Stochastic Processes and their Applications 6 (1977) 25–32.
© North-Holland Publishing Company

ON OPTIMAL RIGHT-OF-WAY POLICIES AT A SINGLE-SERVER STATION WHEN INSERTION OF IDLE TIMES IS PERMITTED

Isaac MEILIJSON
Uri YECHIALI

Department of Statistics, Tel-Aviv University, Ramat Aviv, Tel-Aviv, Israel

Received 7 July 1976
Revised 19 November 1976

A general stream of n types of customers arrives at a Single Server station where service is non-preemptive, the server may undergo Poisson breakdowns and insertion of idle times is allowed. If $\xi(k)$ and $c(k)$ are, respectively, the expected service time and sojourn cost per unit time of a type k customer $(1 \leq k \leq n)$, call k "V.I.P." type if $\xi(k)/c(k) = \min_{1 \leq i \leq n}[\xi(i)/c(i)]$.

We show that any right-of-way service policy can be improved by a policy that grants V.I.P. customers priority over all others, and never inserts idle time when a V.I.P. customer is present.

We further show that if the arrival stream is Poisson, the so-called "$c\mu$" priority rule (applied with no delays) is optimal in the class of all service policies, and not just among those of a priority nature.

G/GI/1 queue	insertion of idle-times
right-of-way policies	"$c\mu$" priority rule

1. Introduction

The subject to be presented is best motivated by the following example.

Customers of three types arrive at a Single Server Station. Interarrival times are either 1 or 9 minutes with equal probabilities, and the type of each arriving customer is 1, 2 or 3 with equal probabilities. Service, which may not be interrupted prior to conclusion, lasts 1 minute for type 1 customers, 3 minutes for type 2 customers and 2 minutes for type 3 customers.

Sojourn costs per unit time are very high for type 1 customers, high for type 2 customers and low for type 3 customers.

The only service routine that makes sense, if we are to minimize total sojourn costs, serves type 1 customers as soon as they arrive, and never faces the possibility of having an arrival during a service period. In other words, one idle minute is always inserted following the arrival of a customer of type 2 or 3. If no customer showed up then the next 8 minutes are used to serve at most two customers of type 2 (if any) and then as many customers of type 3 that will fit in that period.

It has been customary in Queuing theory to deal with static priority rules. The above example shows that optimization considerations might dictate the use of no-priority rules, as well as the insertion of idle time. The only static preferential feature in the example is the priority service with no delay to customers of type 1. The main result of this paper (Theorem 1) is that this feature holds in general for G/GI/1 queues. The preferred type is simply the one for whom the ratio of expected length of service to expected rate of cost is minimal; its determination has nothing to do with the stream, which we allow to be *arbitrary*.

The so called "$c\mu$" priority rule is known to be best for the service of a one-time batch of customers (e.g., Baker [1]) and to be best among the static priority rules for the service of a M/GI/1 queue (e.g. Jaiswal [3]). This last result has been recently strengthened by Klimov [4], Harrison [2], and Meilijson and Weiss [5], who showed the optimality of the "$c\mu$" rule.

Our second result (Theorem 2) is an extension and a new proof of the optimality of the "$c\mu$" rule for the M/GI/1 case. We permit the insertion of idle time — and show that it doesn't pay to do it — and we let the server the possibility of undergoing breakdowns.

2. The model and results

We deal with the control of a non-preemptive G/GI/1 queue with n types of customers. Denote service times by V and holding costs by C. Assume that, given the types, pairs (V, C) corresponding to different customers are independent; assume further that pairs (V, C) of customers of the same type are identically distributed.

Independently of the stream of customers and their service times the server may undergo breakdowns. It is assumed that breakdowns follow a Poisson law, the server is repaired as soon as it breaks down, and repair times are i.i.d. If a breakdown occurs during the service of a customer, the service of this customer resumes as soon as the server is again operative.

The control of this system is carried out at *decision moments*, which may be of three sorts:

(1) the service of a customer has just been completed and the queue is not empty,

(2) a customer has just arrived and the server is idle and operative (we remark that the server may be idle even if there are customers in the queue),

(3) the server has just turned operative, following a breakdown while it was idle, and the queue is not empty.

At a decision moment the server marks as candidate for service one of the customers present in the queue, and he also selects an *idle lag* $0 \leq x \leq \infty$; he then stays idle until the earliest of the following three events:

(1) a customer arrives — this is a new decision moment,

(2) the server breaks down — the end of the repair time is a new decision moment,

(3) x time units have elapsed. In the latter case, he serves the marked customer.

A rule specifying the action to be taken by the server at all decision moments is called a *right-of-way policy.*

For an arbitrary positive constant (henceforth called Tentative Closing Time, TCT) and a right-of-way policy π, define the (random) Actual Closing Time, $\mathrm{ACT}(\pi)$, as the least time t such that $t \geq \mathrm{TCT}$, at which there are no customers present in the system. The *payoff* $\Phi(\pi, \mathrm{TCT})$ of a right-of-way policy π with respect to TCT, is the expected total cost up to $\mathrm{ACT}(\pi)$, using π. The server's aim is to find for each TCT a π that minimizes $\Phi(\pi, \mathrm{TCT})$. Excluding trivial cases, assume that for the policies under consideration, $\mathrm{ACT}(\pi)$ is a.s. finite.

We remark that the usual average cost criterion can not be applied here since steady states need not exist for *general* streams. We will return to this question in Section 5.

Denote by $\xi(i)$ the expected service time and by $c(i)$ the expected holding cost of a customer of type i $(1 \leq i \leq n)$. A customer is V.I.P. if his type i satisfies $\xi(i)/c(i) = \min_{1 \leq j \leq n}[\xi(j)/c(j)]$. A right-of-way policy is a *top-class-policy* if it serves V.I.P. customers as soon as service to them becomes possible.

Theorem 1. *For every* TCT *and every right-of-way policy* π, *there exists a top-class policy* π^* *such that*:

(a) $\Phi(\pi^*, \mathrm{TCT}) \leq \Phi(\pi, \mathrm{TCT})$.

(b) *The three variables*: (1) $\mathrm{ACT}(\pi^*)$, (2) *the amount of time up to* $\mathrm{ACT}(\pi^*)$ *at which customers are present in the station, and* (3) *the amount of inserted idle time up to* $\mathrm{ACT}(\pi^*)$, *are stochastically smaller than the corresponding variables under* π.

Assume, for convenience, that $\xi(1)/c(1) \leq \xi(2)/c(2) \leq \cdots \leq \xi(n)/c(n)$.

Define "cμ" *priority rule*: At every decision moment, mark for service a customer whose type is the least one among those present in the queue, and choose time lag $x = 0$ (i.e., serve him immediately).

The following well known result is an immediate consequence of Theorem 1.

Corollary 1. *If, with probability one, there will be no arrivals beyond some time* t, *switching to the "cμ" rule beyond time* t *will improve any policy.*

Our next result applies to the M/GI/1 case. It generalizes earlier results obtained under related criteria (Klimov [4] and Harrison [2]) in that it permits inserted idle time and server breakdowns. Server breakdowns play here an important role: Theorem 2 follows from Theorem 1 by successively considering the service to preferred types as "repair" of "server breakdowns".

Theorem 2. *In the* M/GI/1 *model, the* "cμ" *priority rules are optimal for all* TCT.

3. Proof of Theorem 1

For a service time V, let $\tilde{V} \geq V$ be the length of time from the beginning of the service until this service has been completed. This time includes the repair of possible breakdowns. It is well known that $\mathbf{E}(\tilde{V})$ is proportional to $\mathbf{E}(V)$. By giving $\tilde{V}$ the role of V, we will assume that the process of breakdowns and their repair is read only at the times at which the server is idle.

For convenience of exposition and without loss of generality, it will be assumed that the service time to a customer will be the same under any two policies being compared. (Work conservation [6].)

We will now express conveniently the total sojourn cost up to ACT under an arbitrary policy π.

A *busy period* is a time interval from a moment a customer arrives to an empty station until the next moment the station is empty. Assume throughout that the station is empty prior to epoch 0. Observe that the server need not be "busy" throughout the entire busy period.

Let $[t_1, t_1^*], [t_2, t_2^*], \ldots, [t_m, t_m^* = \mathrm{ACT}]$ be the consecutive busy periods of π.

The evolution of a busy period $[t_r, t_r^*]$ at which M_r customers were served can be expressed as $\tau_1 V_1 \tau_2 V_2 \cdots \tau_{M_r} V_{M_r}$, where $\tau_1 \geq 0$ is the length of time from t_r until service is started on the first customer, this service lasting V_1. If at the conclusion of the j^{th} service the waiting line is not empty, $\tau_{j+1} \geq 0$ is the length of time from that moment $(t_r + \sum_{i=1}^{j} (\tau_i + V_i))$ until service is started on the next customer, this service lasting V_{j+1}.

For $t \in [t_r, t_r^*]$ denote by $D(t)$ the collection of customers belonging to the busy period $[t_r, t_r^*]$ that completed service before time t.

Denote by $L_i = \sum_{r=1}^{i} M_r$ the total number of customers served under π in the first i busy periods, $i = 1, 2, \ldots, m$. j (with $1 \leq j \leq L_m$) will be the *name* of the j^{th} customer served. C_j is its *holding cost*, V_j is the *length* of its *service*, α_j (β_j) is the length of *time from its arrival* (*exit*) until the *end* of the busy period to which j belongs. $\alpha_j - \beta_j$ is the *sojourn* time of customer j at the station. Let $i(j)$ be the *type* of customer j.

The total cost can now be expressed as:

$$\sum_{j=1}^{L_m} (\alpha_j - \beta_j) C_j = \sum_{j=1}^{L_m} \alpha_j C_j - \sum_{j=1}^{L_m} \beta_j C_j = \sum_{j=1}^{L_m} \alpha_j C_j - \sum_{r=1}^{m} \int_{t_r}^{t_r^*} \Big(\sum_{j \in D(t)} C_j \Big) \mathrm{d}t. \quad (3.1)$$

Say that a policy π' *refines* a policy π if each busy period of π' is a sub-interval of some busy period of π.

Observe that if π' refines π, the value of the first sum in the right-hand side of (3.1) for π' is at most the value for π.

Hence, if we replace π by π' that increases the value of the second sum in the right hand side of (3.1) while at the same time refining π then π' improves π.

Compute $\int_{t_r}^{t_r^*}(\sum_{j\in D(t)} C_j)\mathrm{d}t$:

$$\int_{t_r}^{t_r^*}\Big(\sum_{j\in D(t)} C_j\Big)\mathrm{d}t = \sum_{j=L_{r-1}+1}^{L_r}\ \sum_{l=L_{r-1}+1}^{j-1} C_l(\tau_j + V_j) \quad \Big(\text{where } \sum_{t}^{t-1}(\cdot) = 0\Big). \tag{3.2}$$

Denote by F the σ-field generated by the history of the system up to t_{r-1}^* and by F_j, for $j > L_{r-1}$, the σ-field generated by the history of the system up to whichever comes first: t_r^* or the moment at which π would start serving j.

The event $\{L_r \geqslant j\}$ is F_j-measurable. On $\{L_r \geqslant j\}$ the types $i(l)$ of customers l with $l \leqslant j$ are F_j-measurable, and conditional probabilities given F_j make C_l and V_j independent random variables, distributed each according to its type (that was just said to be F_j-measurable). In addition, on $\{L_r \geqslant j\}$, τ_j is F_j-measurable. Hence, denoting by $I(\alpha)$ the indicator function of the set α,

$$\begin{aligned}
&\mathbf{E}\Big(\sum_{j=L_{r-1}+1}^{L_r}\ \sum_{l=L_{r-1}+1}^{j-1} C_l(\tau_j + V_j)\,\Big|\,F\Big) = \\
&= \mathbf{E}\Big(\sum_{j=L_{r-1}+1}^{\infty}\ \sum_{l=L_{r-1}+1}^{j-1} I\{L_r \geqslant j\}C_l(\tau_j + V_j)\,\Big|\,F\Big) \\
&= \mathbf{E}\Big(\mathbf{E}\Big(\sum_{j=L_{r-1}+1}^{\infty}\ \sum_{l=L_{r-1}+1}^{j-1} I\{L_r \geqslant j\}C_l(\tau_j + V_j)\,\Big|\,F_j\Big)\,\Big|\,F\Big) \\
&= \mathbf{E}\Big(\sum_{j=L_{r-1}+1}^{\infty}\ \sum_{l=L_{r-1}+1}^{j-1} I\{L_r \geqslant j\}c(i(l))(\tau_j + \xi(i(j)))\,\Big|\,F\Big) \\
&= \mathbf{E}\Big(\sum_{j=L_{r-1}+1}^{L_r}\ \sum_{l=L_{r-1}+1}^{j-1} c(i(l))(\tau_j + \xi(i(j)))\,\Big|\,F\Big).
\end{aligned} \tag{3.3}$$

We have thus obtained that

$$\mathbf{E}\Big(\sum\sum C_l(\tau_j + V_j)\,\Big|\,F\Big) = \mathbf{E}\Big(\sum\sum c(i(l))(\tau_j + \xi(i(j))\,\Big|\,F\Big), \tag{3.4}$$

i.e., conditional expectations of total costs during busy periods given the past are unchanged by replacing locally service times and costs by their expected values.

Denote

$$B_r = \sum_j \sum_l c(i(l))(\tau_j + \xi(i(j))). \tag{3.5}$$

The proof will be finished if we produce a policy π^* that does the following

(i) Refines π. (This will automatically yield claim (b) in the theorem.)

(ii) Serves V.I.P. customers as soon as possible.

(iii) When viewing all busy periods of π^* that are subintervals of $[t_r, t_r^*]$ as if they composed one busy period, the value of B_r for π^* is at least that of B_r for π.

If π itself serves V.I.P. customers at no delay, let $\pi^* = \pi$.

Suppose a V.I.P. customer could have been served as j_1^{th} customer at time

$t^{(1)} \in [t_r, t_r^*]$, but π served him instead as j^{th} customer ($j \geq j_1$) at a time $t > t^{(1)}$ (of course, still within the same busy period). Customers $j_1, j_1+1, \ldots, j-1$ are non V.I.P. if $j > j_1$.

Perform the following change: Serve the V.I.P. customer at time $t^{(1)}$, while preserving the inner order of all other service times, repair times, and portions or wholes of inserted idle times. This change is feasible, as justified by (α), (β), (γ) and (δ) below.

(α) The V.I.P. customer was available for service at time $t^{(1)}$ by hypothesis.

(β) All other customers are served not before they would have been served by π, so they have already arrived.

(γ) The particular way in which we realized the breakdown and repair times, and the requirement that service times be the same under all policies, insures the availability for service of all customers at the times allotted to them by the changed policy.

(δ) The information on which the changed policy is to decide what to do (imitating π) is available to it at all times, by (β).

We will now show that the changed π satisfies (iii). Verify that, when replacing π by the changed π, B_r is added the term

$$\Delta B_r = c(i(j))\left\{\left(\Delta\tau + \sum_{l=j_1+1}^{j} \tau_l\right) + \sum_{l=j_1}^{j} c(i(l))\left[\frac{\xi(i(l))}{c(i(l))} - \frac{\xi(i(j))}{c(i(j))}\right]\right\}. \tag{3.6}$$

($\Delta\tau$ is what is missing at time $t^{(1)}$ to complete an inserted idle time or a repair).

Since $\xi(i(j))/c(i(j)) = \min_i(\xi(i)/c(i))$, ΔB_r is positive, as claimed. So the changed π satisfies (iii). It obviously satisfies (i).

Now build π^* as follows: Follow π until completion or until the first time a change as described could be applied. If so, do, and proceed likewise. The satisfaction of (i) for single changes implies the satisfaction of (i) for π^*. (ii) is satisfied by construction. As for (iii): B_r for π^* is the last term of an increasing finite sequence whose first term is B_r for π.

Corollary 2. *If* $\xi(1)/c(1) = \xi(2)/c(2) = \cdots = \xi(n)/c(n)$, *all right-of-way policies that insert no idle time are equivalent and optimal.*

Corollary 3. *Every policy that does not insert idle time can be improved, for every* TCT, *by a top-class policy that does not insert idle time.*

4. The M/GI/1 case. Proof of Theorem 2

By Theorem 1, the V.I.P. customers should be given priority right of way with no delay. Suppose all policies under consideration show this property, i.e., they are top-class policies. Let V be the service time of a non-V.I.P. customer, and denote by $\tilde{V} \geq V$ the time from the beginning of its service until its service has been

concluded and for the first time there are no V.I.P. customers in the station. As remarked in the proof of Theorem 1, $\mathbf{E}(\tilde{V})$ is proportional to $\mathbf{E}(V)$. By using the Poisson nature of the stream of V.I.P. customers and breakdowns, it follows through usual queuing methods, that all V.I.P. customers not included in the foregoing may be viewed as "breakdowns", and their $\tilde{V}$-type service as repair times. The composite breakdown-repair process is again of the kind allowed in Section 2.

Let there be given a top-class policy π. Construct a policy π^* as in the proof of Theorem 1 that (in the absence of V.I.P. customers, that are now part of the breakdown-repair process) serves the lowest type non-V.I.P. customer as soon as possible. By Theorem 1, the payoff of the non-V.I.P. customers under π^* is not greater than that under π. Moreover, every V.I.P. customer that is served by both policies spends under both policies the same amount of time in the station, and every V.I.P. customer served under π^* is also served under π since π^* cuts the idle-breakdown-repair scheme shorter than π. (This is the argument that led to (b) in Theorem 1.) So, the payoff of V.I.P. customers under π^* is also not greater than that under π. Hence, π^* improves π. Proceed by induction.

5. The average cost criterion

Assume throughout this section that the insertion of idle time is not allowed. See Corollary 3.

The average cost criterion is usually applied to systems that admit steady states. When considering general streams, stationary limiting distributions need not exist. However, whenever stationary ergodic situations do exist and busy periods are regenerative, minimizing expected total costs during a busy period (as we did in the preceding sections) is the same as minimizing the expected instantaneous rate of cost of the total system under steady state.

Still, we would like to say something about minimizing rates of cost even when steady states do not exist. The following theorem permits to extend the results of Theorem 1 to the criterion of long run minimal average cost.

Observe that when the insertion of idle times is not permitted, busy periods and idle times do not depend on service policy. This will justify the legitimacy of the assumption in the next theorem.

Theorem 3. *For any $T>0$, let M_T denote the number of arrivals up to time T during the busy period to which T belongs — or $M_T=0$ if T belongs to an idle period.*

Assume

$$\sup_{T>0} \mathbf{E}(M_T^2) = K < \infty.$$

Denote

$$c = \max_{1\le i\le n} c(i), \qquad \xi = \max_{1\le i\le n} \xi(i).$$

Let $\phi(\pi, T)$ be the expected total cost under π up to time T. Then for every policy π there exists a top-class policy π^* such that given $\varepsilon > 0$, whenever $T > KC\xi/\varepsilon$,

$$\frac{\phi(\pi^*, T)}{T} < \frac{\phi(\pi, T)}{T} + \varepsilon.$$

Proof. Consider any policy π. Observe that the top-class policy π^* defined in the proof of Theorem 1 is the same for all TCT in the sense that if $TCT_1 > TCT_2$, the π^* policy defined for TCT_1 agrees with the one defined for TCT_2 — at least until TCT_2. So, let π^* be as above. Use π up to time T, stop the stream at time T and complete the service of all customers present in the system at time T using, say a FIFO regime. The payoff of this policy, viewing T as TCT, exceeds $\phi(\pi^*, T)$ by Theorem 1. On the other hand it is less than $\phi(\pi, T) + KC\xi$, as can be easily checked. The result follows.

References

[1] K.R. Baker, Introduction to Sequencing and Scheduling (John Wiley, 1974).
[2] J.M. Harrison, Dynamic scheduling of a multi-class queue: Small interest rate, SIAM J. Appl. Math. (1975).
[3] M.M. Jaiswal, Priority Queues (Academic Press, 1968).
[4] G.P. Klimov, Time-sharing service systems, I, Theory of Probability and Its Applications, XIX (1974) 532–551.
[5] I. Meilijson and G. Weiss, Multiple feedback at a single server station, Stoch. Proc. Appl. 5 (1977) 195–205.
[6] R.W. Wolff, Work-conserving priorities, J. Appl. Prob. 7 (1970) 327–337.