \qquad \text{(from inequality (13))} \end{aligned}$$

The standard deviation of $\sum_{i=k_N+1}^{k_N+\ell_N} \tilde{S}_i$ is $\sqrt{\ell_N}$, and we have $\sqrt{\ell_N} \leq \sqrt{A_{\tilde{\pi}^N}(q_N)} < \sqrt{(1-\epsilon)q_N} < q_N^{1/2}$. For event (b) not to occur, $\sum_{i=k_N+1}^{k_N+\ell_N} \tilde{S}_i$ must be at least

$$\frac{\epsilon q_N^{3/4}}{6} \frac{1}{q_N^{1/2}} = \frac{\epsilon q_N^{1/4}}{6}$$

standard deviations above its mean, and the probability of this happening is arbitrarily close to zero for sufficiently large N, by the Chebyshev's inequality. Thus, the probability of event (b) is arbitrarily close to 1 for sufficiently large N. 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 no more than v_N [cf. inequality (12)], and it easily follows (as in the case of event (b)) that the probability of event (a) also converges to 1 as N increases. Since these two events are statistically independent, the probability that both events happen converges to 1 as N increases. **Q.E.D.**

The following lemma concerns the average admission rate in interval $[0, q_N)$ for the optimal schedule. Recall that t_N^N is the time of the last admission under the optimal schedule.

Lemma 4 *If $q_N \leq t_N^N$ for each N, then*

$$\liminf_{N\to\infty} \frac{A_{\pi^N}(q_N)}{q_N} \geq 1$$

Proof: The basic method of the proof is that if the number of admissions in $[0, q_N)$ were too small, then the main system would have long idle periods that can be exploited to reduce the costs, thus contradicting optimality.

Suppose now that $\liminf_{N\to\infty} A_{\pi^N}(q_N)/q_N < 1$. Then, there exists some $\epsilon > 0$ and an increasing sequence $\{N_k\}$ such that $A_{\pi^N}(q_{N_k}) < (1-\epsilon)q_{N_k}$. Consider a sufficiently large integer N for which $A_{\pi^N}(q_N) < (1-\epsilon)q_N$. Then, according to Lemma 2, there exists i^* such that for schedule π^N we have $\mu(I_{i^*}) \geq \epsilon q_N^{3/4}/2$ and $i^* \leq q_N^{1/4} - \lceil C \rceil - 2$. Denote $l^* = A_{\pi^N}((i^* + \lceil C \rceil) q_N^{3/4}) + 1$. Then, e_{l^*} is 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 that L_{i^*} is not one of the last $\lceil C \rceil + 2$ intervals, and also because $q_N \leq t_N^N$. We will refer to this customer e_{l^*} as "special customer". We now consider the following modification of the assumed optimal schedule. Under the modified schedule, which we denote by $\grave{\pi}^N$, this special customer is to be admitted at the beginning of the interval L_{i^*}, and the admission times of the other customers are identical to the optimal schedule π^N. (Figure 4) We now compare the costs associated with the optimal schedule and the modified schedule. Regarding the cost incurred in the controller, under the modified schedule only one customer, namely e_{l^*}, has admission time different from under the optimal schedule. In the modified schedule, $\grave{\pi}^N$, customer e_{l^*} is admitted earlier, and the difference of admission times is at least $\lceil C \rceil + 1$ intervals of length $d(q_N)$. Therefore,

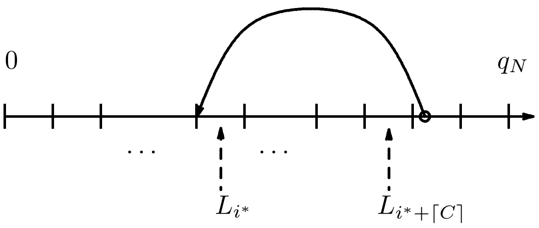

Figure 4: Modification from the optimal schedule – customer e_{l^*} is admitted earlier.

$$f_a(\grave{\pi}^N) - f_a(\pi^N) \leq -(\lceil C \rceil + 1) d(q_N)$$

Denote by $h_N \geq 0$ the difference between the admission time of e_{l^*} under the optimal schedule π^N and $(i^* + \lceil C \rceil) d(q_N)$. Then, we have

$$f_a(\grave{\pi}^N) - f_a(\pi^N) = -[\ \ (\lceil C \rceil + 1) d(q_N) + h_N\ \ ]$$

Let us now consider the cost change in the main system. To do the comparison, we use a coupling argument. We consider the sample paths, $X_{\pi^N}(t)$ and $X_{\grave{\pi}^N}(t)$, resulting from two different schedules π^N and $\grave{\pi}^N$ while keeping the service time of each customer the same. Note that 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, customer e_{l^*} has the lowest service priority, and that other customers have preemptive priority over this special customer. Thus, e_{l^*} cannot be served at any time another customer is in the main system, and the service of customers other than e_{l^*} is not impeded in any way by the service of e_{l^*}. Due to the priority discipline we have assumed, it is clear that all customers except for e_{l^*} have the same response time at the main system. We can therefore compare only the costs incurred by e_{l^*}. Denote by $R_{l^*}(\pi^N)$ and $R_{l^*}(\grave{\pi}^N)$ the response time of the customer e_{l^*} for the optimal schedule π^N and the modified schedule $\grave{\pi}^N$, respectively. Consider the event that the service of customer e_{l^*} is completed during the interval L_{i^*}

under the modified schedule. We denote this event by $\mathcal{G}_N$. Obviously, we have $E[R_{l^*}(\grave{\pi}^N) \mid \mathcal{G}_N] \leq d(q_N)$. Even in the event,$\mathcal{G}_N^c$, that e_{l^*} is not finished during the interval L_i^* under the modified schedule $\grave{\pi}^N$, the difference of the response times $R_{l^*}(\grave{\pi}^N) - R_{l^*}(\pi^N)$ cannot be more than the difference in admission times of the customer e_{l^*}; therefore, we have

$$E[\ R_{l^*}(\grave{\pi}^N) - R_{l^*}(\pi^N) \mid \mathcal{G}_N^c\ ] \leq \quad (\lceil C \rceil + 1)d(q_N) + h_N$$

Therefore,

$$\begin{array}{rl} & f_m(\grave{\pi}^N) - f_m(\pi^N) \\ = & CP(\mathcal{G}_N)E[\ R_{l^*}(\grave{\pi}^N) - R_{l^*}(\pi^N) \mid \mathcal{G}_N\ ] + \\ & \ CP(\mathcal{G}_N^c)E[\ R_{l^*}(\grave{\pi}^N) - R_{l^*}(\pi^N) \mid \mathcal{G}_N^c\ ] \\ \leq & CP(\mathcal{G}_N)d(q_N) + \\ & \ CP(\mathcal{G}_N^c)[\ (\lceil C \rceil + 1)d(q_N) + h_N\ ] \end{array}$$

To summarize,

$$\begin{array}{rl} & f_a(\grave{\pi}^N) - f_a(\pi^N) + f_m(\grave{\pi}^N) - f_m(\pi^N) \qquad (14) \\ \leq & -[\ (\lceil C \rceil + 1)d(q_N) + h_N\ ] + CP(\mathcal{G}_N)d(q_N) + \\ & \ CP(\mathcal{G}_N^c)\,[(\lceil C \rceil + 1)d(q_N) + h_N\ ] \\ = & h_N\ \ [\ -1 + CP(\mathcal{G}_N^c)\ ] + \\ & d(q_N)\ \ \{\ -\lceil C \rceil - 1 + CP(\mathcal{G}_N)\} + \\ & d(q_N)CP(\mathcal{G}_N^c)(\lceil C \rceil + 1) \end{array}$$

Now,

$$\begin{array}{rcl} P(\mathcal{G}_N) & = & P[S_{l^*} \leq \mu(\{t \in L_{i^*} | X_{\pi^N}(t) = 0\})\ ] \\ & \geq & P[S_{l^*} \leq \epsilon q_N^{3/4}/6\ ] \cdot \\ & & P[\ \mu(\{t \in L_{i^*} \mid X_{\tilde{\pi}^N}(t) = 0\ \}) > \epsilon q_N^{3/4}/6\ ] \end{array}$$

For sufficiently large N, $P[S_{l^*} \leq \epsilon q_N^{3/4}/6\ ] = 1 - \exp(-\epsilon q_N^{3/4}/6)$ is arbitrarily close to 1, considering that q_N increases to ∞. Also, Lemma 3 states that $P[\ \mu(\{t \in L_{i^*} \mid X_{\tilde{\pi}^N}(t) = 0\ \}) > \epsilon q_N^{3/4}/6\ ]$ is arbitrarily close to 1 for sufficiently large N. Therefore, as N grows, $P(\mathcal{G}_N)$ converges to 1, and $P(\mathcal{G}_N^c)$ converges to 0. Therefore, the cost change associated with the modification from the optimal schedule, as expressed in (14), becomes negative. This contradicts optimality of the schedule π^N. **Q.E.D.**

Theorem 1

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

Proof: Use $q_N = t_N^N$ in Lemma 4. From Lemma 1 we have $t_N^N \to \infty$ as $N \to \infty$, and by definition we have $A_{\pi^N}(t_N^N) = N - 1$. Therefore, Lemma 4 implies $\liminf_{N\to\infty} \frac{N-1}{t_N^N} \geq 1$, so

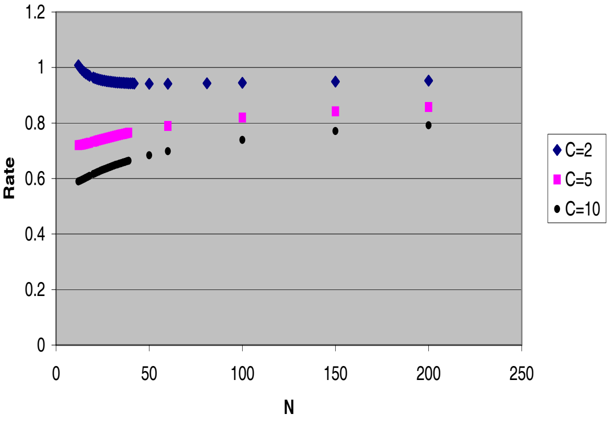

Figure 5: N/t_N^N vs. N

$$\liminf_{N\to\infty} \frac{N}{t_N^N} \quad \geq \quad 1 \qquad (15)$$

From Lemma 1 we also have $\liminf_{N\to\infty} \frac{t_N^N}{N} \geq 1$, which is equivalent to

$$\limsup_{N\to\infty} \frac{N}{t_N^N} \quad \leq \quad 1 \qquad (16)$$

Inequalities (15) and (16) imply

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

Q.E.D.

A natural question raised after the convergence result is how fast the convergence takes place. In order to gain some idea on the convergence speed, numerical methods were used to obtain minima for different numbers of customers (N's) and different values of C's. The optimal admission rates obtained from these numerical results are presented in Fig. 5. As the mathematical optimization problem is over positive orthant of the $(N-1)$-dimensional real vector space, the numerical computation for large N becomes prohibitive. However, the numerical evaluation for relatively small values of N provides good insights.

References

[1] IPN technical information. http://www.ipnsig.org/techinfo.htm.

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

[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] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, E. Travis, and H. Weiss. Interplanetary internet (IPN): Architectural definition. http://www.ietf.org/internet-drafts/draft-irtf-ipnrg-arch-00.txt, May 2001.

[5] A. W. Drake. *Fundamentals of Applied Probability Theory.* McGraw–Hill, New York, 1967.

[6] L. Kleinrock. *Queueing Systems*, volume 1. John Wiley & Sons, New York, 1975.

[7] L. Kleinrock. *Queueing Systems*, volume 2. John Wiley & Sons, New York, 1976.

[8] D. C. Lee. Optimally scheduling admission of N customers from low–price buffer to high-price queue, part II: Edge effect. *Proc. 6th WSEAS International Conference on Telecommunications and Informatics*, May 2004, Cancun, Mexico.

[9] C. D. Pegden and M. Rosenshine. Scheduling arrivals to queues. *Computers & Operations Research*, 17(4):343–348, 1990.

[10] S. M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, 1983.

[11] W. Rudin. *Principles of Mathematical Analysis.* McGraw–Hill, New York, third edition, 1976.

[12] İ. E. Telatar and R. G. Gallager. Combining queueing theory with information theory for multiaccess. *IEEE Journal on Selected Areas in Communications*, 13(6):963–969, August 1995.

[13] P. Wang. Static and dynamic scheduling of customer arrivals to a single–server system. *Naval Research Logistics*, 40:345–360, 1993.

[14] P. P. Wang. Optimally scheduling N customer arrival times for a single–server system. *Computers & Operations Research*, 24(8):703–716, 1997.