QUEUEING PERFORMANCE WITH IMPATIENT CUSTOMERS

Z-X. Zhao, S.S. Panwar, D. Towsley

COINS Technical Report 91-18
February 1991

QUEUEING PERFORMANCE WITH IMPATIENT CUSTOMERS*

ZHENG-XUE ZHAO † SHIVENDRA S. PANWAR †
DON TOWSLEY ‡

†*Polytechnic University, Brooklyn, New York 11201*
‡*University of Massachusetts, Amherst, Massachusetts 01003*

Abstract

We consider the problem of scheduling impatient customers in a non-preemptive G/GI/1 queue. Every customer has a random deadline to the beginning of its service. Given the distribution of the customer deadlines (rather than their exact values), a scheduling policy decides the customer service order and also which customer(s) to reject, since those whose deadlines have expired do not leave the queue automatically. Our objective is to find an optimal policy which maximizes the number of customers served before their deadlines. We show that LIFO (last-in first-out) is an optimal service order when the deadlines are i.i.d. random variables with a concave cumulative distribution function. After analyzing the rejection strategy, we claim that there is an optimal policy in the LIFO-TO (time-out) class, as defined in the paper. For the M/GI/1 queue, we further prove that unforced idle times are not allowed under this optimal policy. We also show that the optimal LIFO-TO policy assigns a fixed critical time (i.e., its maximum waiting time) to every customer. When the customer waiting times are unknown, we show that the optimal policy for a M/M/1 queue becomes the LIFO-PO (push-out) policy, with a fixed buffer size used as a rejection threshold. Among other applications, this may be applied in determining scheduling and buffer management policies for the time-critical cells in an ATM (Asynchronous Transfer Mode) network.

1 Introduction

Some queueing performance is significantly affected by the behavior of impatient customers. These customers should be served before their respective deadlines. A customer which exceeds its deadline will either leave the queue without service or stay in the queue to get unsuccessful service. One application of this problem is the transmission of time-constrained messages over a communication channel. These messages have to reach their receivers within a certain time interval of their transmission or they are useless to the receivers and considered lost. Two possible scenarios are often encountered in this kind of queueing system[1]. The first is that the server of the queue is aware of each customer's deadline. The messages whose delay times exceed their deadlines are discarded without transmission. In the second scenario, the server is only aware of the deadline distribution of the customers. Therefore some server work is useless because of the expiration of customers' deadlines. For example, there may be a delay before a dial tone in an overloaded call processing system. If some people start dialing before a dial tone is heard, the system will not receive all the digits dialed. However, the call is still processed and an unsuccessful call results[4]. A similar case can arise in dealing with time-critical voice or video cells in an ATM (Asynchronous Transfer Mode) network.

When the customers' deadlines are available, the shortest time to extinction (STE) and the shortest time to extinction with inserted idle time (STEI) were proved to be optimal under certain conditions. These policies maximize the fraction of customers served within their respective deadlines out of an arrival stream. The results for single server queues can be found in [7], [8] and [2]. In [13], earlier results are extended to multi-server queues.

This paper is devoted to determining the optimal policies when only the deadline cumulative distribution is known. Without knowing the deadline of every specific customer, the control action is to decide, at appropriate decision instants, which customer to serve and which customer(s) to reject. The rejection is necessary since the customers whose deadlines have expired do not leave the queue automatically. Therefore a customer could be

*This work was supported in part by the National Science Foundation under grant NCR-8909719, and by the New York State Center for Advanced Technology in Telecommunications (CATT), Polytechnic University, Brooklyn, New York.

To be presented at INFOCOM '91.

either served in an order decided by a service discipline or discarded by a rejection scheme. From now on, we use the term queueing "policy" to represent the combination of the service discipline and rejection scheme in a queue. The following notation is used for some specific queueing policies in this paper:

(*i*) FIFO(or LIFO)-BL: first-in first-out (or last-in first-out) service discipline; a customer arriving to see a "full" buffer leaves immediately (blocked).

(*ii*) FIFO(or LIFO)-PO: first-in first-out (or last-in first-out) service discipline; a customer arriving to see a "full" buffer pushes out the "oldest" customer (the one with the longest waiting time) in the buffer and joins the queue.

(*iii*) FIFO(or LIFO)-TO: first-in first-out (or last-in first-out) service discipline; every arriving customer joins the buffer but will leave at a critical time T after its arrival if it is still in the buffer at that time (time-out).

More precisely, the above notation is used for queueing policy classes. Those classes consist of the queueing policies with "full" buffer size (for BL and PO schemes) or critical time (for TO scheme) varying with the state of the queue.

Increasing interest has been shown in the performance evaluation of such queueing systems[4]. Optimal service disciplines were found in [4], [6] and [10] for the M/G/1 and G/G/1 non-preemptive queues with different reward functions when no rejections are allowed. [5] and [12] discussed the optimal control problem for a non-preemptive M/M/1/k overloaded queue under FIFO-BL. It is proved that a fixed threshold type rejection decision is optimal for a BL scheme. This optimal policy maximizes the reward associated with the successful service of customers. Other work on this issue focuses on the performance evaluation of various queueing policies. The TO rejection scheme was proposed in [4] to improve the queueing performance when an appropriate constant critical time is assigned to every customer. Under the different policies, delay distributions have been compared in [3] for the M/M/1 non-preemptive queue after matching the throughput of successfully served customers. This throughput is also maximized with respect to critical time or buffer size.

This paper is organized as follows. Section 2 contains a model of the system and some general results on the structure of an optimal policy. This analysis leads to an optimal stationary policy in Section 3 when the arrival process is Poisson. Next in Section 4, we determine the optimal policy for a M/M/1 queue under a reduced information structure. Our results are summarized in Section 5.

2 Model and policy analysis

We consider a simple non-preemptive G/GI/1 queue with either a finite or an infinite buffer as our model. The i-th customer C_i arrives at the instant a_i and generates a random deadline d_i to the beginning of its service with common distribution function $F_d(.)$ on the set of positive real numbers. $\{d_i\}_{i=1}^{\infty}$ are all independent and, unless noted otherwise, $F_d(.)$ is a non-decreasing concave function. Customer service times are independent and identically distributed. The deadline of a customer may expire while waiting in the queue. If a customer with an expired deadline is served, then its service is considered unsuccessful. Our objective is to choose a queueing policy π such that the fraction of customers getting served before their deadlines can be maximized.

Let $G_\pi(S_0)$ denote the process $\{G_\pi^t(S_0), t \geq 0\}$ with initial state S_0, where $G_\pi^t(S_0)$ is the number of customers served successfully under π by time $t \geq 0$. A policy π is better than another policy π' if $G_\pi(S_0) \geq_{st} G_{\pi'}(S_0)$ for all S_0, where $\geq_{st}$ is a stochastic order relation to be defined shortly. A policy π^* is optimal if

$$G_{\pi^*}(S_0) \geq_{st} G_\pi(S_0) \qquad \forall \pi \text{ and } S_0.$$

Using the standard notation[9], we say that a random variable X is stochastically larger than another random variable Y, written $X \geq_{st} Y$, if

$$\Pr(X > c) \geq \Pr(Y > c) \qquad \text{for all } c.$$

For random vectors, $\underline{X} = (X_1, \ldots, X_k)$ is stochastically greater than $\underline{Y} = (Y_1, \ldots, Y_k)$, if for all increasing functions f

$$\mathrm{E}[f(\underline{X})] \geq \mathrm{E}[f(\underline{Y})]. \tag{1}$$

The order relations for stochastic processes can be extended from the definitions for random vectors. We say that the process $\{X(t), t \geq 0\}$ is stochastically greater than the process $\{Y(t), t \geq 0\}$, and write $X(t) \geq_{st} Y(t)$ if

$$(X(t_1), \ldots, X(t_k)) \geq_{st} (Y(t_1), \ldots, Y(t_k))$$

for all $t_1, \ldots, t_k$ and k.

Back to our problem, let $\{b_i\}_{i=1}^{\infty}$ be the sequence of positive random variables which are used to assign service times to customers according to their service order. Then $\tilde{s} = (\{a_i\}_{i=1}^{\infty}, \{b_i\}_{i=1}^{\infty})$ is referred to as an input sample. Consider an arbitrary sample path $\tilde{s}$ applied to the queue starting from S_0. Let W_i denote the waiting time of customer C_i from its arrival instant to the beginning of its service under π. W_i is infinite if C_i is rejected. Then

$$G_\pi^t(S_0, \tilde{s}) \;=\; \sum_{i=1}^{n_\pi^t(S_0,\tilde{s})} \Pr(W_i < d_i | \pi, S_0, \tilde{s})$$

$$= \sum_{i=1}^{n_\pi^t(S_0,\tilde{s})} [1 - F_d(W_i)], \qquad (2)$$

where $n_\pi^t(S_0,\tilde{s})$ is the random number of customers which depart from the buffer, either for service or as a result of a rejection, by time t. Equation (2) allows us to analyze the queueing performance by computing the sum of the successful service probabilities of all customers in an arbitrary sample path till time t. Thus we can use the same sample path to compare the performances under different queueing policies.

In the following theorem, we appropriately exchange the service order under one policy to improve its queueing performance[14]. Similar results are also given in [4], [10] and [6] for queues with no rejections.

Theorem 2.1 *For a non-preemptive G/GI/1 queue, if the customers' deadlines are independent and identically distributed and $F_d(.)$ is a concave function, there exists a policy with LIFO service discipline which is at least as good as any non-LIFO policy.*

Proof: Assume an arbitrary policy π does not serve customers from an arbitrary sample path $\tilde{s}$ in LIFO order. Let C_i, C_j, with waiting times $W_i \geq W_j$, be among the customers available for service at time t_i. C_i gets served first and C_j is either served after time period $\tau \geq 0$ or rejected at some point $t_i+\tau$ under π. We can construct a new policy π' that is identical to π except that it serves C_j first instead of C_i. The service order of customers other than C_i and C_j is the same under both policies. Then π' is as good as π before time t_i. Since the same arbitrary sample path (containing C_i, C_j) is scheduled by π and π', and customer service time is independent of all the other random factors including the customer itself, the service time of C_j under π' is the same as the service time of C_i under π. We have the following two cases:

Case 1: If C_j is rejected by π at $t_i + \tau$, π' will reject C_i. We have

$$\begin{aligned} & G_{\pi'}^t(S_0,\tilde{s}) - G_\pi^t(S_0,\tilde{s}) \\ = \; & \Pr(W_j < d_j|\pi', S_0, \tilde{s}) - \Pr(W_i < d_i|\pi, S_0, \tilde{s}) \\ = \; & F_d(W_i) - F_d(W_j), \qquad t \in [t_i, \infty). \end{aligned} \qquad (3)$$

Since $W_i \geq W_j$ and $F_d(.)$ is a non-decreasing function, equation (3) is not negative. π' is at least as good as π in this case.

Case 2: If π chooses to serve C_j at $t_i + \tau$, π' chooses to serve C_i instead at that time. Then

$$\begin{aligned} & G_{\pi'}^t(S_0,\tilde{s}) - G_\pi^t(S_0,\tilde{s}) \\ = \; & \begin{cases} \Pr(W_j < d_j|\pi', S_0, \tilde{s}) \\ \quad - \Pr(W_i < d_i|\pi, S_0, \tilde{s}), \\ \qquad\qquad t \in [t_i, t_i + \tau); \end{cases} \\ = \; & \begin{cases} \Pr(W_j < d_j|\pi', S_0, \tilde{s}) \\ \quad + \Pr(W_i + \tau < d_i|\pi', S_0, \tilde{s}) \\ \qquad - \Pr(W_i < d_i|\pi, S_0, \tilde{s}) \\ \qquad\quad - \Pr(W_j + \tau < d_j|\pi, S_0, \tilde{s}), \\ \qquad\qquad t \in [t_i + \tau, \infty) \end{cases} \\ = \; & \begin{cases} F_d(W_i) - F_d(W_j), \\ \qquad\qquad t \in [t_i, t_i + \tau); \\ F_d(W_i) + F_d(W_j + \tau) \\ \quad - F_d(W_j) - F_d(W_i + \tau), \\ \qquad\qquad t \in [t_i + \tau, \infty). \end{cases} \end{aligned} \qquad (4)$$

Since $\tau \geq 0$ and $W_j \leq W_i$, we have

$$W_j + \tau \leq W_i + \tau.$$

$F_d(.)$ is a non-decreasing concave function, so equation (4) is not negative. Therefore π' is also at least as good as π in this case as well. ∎

Theorem 2.1 shows that there always exists a LIFO policy which can perform at least as well as any non-LIFO policy by employing a work-conserving rejection scheme, although this may imply some changes in the original scheme. Therefore there is an optimal policy belong to the class of policies using LIFO service discipline.

Next we consider the rejection scheme for an optimal queueing policy. When the arrival rate to a queueing system is larger than the service rate of server, some customers may never reach the server if a rejection scheme is not applied to the system. These customers stay in the buffer forever which is equivalent to being rejected. In general, a customer waiting in the buffer for a very long time will, with a probability approaching one, have an expired deadline. Since serving this customer could result in useless server work, it is worth rejecting it and serving another customer with shorter waiting time, or even waiting for a new arrival and serving it. We are interested in finding the optimal rejection scheme under one of the following assumptions:

($A1$) The customer waiting times are available to the server.

($A2$) Only the buffer occupancy is known to the server.

The following lemma will help us determine an optimal rejection scheme under either of the above two assumptions.

Lemma 2.2 *For any deadline distribution $F_d(.)$, there exists an optimal queueing policy which does not reject*

a customer with given waiting time while another customer present in the buffer with a longer waiting time gets served later.

Proof: Consider an arbitrary input sample path $\tilde{s}$. Assume a customer C_i is rejected by an optimal policy π, while customer C_j with waiting time $W_j \geq W_i$ is in the buffer. If C_j completes its service later, we can have another policy π' under which C_j is rejected and C_i is served. π' is at least as good as π, because the successful probability is a non-increasing function of waiting time (see Case 1 in proof of Theorem 2.1). ∎

Lemma 2.2 implies that an optimal policy π^* could reject all the customers with waiting time longer than W_i whenever C_i is rejected. Since those customers will be rejected eventually, they would not affect π^* anyway. From the approach used in the proof of the above lemma, we also find that a policy can be improved by making it reject a customer with the largest waiting time instead of a customer with a shorter waiting time. BL schemes reject new arrivals when the queue is full. If we use a PO scheme instead of a BL scheme so that the altered policy pushes out the "oldest" customer instead of blocking at every rejection moment, we have the following corollary.

Corollary 2.3 *For a non-preemptive G/GI/1 queue, there exists a policy using the PO rejection scheme which is at least as good as the one using the BL scheme. This is true for any customer deadline distribution $F_d(.)$.*

As has been proved in Theorem 2.1, an optimal policy can schedule customers in LIFO order. On the other hand, a rejection leads to the rejection of all the "older" customers by Lemma 2.2. These results are derived by improving the queueing performance under an arbitrary queueing policy. With assumption $A1$, customer waiting times are known to the server. Therefore for customers with a concave deadline distribution, an optimal policy could exist in the LIFO-TO policy class. The critical time T of this LIFO-TO policy could potentially vary with the state of the queue.

3 Optimal stationary policy for the M/GI/1 queue

Now let the arrival process be Poisson with arrival rate λ and independent of the customer service times and deadlines. With the *memoryless* property of the inter-arrival times, any queueing state could be treated as an initial state expecting the next arrival with arrival rate λ. Consider the queue at a decision instant $t \geq 0$, whenever a service completion occurs or an arrival joins an empty queue, the server is always idle at this instant. Let $S(t) = (w_1(t), w_2(t), \ldots, w_k(t))$ denote the complete state of an M/GI/1 non-preemptive queue at $t \geq t_0$ when there are k customers in the buffer. These customers are labeled as $c_1(t), c_2(t), \ldots, c_k(t)$ and customer $c_i(t)$ having been in the buffer for $w_i(t)$ time units, $w_i(t) \leq w_{i+1}(t)$, $i \geq 1$. Note that the customer corresponding to $c_i(t)$ can vary with the time, as well its current waiting time $w_i(t)$, in contrast to C_i and W_i defined earlier. When there is no confusion, we will omit the argument t in $c_i(t)$ and $w_i(t)$ for notational convenience. We define the set of feasible states to be

$$S(t) = \begin{cases} \{\Phi\} & \text{empty buffer;} \\ \{(w_1, w_2, \ldots, w_k) | 0 \leq w_1 \leq w_2 \leq \ldots \leq w_k, k = 1, 2, \ldots\} & \text{otherwise.} \end{cases} \quad (5)$$

It is also useful to assign order to the waiting rooms in the buffer, which starts from the first waiting room occupied by the customer with the shortest waiting time. Then c_1 is the customer in that first room at time t.

We are going to find an optimal *stationary* policy for this M/GI/1 queue under assumption $A1$. For an optimal LIFO-TO policy, its rejection scheme can be emulated by the following "delayed" scheme which makes use of Lemma 2.2. Let a rejected customer leave the buffer either when it reaches the server in LIFO order or when another customer at the first waiting room is rejected. In other words, we can tag every rejected customer and keep it in the buffer until either of the rejection moments described above. With LIFO order, a "delayed" rejection always throws away the customer with the shortest waiting time and thus results in an empty queue.

While a customer c_1 reaches server under LIFO order, an optimal policy could either serve or reject it. It is also possible to insert an unforced idle period to the service sequence. This means that c_1 is held in the buffer while the server is kept idle until a fresh arrival, at which point either the new customer is served or another unforced idle time period may commence. As is proved in the next lemma, there is no advantage in allowing unforced idle times.

Lemma 3.1 *For a non-preemptive M/GI/1 queue, an optimal stationary LIFO-TO queueing policy will not allow unforced idle times.*

The proof is in the Appendix.

From the above lemma, we conclude that an optimal queueing policy belongs to LIFO-TO class without unforced idle periods. Under this optimal policy, the server treats customers in LIFO order, either serves a customer or rejects it. We further define a decision function $u_\pi : S(\tau) \mapsto \{0, 1\}$ associated with a stationary

policy π at decision instant τ as follows:

$$u_\pi[S(\tau)] = \begin{cases} 0 & \pi \text{ rejects } c_1; \\ 1 & \pi \text{ serves } c_1. \end{cases} \quad (6)$$

Because a rejection can bring the state equivalent to the null set (an empty queue), for any stationary policy π and $\tau \geq 0$, we should have

$$u_{\pi^*}[S(\tau)] = \begin{cases} 0 \text{ when } G_{\pi^*}(\{\Phi\}) \geq_{st} G_\pi(S(\tau)); \\ 1 \text{ when } G_{\pi^*}(S(\tau)) \geq_{st} G_{\pi^*}(\{\Phi\}). \end{cases} \quad (7)$$

Theorem 3.2 *Consider a non-preemptive M/GI/1 queue under assumption A1. If the customers' deadlines are independent and identically distributed with a concave function $F_d(.)$, there exists an optimal stationary policy in LIFO-TO class with a fixed critical time T. Unforced idle time is not allowed under this policy.*

Proof: From Lemma 2.2, an optimal policy could be LIFO-TO type. Lemma 3.1 shows that unforced idle times provide no advantage. Here we are going to prove that a fixed critical time T is optimal. Let w_{i1} be the smallest waiting time for which $u_{\pi^*}[S(t_i)] = 0$ with $S(t_i) = (w_{i1}, w_{i2}, \ldots)$ at t_i. Suppose T is not fixed, then there exists a state $S(t_j) = (w_{j1}, \ldots, w_{jm}, w_{j(m+1)}, \ldots)$ with $w_{j1} \geq w_{i1}$, and $u_{\pi^*}[S(t_j)] = 1$ at t_j. If c_{jm}, $m \geq 1$, is served at time $t_j + \sigma$, and $c_{j(m+1)}, \ldots$ are rejected later, then

$$G^t_{\pi^*}(S(t_j + \sigma)) \geq_{st} G^t_{\pi^*}(\{\Phi\}) \quad (8)$$

with $S(t_j + \sigma) = (w_{jm} + \sigma, w_{j(m+1)} + \sigma, \ldots)$. At the next decision instant, π^* will either reject $c_{j(m+1)}, \ldots$ if no customer arrives during the service to c_{jm}, or serve the latest arrival.

Assume another stationary policy π' is identical to π^* except that it serves c_{i1} and rejects $c_{i2}, \ldots$ at t_i. Hence from equation (7)

$$G^t_{\pi^*}(\{\Phi\}) \geq_{st} G^t_{\pi'}(S(t_i)). \quad (9)$$

In a manner similar to the policy π^* after time $t_j + \sigma$, π' will next either reject $c_2, \ldots$ or serve the latest arrival at its next decision instant. Let us consider an increasing function $f(G^t_\pi(S_0)) = G^t_\pi(S_0)$ as the one used in inequality (1). We have

$$\begin{aligned} & \mathrm{E}\left[G^t_{\pi'}(S(t_i))\right] - \mathrm{E}\left[G^t_{\pi^*}(S(t_j + \sigma))\right] \\ = \; & \mathrm{E}\left[G^t_{\pi'}(S(t_i), \tilde{s})\right] - \mathrm{E}\left[G^t_{\pi^*}(S(t_j + \sigma), \tilde{s})\right] \\ = \; & F_d(w_{jm} + \sigma) - F_d(w_{i1}), \end{aligned} \quad (10)$$

since only the successful service probabilities of c_{i1}, c_{jm} are different. Consider

$$w_{i1} \leq w_{j1} < w_{jm} + \sigma \quad (11)$$

and $F_d(.)$ is non-decreasing in waiting time, so equation (10) is non-negative. Then from inequality (8),

$$\begin{aligned} \mathrm{E}\left[G^t_{\pi'}(S(t_i))\right] & \geq \mathrm{E}\left[G^t_{\pi^*}(S(t_j + \sigma))\right] \\ & \geq \mathrm{E}\left[G^t_{\pi^*}(\{\Phi\})\right]. \end{aligned} \quad (12)$$

Any contradiction between inequalities (9) and (12) shows that c_{jm} should not be served under π^*. To avoid this contradiction, the "$\geq$" symbols in inequality (12) actually should be "$=$" symbols, which matches with a change from "$\geq_{st}$" to "$=_{st}$" in inequality (9). Since inequality (12) will still be true if any other increasing function of G is chosen for equation (10), we have

$$G_{\pi^*}(S(t_j + \sigma)) =_{st} G_{\pi^*}(\{\Phi\}).$$

Then from equation (7), c_{jm} should be rejected after $t_j + \sigma$.

By an argument similar to the one shown above for c_{jm}, $c_{j(m-1)}$ and then $c_{j(m-2)}, \ldots, c_{j1}$ should not be served either. As a consequence, the optimal policy π^* should not serve any customers with waiting time longer than w_{i1}. This implies that the critical time T is a constant. ∎

Note that the above proof does not depend on the concavity of $F_d(.)$. This leads to the following corollary.

Corollary 3.3 *For any customer deadline distribution $F_d(.)$, an optimal stationary LIFO-TO policy with no unforced idle times for an M/GI/1 non-preemptive queue has a fixed critical time T.*

When λ, $F_d(.)$ and the service time distribution are given, the fixed critical time T in an optimal LIFO-TO policy can be determined. Therefore a customer with waiting time exceeding T can be rejected immediately.

4 The M/M/1 queue with a reduced information structure

Buffer size is the only information that can be used by a queueing policy under assumption *A2*. When the customer service times are exponentially distributed, we can extend the LIFO-TO results in Section 2 and Section 3 to LIFO-PO policies. Since we can compare the customer waiting times by their arrival order, Theorem 2.1, Lemma 2.2 and Lemma 3.1 still hold here. Thus LIFO is still the service order for customers and unforced idle times are not allowed. A TO rejection scheme cannot be implemented because the customer waiting times are unknown. Under a PO scheme, a customer reaching some

"end" position of a queue may be pushed out by an arrival. This is reasonable since the rejected customer is expected to have the longest waiting time.

To emulate a policy under assumption $A2$ by the "delayed" rejection scheme, we may again tag the rejected customers. The tagged customers leave the buffer only at the moment when a customer in the first waiting room is rejected. We define a customer's push-up index η as follows in order to estimate its waiting time at the decision instants. We assign a push-up index $\eta = 0$ to every arriving customer. If a customer visits the k-th waiting room but no room beyond it, then $\eta \triangleq k$. Let W^k denote the random waiting time of a customer with push-up index k, $k = 0, 1, \ldots$, from the time of its arrival to the time it reaches the server just before service or rejection. Then $W^0 = 0$ since an arrival joining an empty queue could get served immediately. We have the following stochastic order relations[9][11] for W^η's.

Lemma 4.1 *For the M/M/1 queue, W^η has the stochastic monotonicity property with respect to the push-up index η, i.e. $W^0 \leq_{st} W^1 \ldots \leq_{st} W^{k-1} \leq_{st} W^k \ldots$.*

Proof: Since a rejection results in an empty queue, the queueing system starts with the initial null set state after every rejection. Thus our M/M/1 model still gives a Markov process between any two rejections. We use the induction method to prove the above lemma as follows.
Basic Step: When $\eta = 1$, W^1 is the residual service time conditionally distributed on the event that the ongoing service completes before the next arrival instant. That is

$$\begin{aligned} & \Pr(W^1 \leq t) \\ = \; & \Pr(\text{ Residual service time} \leq t \mid \text{Ongoing service completes before next arrival }) \\ = \; & 1 - e^{-(\lambda+\mu)t} \qquad t \geq 0. \end{aligned}$$

Because of the memoryless property of Markov process and the LIFO customer service order, W^1 is independent of the service time received by the customer present in the server (or C^1's arrival instant) and the buffer occupancy at C^1's arrival instant. Obviously, we have $W^1 \geq_{st} W^0 = 0$.
Inductive Step: Now assume $W^k \geq_{st} W^{k-1}$, $k > 1$, and W^k is independent of C^k's arrival instant and the buffer occupancy at that moment. For a customer C^{k+1}, its waiting time W^{k+1} can be decomposed into a sequence of random intervals as follows. Since its push-up index $k+1 > 1$, C^{k+1} should wait for a new customer pushing it up to the second waiting room after its arrival and visit the $(k+1)$-th room at least once. On the other hand, any customer arriving during C^{k+1}'s waiting period can at most reach up to the k-th waiting room, and it would be scheduled before C^{k+1} under LIFO service order. Let α denote the first time period that C^{k+1} spends in the first room after its arrival and before it is pushed up by a new arrival. Then α is the interarrival period conditionally distributed on the event that the new arrival comes before the completion of ongoing service. We have

$$\begin{aligned} & \Pr(\alpha \leq t) \\ = \; & \Pr(\text{ Interarrival time} \leq t \mid \text{Next arrival comes before ongoing service completion }) \\ = \; & 1 - e^{-(\lambda+\mu)t} \qquad t \geq 0. \end{aligned}$$

α has the same exponential distribution function as W^1. Let $W_{(k+1)i}$ denote the i-th sub-waiting period, which starts at an arrival instant when C^{k+1} is pushed up to the second room by a new arrival for the i-th time, and ends at the moment when that new customer begins its service. Under LIFO service order, $W_{(k+1)i}$ is equal to the waiting time of that new customer to the beginning of its service. This waiting period can be estimated by the new customer's push-up index, which is less than $k+1$ since the new customer can at most reach up to the k-th waiting room. Thus $W_{(k+1)i}$ take values from $\{W^1, W^2, \ldots, W^k\}$. Between any two sub-waiting periods, C^{k+1} could stay in the first room for α time units as if it just joined buffer. Assume during the n-th sub-waiting period, $n \geq 1$, C^{k+1} reaches the $(k+1)$-th waiting room for its last time, then $W_{(k+1)n} = W^k$. After this last sub-waiting period, we define a residual waiting period of C^{k+1}, ΔW^{k+1}. ΔW^{k+1} only can take its value from $\{W^1, W^2, \ldots, W^k\}$ because C^{k+1} visits the $(k+1)$-th room for the last time during the n-th sub-waiting period.

In Figure 1, we give a sample waiting time structure of C^4 as an example. During W^4, there are three sub-waiting periods: $W_{51} = W^3$, $W_{52} = W^1$ and $W_{53} = W^3$. During W_{53}, C^4 is pushed up from the first waiting room to the 4-th room for the second and also the last time. After that period, C^4 is pushed up and down several times and $\Delta W^5 = W^3$.

Now we have

$$\begin{aligned} & W^{k+1} \\ = \; & \sum_{i=1}^{n} (\alpha + W_{(k+1)i}) + \Delta W^{k+1} \\ = \; & \alpha + W^k + \sum_{i=1}^{n-1} (\alpha + W_{(k+1)i}) + \Delta W^{k+1}. \end{aligned} \tag{13}$$

Again from memoryless property of Markov process, n is independent of C^{k+1}'s arrival instant and the buffer occupancy at that moment. Therefore from equation (13), W^{k+1} is also independent of above random factors. Since

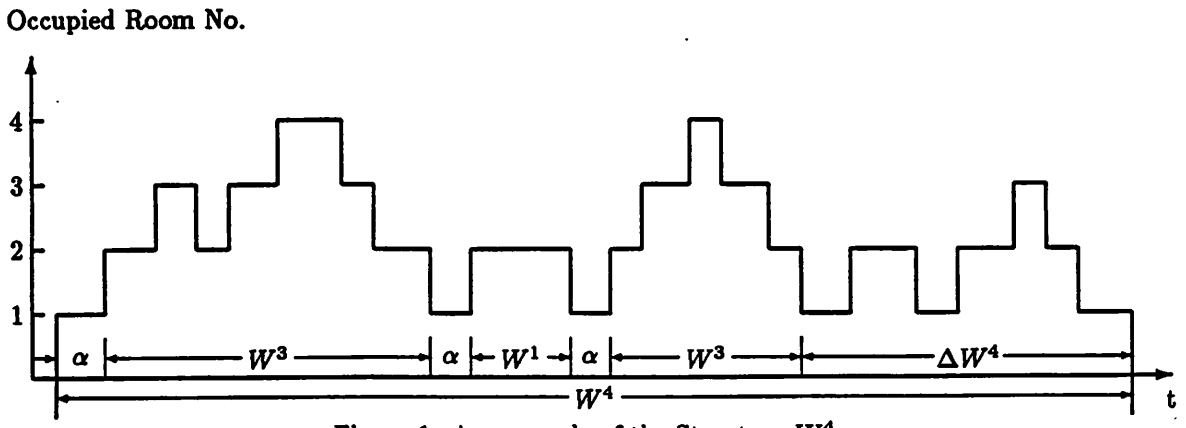

Figure 1: An example of the Structure W^4

W^{k+1} is the summation of non-negative random variables α, W^k and ΔW^{k+1}, from equation (13) we have

$$W^k > c \Longrightarrow W^{k+1} > c,$$

and then

$$\Pr(W^k > c) \leq \Pr(W^{k+1} > c),$$

which follows $W^k \leq_{st} W^{k+1}$. ∎

Now the queueing state becomes

$$S(t) = \begin{cases} \{\Phi\} & \text{empty buffer;} \\ \{(w^{\eta_1}, w^{\eta_2}, \ldots, w^{\eta_k}) | 0 \leq w^{\eta_1} \leq \\ \quad w^{\eta_2} \leq \ldots \leq w^{\eta_k}, k = 1, 2, \ldots\} & \\ & \text{otherwise,} \end{cases}$$

where w^{η_i} is the waiting time of c_i, with push-up index η_i, from its arrival instant to t. At a time moment t, $\{\eta_i\}_{i=1}^k$ are positive integers increasing in i and $\eta_i \geq i$ for all i. The discontinuity of push-up indices indicates that the corresponding customer(s) was pushed away from the server by new arrival(s). The customer waiting times at the different times can be stochastically ordered by their push-up indices (Lemma 4.1). Using the coupling method[9][11], a set of random variables, $w'^{\eta_1} \leq w'^{\eta_2} \ldots \leq w'^{\eta_i} \ldots \leq w'^{\eta_k}$, can be used to analyze the queueing system, where w'^{η_i} has the same distribution as w^{η_i} for all i. By using $\{w'^{\eta_i}\}_{i=1}^k$ as customer waiting times, we can show, in a manner similar to the one used in Theorem 3.2, that LIFO-PO with fixed buffer size is optimal.

Theorem 4.2 *Consider a non-preemptive M/M/1 queue under assumption A2. If the customers' deadlines are independent and identically distributed with a concave function $F_d(.)$, there exists an optimal stationary policy in LIFO-PO class with a fixed buffer size used as a rejection threshold. Unforced idle times are not allowed under this policy.*

Proof: Since the approach is similar to the one in proof of Theorem 3.2, we only point out the essential differences here. Now the queueing state consists of a set of waiting times $\{w^{\eta_i}\}_{i=1}^k$. The TO scheme becomes a rejection scheme which pushes out a set of customers according to their push-up indices. We need to prove that this rejection scheme is a PO scheme with a fixed threshold on the push-up index to reject customers.

Assume that $u_{\pi^*}[S(t_i)] = 0$ with $S(t_i) = (w^{\eta_{i1}}, w^{\eta_{i2}}, \ldots)$ under an optimal policy π^*, and the buffer size is not fixed. Then there exist a state $S(t_j) = (w^{\eta_{j1}}, \ldots, w^{\eta_{jm}}, w^{\eta_{j(m+1)}}, \ldots)$ with $\eta_{j1} \geq \eta_{i1}$, i.e. $w^{\eta_{j1}} \geq_{st} w^{\eta_{i1}}$, and $u_{\pi^*}[S(t_j)] = 1$. Suppose c^{jm}, $m \geq 1$, is served at $t_j + \sigma$ and $c^{j(m+1)}, \ldots$ are rejected later, then π^* will either reject $c^{j(m+1)}, \ldots$ or serve the latest arrival arriving during the service to c^{jm} at the next decision instant.

Let π' be identical to π^* except that it serves c^{i1} and rejects $c^{i2}, \ldots$ at t_i. In a manner similar to policy π^* after time $t_j + \sigma$, at its next decision instant π' will either make a rejection leading to an empty queue, or serve the latest arrival. By using the monotonicity property of W^η and the coupling method, we have

$$w^{\eta_{i1}} \leq_{st} w^{\eta_{j1}} \leq_{st} w^{\eta_{jm}} + \sigma. \tag{14}$$

Then $w'^{\eta_{i1}}$ and $w'^{\eta_{jm}}$ can be chosen such that

$$w'^{\eta_{i1}} \leq w'^{\eta_{jm}} + \sigma. \tag{15}$$

instead of $w^{\eta_{i1}}$ and $w^{\eta_{jm}}$ in the analysis.

As has been shown in proof of Theorem 3.2, inequalities (8) and (9) also hold here. When $w^{\eta_{i1}}$ and $w^{\eta_{jm}}$ are

used instead of w_{i1} and w_{jm}, equation (10) is still non-negative because of inequality (15). This again results in a contradiction in inequality(12) with inequality (9).

By an argument similar to the one shown above for c^{jm}, $c^{j(m-1)}$ and then $c^{j(m-2)}$, ..., c^{j1}, should not be served either. Since π^* could not serve any customers with waiting time longer than $w^{\eta_{i1}} =_{st} w^{\eta_{i1}}$, the TO scheme should reject all the customer with push-up index larger than or equal to η_{i1}. This leads to a PO scheme with a fixed threshold. ∎

The concavity of $F_d(.)$ is not used in the above proof. This leads to the following corollary.

Corollary 4.3 *For any customer deadline distribution $F_d(.)$, an optimal stationary LIFO-PO policy with no unforced idle times for an M/M/1 non-preemptive queue uses a fixed buffer size as a rejection threshold.*

5 Summary

We discussed the optimality of queueing policies for non-preemptive queues with impatient customers. The deadlines of these customers are set to the beginnings of their services, and only the customer deadline distribution is given. In general, PO is at least as good as BL when the service discipline is specified. For example, under a FIFO discipline, it is better to push out a customer at the head of queue rather block a new arrival. If the deadline distribution is a concave function, an optimal policy for the G/GI/1 queue exists in LIFO-TO policy class. With a Poisson arrival process, this optimal policy does not insert unforced idle times, and it's critical time is a fixed constant for a given set of system parameters. Furthermore, if the customer waiting times are not available, an optimal policy for the M/M/1 queue is in LIFO-PO class with a fixed finite buffer size. Though not presented here, these results can be extended when the customer deadlines are set to the ends of their services and also to some equivalent multi-server queues[16].

Appendix

Proof of Lemma 3.1: Suppose π^* is an optimal LIFO-TO policy for which the hypothesis is not satisfied. If an unforced idle period of σ time units is inserted in the service sequence from an arbitrary sample path at state $S(t_0) = (w_1, w_2, \ldots, w_k)$, then

$$G_{\pi^*}(S(t_0)) \geq_{st} G_{\pi}(S(t_0)) \qquad \forall \pi \text{ and } t \geq t_0. \quad (16)$$

Under LIFO service order, all new arrivals are scheduled before $c_1, c_2, \ldots, c_k$. If any one of them is rejected, $c_1, c_2, \ldots, c_k$ would be rejected anyway (Lemma 2.2). Otherwise all the new arrivals do get served in a time period of θ time units, and then c_1 reaches the server again at $t_0 + (\sigma + \theta) = t_0 + \delta$ with $c_2, \ldots, c_k$ behind it. The queueing state becomes $S(t_0 + \delta) = (w_1 + \delta, w_2 + \delta, \ldots, w_k + \delta)$, which is the state $S(t_0)$ time-shifted by δ. At $t_0 + \delta$, π^* can decide to serve c_1 or reject it, or even insert another unforced idle time. We first show by contradiction that serving c_1 is not an optimal decision.

Let $\{C_s\}$ denote the set of customers $\{c_1, c_2, \ldots, c_k\}$, which are either served or rejected eventually under π^*. Assume $c_i \in \{C_s\}$, $0 \leq i \leq k$, leaves the buffer τ_i time units after $t_0 + \delta$ when the customer deadlines are set to the beginning of their services. With LIFO service order, $\tau_1 \geq 0$, $\tau_i \leq \tau_j$ when $i < j$. We further define $\tau_i = \infty$ if c_i is rejected. Also, let $C_i \notin \{C_s\}$ be a new arrival from $\tilde{s}$ with a waiting time W_i at the beginning of its service. Consider a policy π' which satisfies inequality (16). It serves c_1 at t_0. After t_0, it behaves like π^* after $t_0 + \delta$. Because the interarrival time is memoryless and the service times are independent and identically distributed, we can couple the sample path after $t_0 + \delta$ under π^* with that after t_0 under π'. Suppose π' schedules new arrivals after t_0 in the same manner as π^* schedules new arrivals after $t_0 + \delta$. Then the same set of customers in $\{C_s\}$ are scheduled under both π' and π^*, but π' schedules customers in $\{C_s\}$ when they have shorter waiting times. To compare the queueing performances stochastically, let us choose an increasing function $f(G^t_\pi(S_0)) = G^t_\pi(S_0)$ as the one used in inequality (1) when $k = 1$. From equation (2) and inequality (16) we have

$$\begin{aligned}
& \mathrm{E}\left[G^t_{\pi^*}(S(t_0))\right] - \mathrm{E}\left[G^t_{\pi'}(S(t_0))\right] \\
= \; & \mathrm{E}\left[G^t_{\pi^*}(S(t_0), \tilde{s})\right] - \mathrm{E}\left[G^t_{\pi'}(S(t_0), \tilde{s})\right] \\
= \; & \mathrm{E}\Bigg[\sum_{c_i \in \{C_s\}, t_i \in [t_0, t]} [\Pr(w_i + \delta + \tau_i < d_i | \pi^*, S(t_0), \tilde{s}) \\
& \qquad - \Pr(w_i + \tau_i < d_i | \pi', S(t_0), \tilde{s})] \\
& \quad + \sum_{C_i \notin \{C_s\}, t_i \in [t_0, t]} [\Pr(W_i < d_i | \pi^*, S(t_0), \tilde{s}) \\
& \qquad - \Pr(W_i < d_i | \pi', S(t_0), \tilde{s})] \Bigg] \\
\geq \; & 0 \qquad \forall t \geq t_0, \qquad (17)
\end{aligned}$$

where t_i is the service beginning instant of the corresponding customer c_i or C_i.

We further construct another policy π'' which is identical to π^* in interval $[t_0, t_0 + \delta)$. At $t_0 + \delta$, π'' inserts another unforced idle period σ'', and then it behaves just like π^* after the insertion of the idle time σ at t_0. Note that π'' can make all the decisions after $t_0 + \delta$ as if the system with initial state $S(t_0)$ was governed by π^*.

Based on the same reasons as shown above for π^* and π', we also can couple the sample path under π'' after $t_0+\delta$ with that under π^* after t_0. As customers in $\{C_s\}$ could reach the server under π^* at $t_0+\delta$, they are scheduled by π'' at $t_0+\delta+\delta''$ under the coupled sample path. Here δ'' and δ are identically distributed, but δ'' is independent of δ because of the memoryless property. Thus we can have an expression similar to expression (17) for all $t \geq t_0+\delta$ as follows:

$$\begin{aligned}
&\mathrm{E}\Big[G^t_{\pi''}(S(t_0+\delta))\Big] - \mathrm{E}\Big[G^t_{\pi^*}(S(t_0+\delta))\Big] \\
&= \mathrm{E}\Big[G^t_{\pi''}(S(t_0+\delta),\tilde{s})\Big] - \mathrm{E}\Big[G^t_{\pi^*}(S(t_0+\delta),\tilde{s})\Big] \\
&= \mathrm{E}\Bigg[\sum_{c_i\in\{C_s\},t_i\in[t_0+\delta,t]} [\Pr(w_i+\delta+\delta''+\tau_i<d_i \\
&\qquad |\pi'',S(t_0+\delta),\tilde{s}) \\
&\qquad -\Pr(w_i+\delta+\tau_i<d_i|\pi^*,S(t_0+\delta),\tilde{s})] \\
&\quad + \sum_{C_i\notin\{C_s\},t_i\in[t_0+\delta,t]} [\Pr(W_i<d_i|\pi'',S(t_0+\delta),\tilde{s}) \\
&\qquad -\Pr(W_i<d_i|\pi^*,S(t_0+\delta),\tilde{s})]\Bigg]. \qquad (18)
\end{aligned}$$

The second term in expression (17) and (18) can be rewritten as

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{C_i\notin\{C_s\},t_i\in[0,t]} [\Pr(W_i<d_i|\pi_1,S_0,\tilde{s}) \\
&\qquad -\Pr(W_i<d_i|\pi_2,S_0,\tilde{s})]\Bigg], \qquad (19)
\end{aligned}$$

where S_0 is the initial state for sample path $\tilde{s}$, and π_1,π_2 are the two corresponding policies. Equation (19) is the performance difference between π_1 and π_2 awarded by the services to new arrivals ($\notin \{C_s\}$) after 0. With the memoryless interarrival times and the couplings of sample path and policy, expression (19) and the second term in expressions (17) and (18) are actually equal. The rest of the terms in expressions (17) and (18) have the form

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{c_i\in\{C_s\},t_i\in[0,t]} [\Pr(w_i+\nu+\delta<d_i|\pi_1,S_0,\tilde{s}) \\
&\qquad -\Pr(w_i+\nu<d_i|\pi_2,S_0,\tilde{s})]\Bigg] \\
&= \mathrm{E}\Bigg[\sum_{c_i\in\{C_s\},t_i\in[0,t]} [(1-F_d(w_i+\nu+\delta)) \\
&\qquad -(1-F_d(w_i+\nu))]\Bigg] \qquad (20)
\end{aligned}$$

with $\nu \geq 0$ as the time shift period and S_0 as the initial state. Since $1-F_d(.)$ is a convex non-increasing function, equation (20) is non-positive and non-decreasing in ν. Hence

$$\begin{aligned}
&\mathrm{E}\Big[G^{t+\delta}_{\pi''}(S(t_0+\delta))\Big] - \mathrm{E}\Big[G^{t+\delta}_{\pi^*}(S(t_0+\delta))\Big] \\
&\geq \mathrm{E}\Big[G^t_{\pi^*}(S(t_0))\Big] - \mathrm{E}\Big[G^t_{\pi}(S(t_0))\Big] \\
&\geq 0 \qquad \forall t \geq t_0. \qquad (21)
\end{aligned}$$

Equation (21) shows that π'' could be better than π^* for the increasing function $f(G^t_\pi(S_0)) = G^t_\pi(S_0)$ when $t \geq t_0+\delta$. This contradicts inequality (16).

We have just shown that c_1 will not be served at $t_0+\delta$. Thus either it is rejected or another unforced idle period is inserted at this instant. If there is an insertion at $t_0+\delta$, then based on the same reasoning as above, an unforced idle period will always be inserted whenever c_1 reaches the server. Thus c_1 will never be served. As a result, an optimal policy actually would not lose anything by rejecting $c_1, c_2, \ldots, c_k$, which leads to a forced idle period at time t_0. ∎

References

[1] F. Baccelli and G. Hebuterne, "On Queues with Impatient Customers", *PERFORMANCE '81*, North-Holland Publishing Company, 1981, *pp.* 159-179.

[2] P. Bhattacharya and A. Ephremides, "Optimal Scheduling with Strict Deadlines", *IEEE Transactions on Automatic Control*, Vol. 34, No. 7, July 1989, *pp.* 721-728.

[3] B. T. Doshi and H. Heffes, "Overload Performance of Several Processor Queueing Disciplines for the M/M/1 Queue", *IEEE Transactions on Communications*, Vol. Com-34, No. 6, June 1986, *pp.* 538-546.

[4] B. T. Doshi and E. H. Lipper, "Comparisons of Service Disciplines in a Queueing System with Delay Dependent Customer Behavior", *Applied Probability-Computer Science: The Interface*, Vol II, R. L. Disney and T. J. Ott, Eds. Cambridge, MA: Birkhauser, 1982, *pp.* 269-301.

[5] P. R. de Waal, "Performance Analysis and Optimal Control of an M/M/1/k Queueing System with Impatient Customers", *Report OS-R8713*, Center for Mathematics and Computer Science, Amsterdam, The Netherlands, 1987.

[6] M. H. Kallmes, D. Towsley and C. G. Cassandras, "Optimality of the Last-In-First-Out (LIFO) Service discipline in Queueing Systems with Real-Time

Constraints", *Proceeding of the 28th Conference on Decision and Control*, December 1989, *pp.* 1073-1074.

[7] S. S. Panwar, "Time-Constrained and Multiaccess Communications", Ph.D. Dissertation, University of Massachusetts, Amherst, 1985.

[8] S. S. Panwar, D. Towsley and J. K. Wolf, "Optimal Scheduling Policies for a Class of Queues with Customers Deadlines to the Beginning of Service", *Journal of the Association for Computing Machinery*, Vol. 35, No. 4, October 1988, *pp.* 832-844.

[9] S. M. Ross, "Stochastic Processes", John Wiley & Sons, 1983.

[10] J. G. Shanthikumar and U. Sumita, "Convex Ordering of Sojourn Times in Single-Server Queues: Extremal Properties of FIFO and LIFO Service Disciplines", *Journal of Applied Probability*, Vol. 24, 1987, *pp.* 737-748.

[11] D. Stoyan, "Comparison Methods for Queues and Other Stochastic Models", John Wiley & Sons, 1983.

[12] J. P. C. Blanc, P. R. de Waal, P. Nain and D. Towsley, "A New Device for the Synthesis Problem of Optimal Control of Admission to an M/M/c Queue", *COINS Technical Report 90-90*, Department of Computer and Information Science, University of Massachusetts, October 1990.

[13] D. Towsley and S. S. Panwar, "On the Optimality of the STE Rule for Multiple Server Queues that Serve Customers with Deadlines", *COINS Technical Report 88-81*, Department of Computer and Information Science, University of Massachusetts, July 1988.

[14] J. Walrand, "Queueing Networks", Prentice-Hall, 1988.

[15] R. W. Wolff, "Stochastic Modeling and the Theory of Queues", Prentice-Hall, 1989.

[16] Z. Zhao, "Optimal Scheduling Policies for Real-Time Messages", Ph.D Dissertation, Polytechnic University, Brooklyn, New York, Under preparation.