STOCHASTIC SCHEDULING IN A MULTICLASS G/G/1 QUEUE

P. NAIN and D. TOWSLEY

COINS Technical Report 92-56
August 1992

STOCHASTIC SCHEDULING IN A MULTICLASS G/G/1 QUEUE

P. Nain* and D. Towsley†

August 25, 1992

Abstract

We address the problem of scheduling customers in a multiclass G/G/1 queue so as to minimize a weighted sum of the workloads of the different classes. We establish that the nonidling, preemptive, fixed priority policy that schedules customers belonging to the class having the maximum weight minimizes the cost function pathwise at any point in time. This result is based on the application of elementary forward induction arguments and is shown to hold for a very general class of policies. A new proof for the optimality of the μc-rule in the multiclass G/M/1 queue is then obtained as an easy corollary of the first result.

Keywords: Stochastic scheduling; Pathwise argument; Stochastic ordering; μc-rule; Optimal control of queues.

1 Introduction

Consider a G/G/1 queue with K classes of customers where the K arrival processes and K service processes may be arbitrary. We consider the problem of scheduling customers so as to minimize a weighted sum of the workloads of the different classes. We establish that the nonidling, preemptive, fixed priority policy that schedules customers belonging to the class having the maximum weight minimizes the cost function at any point in time pathwise. This result is based on the application of elementary forward induction arguments and is shown to hold for a very general class of policies. Last, the classical result (Baras et al. [4], Buyukkoc et al. [6], Nain [5]; see also Hirayama et al. [3] for further results on the multiclass G/DFR/1 queue that are not covered in the present paper) regarding the optimality of the μc rule for the G/M/1 queue is established as a simple consequence of this sample path property.

*INRIA, B.P. 109, 06565 Valbonne Cedex, France. This work was done when this author was visiting the University of Massachusetts, Amherst MA, USA, during the Spring of 1991.

†Department of Computer Science, University of Massachusetts, Amherst MA 01003, USA. This author was supported in part NSF under grants ASC-8802764 and NCR-9116183.

2 The Model

In this section we construct a model that captures the behavior of the multiclass G/G/1 queue loosely described in the introduction. An equivalent and somewhat more convenient way to view this queueing system is to assume that there are K queues attended by a single server and that customers of class i, $i = 1, 2, \ldots, K$, are routed to queue i upon arrival.

For notational convenience, we shall assume from now on that within each queue customers are served in the first-in-first-out order (see remark 2.1) and that the customer in position 1 in any queue is the oldest one among customers in that queue (i.e., the customer in position 1 is either the next eligible customer for service if the server is not attending the queue, or the customer in service if the server is serving that queue).

Let $\mathbb{N}$ be the set of all nonnegative integers and let $\mathbb{R} := (-\infty, +\infty)$, $\mathbb{R}_+ := [0, \infty)$ and $\overline{\mathbb{R}}_+ := [0, \infty]$. Define $\mathbf{S} := \{0\} \cup \{(x_1, \ldots, x_n), x_i > 0, i = 1, 2 \ldots, n, n \geq 1\}$ to be the set that contains all vectors with strictly positive components as well as the scalar number 0.

To describe this model, one starts with a probability triple $(\Omega, \mathcal{F}, P)$, where the state space Ω defined as

$$\Omega := \mathbb{N}^K \times \mathbf{S}^K \times \left\{\mathbb{R}_+^2 \times \{1, 2, \ldots, K\}\right\}^{\mathbb{N}} \times [0, 1]^{\mathbb{N}} \times [0, 1]^{\mathbb{N}}, \tag{2.1}$$

simultaneously carries

- an $\mathbb{N}^K$-valued random variable (RV) $Q := (Q_1, Q_2, \cdots, Q_K)$, where Q_i describes the number of customers in queue i at time $t = 0$;
- an $\mathbf{S}^K$-valued RV $W := (W_1, W_2, \ldots, W_K)$ with $W_i := (W_{i,1}, W_{i,2}, \ldots, W_{i,Q_i})$ if $Q_i > 0$ and with $W_i = 0$ if $Q_i = 0$, where $W_{i,j}$ describes the service requirement of customer in position j in queue i at time $t = 0$;
- a sequence $\{A_n, S_n, C_n\}_1^\infty$ of $\mathbb{R}_+^2 \times \{1, 2, \ldots, K\}$-valued RV's such that $0 < A_1 < A_2 < \cdots < A_n < A_{n+1} < \cdots$ a.s. and $S_n > 0$ a.s. for all $n \geq 1$, where A_n, S_n and C_n represent the arrival time, service requirement and class, respectively, of the n-th customer to join the system;
- two sequences of $[0, 1]$-valued RV's $\{\alpha_n\}_1^\infty$ and $\{\beta_n\}_1^\infty$. These sequences will be used to construct randomized scheduling policies.

In the following, any sample path $\omega \in \Omega$ will be written in the form

$$\omega = \left(\omega^1, \omega^2, \left\{\omega_{n,1}^3, \omega_{n,2}^3, \omega_{n,3}^3\right\}_1^\infty, \left\{\omega_n^4\right\}_1^\infty, \left\{\omega_n^5\right\}_1^\infty\right), \tag{2.2}$$

with $\omega^1 \in \mathbb{N}^K$, $\omega^2 \in \mathbf{S}^K$, $\omega_{n,1}^3, \omega_{n,2}^3 \in \mathbb{R}_+$, $\omega_{n,3}^3 \in \{1, 2, \ldots, K\}$, $\omega_n^4, \omega_n^5 \in [0, 1]$ for all $n \geq 1$.

Further notations are needed at this point. Let $\mathbf{H}_1 := \Omega$, $\mathbf{K}_1 := \Omega \times \{0,1,\ldots,K\}$, $\mathbf{H}_{n+1} := \mathbf{H}_n \times \{0,1,\ldots,K\} \times \overline{\mathbb{R}}_+^2 \times \mathbb{N}^K \times \mathbf{S}^K$, $\mathbf{K}_{n+1} := \mathbf{K}_n \times \overline{\mathbb{R}}_+^2 \times \mathbb{N}^K \times \mathbf{S}^K \times \{0,1,\ldots,K\}$ for $n \geq 2$. Any element $h_n \in \mathbf{H}_n$ will be written in the form

$$h_1 = \omega; \tag{2.3}$$

$$h_n = (\omega; u_1, i_1, t_1, q_2, v_2, u_2, i_2, t_2, \ldots, q_n, v_n), \; n \geq 2, \tag{2.4}$$

with $\omega \in \Omega$, $u_j \in \{0,1,\ldots,K\}$, $i_j, t_j \in\in \overline{\mathbb{R}}_+$ for $j \geq 1$ and $q_j := (q_j^1, q_j^2, \ldots, q_j^K) \in \mathbb{N}^K$, $v_j \in \mathbf{S}^K$ for all $j \geq 2$. Similarly, any element $k_n \in \mathbf{K}_n$ will be written in the form

$$k_1 = (\omega; u_1); \tag{2.5}$$

$$k_n = (\omega; u_1, i_1, t_1, q_2, v_2, u_2, i_2, t_2, \ldots, q_n, v_n, u_n), \; n \geq 2. \tag{2.6}$$

A scheduling policy π is defined to be any sequence $\{\pi_n^1, \pi_n^2\}_n$ of mappings

$$\pi_n^1 : \quad \mathbf{H}_n \to \{0,1,\ldots,K\};$$

$$\pi_n^2 : \quad \mathbf{K}_n \to \overline{\mathbb{R}}_+,$$

such that $\pi_n^1(h_n) \neq i$ if $q_n^i = 0$ for $1 \leq i \leq K$ and $\pi_n^1(h_n) = 0$ if $q_n = 0$, for all $n \geq 1$ (by convention $q_1 := \omega^1$). Let Π be the collection of all scheduling policies.

Let us briefly comment on the definition of a scheduling policy. Given the information h_n available at the n-th decision epoch (see below) to the decision-maker, $\pi_n^1(h_n)$ gives the class of customers that is elected to receive the server's attention until the next decision epoch if $\pi_n^1(h_n) \in \{1,2,\ldots,K\}$; if $\pi_n^1(h_n) = 0$, then the decision is to idle the server until the next decision epoch. The mapping π_n^2 is used to determine the time of the $(n+1)$-th decision (see below).

For every scheduling policy $\pi \in \Pi$, we generate five sequences $\{Q^\pi(t), t \geq 0\}$ $\{W^\pi(t), t \geq 0\}$, $\{U_n^\pi\}_1^\infty$, $\{T_n^\pi\}_1^\infty$ and $\{I_n^\pi\}_1^\infty$ of RV's such that for all $n \geq 1$, $t \geq 0$,

- $Q^\pi(t) := (Q_1^\pi(t), Q_2^\pi(t), \ldots, Q_K^\pi(t)) \in \mathbb{N}^K$, where $Q_i^\pi(t)$ gives the number of customers in queue i under policy π at time t, including the customer in service, if any, for all $i \in \{1,2,\ldots,K\}$;
- $W^\pi(t) := (W_1^\pi(t), W_2^\pi(t), \ldots, W_K^\pi(t)) \in \mathbf{S}^K$, where $W_i^\pi(t) := (W_{i,1}^\pi(t), W_{i,2}^\pi(t), \ldots, W_{i,Q_i^\pi(t)}^\pi(t))$ if $Q_i^\pi(t) > 0$ and $W_i^\pi(t) := 0$ if $Q_i^\pi(t) = 0$, $1 \leq i \leq K$, with the interpretation that $W_{i,j}^\pi(t)$ is the service requirement of the customer in position j in queue i under policy π at time t if $Q_i^\pi(t) > 0$ for all $j = 1,2,\ldots,Q_i^\pi(t)$;
- U_n^π gives the n-th action taken when policy π is used;
- T_n^π gives the occurrence time of the n-th decision when the policy π is used. We shall assume that $T_1^\pi = 0$ for all $\pi \in \Pi$ (i.e., the first decision is always made at time 0);

- I_n^π is used to generate the RV T_{n+1}^π (see below).

These RV's are recursively defined as follows:

$$
\begin{aligned}
U_1^\pi &:= \pi_1^1\Big(Q, W, \{A_m, S_m, C_m\}_1^\infty, \{\alpha_m\}_1^\infty, \{\beta_m\}_1^\infty\Big); && (2.7)\\
U_n^\pi &:= \pi_n^1\Big(Q, W, \{A_m, S_m, C_m\}_1^\infty, \{\alpha_m\}_1^\infty, \{\beta_m\}_1^\infty; \\
&\qquad U_1^\pi, I_1^\pi, T_2^\pi, Q^\pi(T_2^\pi), W^\pi(T_2^\pi), \ldots, U_{n-1}^\pi, I_{n-1}^\pi, T_n^\pi, Q^\pi(T_n^\pi), W^\pi(T_n^\pi)\Big), \quad n \geq 2; && (2.8)\\
T_1^\pi &:= 0;\\
T_{n+1}^\pi &:= \min\Big\{\inf\{A_m, m \geq 1 \,:\, A_m > T_n^\pi\},\\
&\qquad T_n^\pi + 1(U_n^\pi = 0)\, I_n^\pi + \sum_{i=1}^{K} 1(U_n^\pi = i) W_{i,1}^\pi(T_n^\pi), T_n^\pi + I_n^\pi\Big\}, \quad n \geq 1; && (2.9)\\
I_n^\pi &:= \pi_n^2\Big(Q, W, \{A_m, S_m, C_m\}_1^\infty \{\alpha_m\}_1^\infty, \{\beta_m\}_1^\infty; U_1^\pi, I_2^\pi, T_2^\pi,\\
&\qquad \ldots, Q^\pi(T_{n-1}^\pi), W^\pi(T_{n-1}^\pi), U_{n-1}^\pi, I_{n-1}^\pi, T_n^\pi, Q^\pi(T_n^\pi), W^\pi(T_n^\pi), U_n^\pi\Big), \quad n \geq 1. && (2.10)
\end{aligned}
$$

The $(n+1)$-th decision epoch occurs either at the time of an arrival, a service completion, or after I_n^π time units beyond the n-th decision epoch, whichever occurs first. Here I_n^π is the length of time that the scheduling policy allows the server to idle (if $U_n^\pi = 0$) or after which it may preempt the customer in service (if $U_n^\pi \in \{1, 2, \ldots, K\}$). This definition of the decision epochs will allow one to consider arbitrary (possibly randomized) preemptive and idling policies. Last, it is worth observing from the above definitions that scheduling policies that may know (in particular) future arrival times and future service times — usually referred to as anticipative policies — are also allowed here.

It remains to construct the queue-length process $\{Q^\pi(t),\, t \geq 0\}$ and the workload process $\{W^\pi(t),\, t \geq 0\}$. The RV $Q^\pi(t)$ is defined as follows:

$$
\begin{aligned}
Q^\pi(0) &:= Q;\\
Q_i^\pi(T_{n+1}^\pi) &:= Q_i^\pi(T_n^\pi) + \sum_{m \geq 1} 1\left((A_m, C_m) = (T_{n+1}^\pi, i)\right)\\
&\qquad -1\left(U_n^\pi = i, W_{i,1}^\pi(T_n^\pi) = T_{n+1}^\pi - T_n^\pi\right), \quad n \geq 1,\, i = 1, 2, \ldots, K; && (2.11)\\
Q^\pi(t) &:= \sum_{n \geq 1} Q^\pi(T_n^\pi)\, 1(T_n^\pi \leq t < T_{n+1}^\pi), \; t \geq 0. && (2.12)
\end{aligned}
$$

On the other hand, the RV $W^\pi(t)$ is defined as follows:

$$
\begin{aligned}
W^\pi(0) &:= W;\\
W_i^\pi(T_{n+1}^\pi) &:= \left(W_{i,1}^\pi(T_n^\pi) - 1(U_n^\pi = i)\, (T_{n+1}^\pi - T_n^\pi), W_{i,2}^\pi(T_n^\pi), \ldots, W_{i,Q_i^\pi(T_n^\pi)}^\pi(T_n^\pi),\right.
\end{aligned}
$$

$$\sum_{m\geq 1} S_m \mathbf{1}\left((A_m, C_m) = (T^\pi_{n+1}, i)\right)\Big), \quad n \geq 1, i = 1,2,\ldots,K; \tag{2.13}$$

$$\begin{aligned} W^\pi_i(t) \; &:= \; \Big(W^\pi_{i,1}(T^\pi_n) - \mathbf{1}(U^\pi_n = i)\,(t - T^\pi_n), W^\pi_{i,2}(T^\pi_n), \\ & \ldots, W^\pi_{i,Q^\pi_i(T^\pi_n)}(T^\pi_n)\Big) \;\text{ if }\; T^\pi_n \leq t < T^\pi_{n+1}, n \geq 1, i = 1,2,\ldots,K, \end{aligned} \tag{2.14}$$

for all $t \geq 0$, where (2.13) and (2.14) must read with the abuse of notation $(0, x_1, \ldots, x_k) = (x_1, \ldots, x_k, 0) = (0, x_1, \ldots, x_k, 0) = (x_1, \ldots, x_k)$ for all $k \geq 1$ and $(0) = (0,0) = 0$, so as to be consistent with the definition of the set $\mathbf{S}$.

Observe that, by construction, the sample paths of both the queue-length and the workload processes are right-continuous with left limits. It is also worth noticing from (2.12) and (2.14) that $Q^\pi(t)$ and $W^\pi(t)$ are well defined for all $t > 0$ if and only if the nondecreasing sequence $\{T^\pi_n\}_n$ of decision epochs satisfies

$$\lim_{n\to\infty} T^\pi_n = +\infty \text{ a.s.} \tag{2.15}$$

We conclude this section by commenting on the role of the sequences $\{\alpha_n\}_1^\infty$ and $\{\beta_n\}_1^\infty$. As already mentioned, these sequences may be used to generate randomized policies. For the sake of illustration, let us consider the following example.

Let π be a policy such that if all queues are non-empty at the n-th decision epoch, then the server is allocated to queue i with probability $p_{n,i}$ for all $i = 1,2,\ldots,K$ and is kept idle till the next decision epoch with probability $1 - \sum_{i=1}^K p_{n,i}$, $n \geq 1$ (observe that the above description only partially defines π since nothing is said as to the behavior of this policy when at least one queue is empty). Let us show how this behavior can be captured within the setting developed in this section.

Fix $\omega \in \Omega$ and assume that the sequence $\{\alpha_n\}_1^\infty$ is a renewal sequence of uniformly distributed RV's, further independent of the RV's Q, W, $\{A_n, S_n, C_n\}_1^\infty$ and $\{\beta_n\}_1^\infty$. Then, it suffices to set

$$\pi_n(h_n) = \begin{cases} i, & \text{if } \sum_{j=1}^{i-1} p_{n,j} \leq \omega_n^4 < \sum_{j=1}^{i} p_{n,j}; \\ 0, & \text{if } 1 - \sum_{i=1}^K p_{n,i} \leq w_n^4 \leq 1, \end{cases} \tag{2.16}$$

for all $h_n \in \mathbf{H}_n$ so as to reflect the (partial) behavior of the policy π. Indeed, by construction of the RV U^π_n (see (2.7)-(2.8)) it is seen that for $i = 1,2,\ldots,K$

$$\begin{aligned} & P\left(U^\pi_n = i \,|\, Q^\pi_j(T^\pi_n) > 0, j = 1,2,\ldots,K\right) \\ = \;& P\left(\pi^1_n(H_n) = i \,|\, Q^\pi_j(T^\pi_n) > 0, j = 1,2,\ldots,K\right), \\ = \;& P\left(\sum_{j=1}^{i-1} p_{n,j} \leq \alpha_n < \sum_{j=1}^{i} p_{n,j}\right), \quad \text{from (2.16)} \\ = \;& p_{n,i}, \end{aligned} \tag{2.17}$$

where in (2.17) the RV H_n denotes the argument of the mapping π^1_n in (2.7)-(2.8). Similarly, it is seen that $P\left(U^\pi_n = 0 \,|\, Q^\pi_j(T^\pi_n) > 0, j = 1,2,\ldots,K\right) = 1 - \sum_{i=1}^K p_{n,i}$.

The sequence $\{\beta_n\}_1^\infty$ may be used in the definition of the mappings $\{\pi_n^2\}_1^\infty$ to construct random idle periods (see (2.9), (2.10)).

Remark 2.1 The assumption that the order of service within each queue is first-in-first-out is only used in the construction of the queue length process (see (2.11)-(2.12)) and of the workload process (see (2.13)-(2.14)). In particular, it will not affect the generality of the results in sections 3 and 4 since only the total workload in each queue is considered in these sections. If one wants to relax this assumption, then the scheduling policy must also specify which customer should be served in the queue (if any) that has been elected to receive the server's attention. This can be achieved, for instance, by introducing a third component, π_n^3, in the definition of a scheduling policy π_n for all $n \geq 1$.

3 Scheduling in the G/G/1 Queue

In this section we consider a cost function corresponding to the weighted sum of the workloads of the different classes. We show that the preemptive fixed priority policy that assigns priority in decreasing order of the weights minimizes the cost function pathwise at every point in time.

Let $\gamma := \{\gamma_n^1, \gamma_n^2\}_1^\infty \in \Pi$ be the nonidling and preemptive policy that always allocates the server to class i customers when there are no longer class $j < i$ customers in the system, $i = 1, 2, \ldots, K$. In other words, $\gamma_n^1(h_n) = i$ for all $h_n \in \mathbf{H}_n$ such that $q_{n,j} = 0$ for $j = 1, 2, \ldots, i-1$ and $q_{n,i} \neq 0$, for all $i = 1, 2, \ldots, K$, $n \geq 1$. Let

$$V_i^\pi(t) := \sum_{j=1}^{Q_i^\pi(t)} W_{i,j}^\pi(t),$$

be the total workload due to class i customers at time $t \geq 0$, $i = 1, 2, \ldots, K$.

Let r_i, $i = 1, 2, \ldots, K$ be given real numbers such that $r_1 \geq r_2 \geq \cdots \geq r_K \geq 0$. We shall show the following result:

Proposition 3.1 *Assume that condition (2.15) holds. Then, for every sample path $\omega \in \Omega$,*

$$\sum_{i=1}^{k} r_i V_i^\gamma(t) \leq \sum_{i=1}^{k} r_i V_i^\pi(t), \tag{3.1}$$

for all $k = 1, 2, \ldots, K$, $t \geq 0$, $\pi \in \Pi$.

Recall that a real-valued RV X is smaller than a real-valued RV Y in the sense of stochastic ordering (written $X \leq_{st} Y$) if $E[f(X)] \leq E[f(Y)]$ for all nondecreasing mappings $f : \mathbb{R} \to \mathbb{R}$ such that the expectations exist. Proposition 3.1 yields the following result:

Corollary 3.1 *For all $t \geq 0$, $\pi \in \Pi$,*

$$\sum_{i=1}^{K} r_i V_i^{\gamma}(t) \leq_{st} \sum_{i=1}^{K} r_i V_i^{\pi}(t).$$

Proposition 3.1 follows from the following two lemmas:

Lemma 3.1 *Let $N > 0$ be an arbitrary integer and let $(X_1, \ldots, X_N) \in \mathbb{R}^N$ and $(Y_1, \ldots, Y_N) \in \mathbb{R}^N$ be two vectors such that $\sum_{i=1}^{n} X_i \leq \sum_{i=1}^{n} Y_i$ for all $n = 1, 2 \ldots, N$. Then,*

$$\sum_{i=1}^{N} c_i X_i \leq \sum_{i=1}^{N} c_i Y_i, \tag{3.2}$$

for any sequence $\{c_i\}_{i=1}^{N}$ such that $c_1 \geq c_2 \geq \cdots \geq c_N \geq 0$.

Proof. The proof is by induction in N. Inequality (3.2) is trivially true when $N = 1$. Assume that it is true for $N = 1, 2, \ldots, m-1$ and let us show that it is still true for $N = m$.

We have

$$\sum_{i=1}^{m} (Y_i - X_i) c_i = \sum_{i=1}^{m-1} (Y_i - X_i)(c_i - c_m) + c_m \sum_{i=1}^{m} (Y_i - X_i),$$

which is nonnegative from the induction hypothesis, which concludes the proof. ∎

Lemma 3.2 *Assume that (2.15) holds. Then, for every sample path $\omega \in \Omega$,*

$$\sum_{i=1}^{k} V_i^{\gamma}(t) \leq \sum_{i=1}^{k} V_i^{\pi}(t), \tag{3.3}$$

for all $k = 1, 2, \ldots, K$, $t \geq 0$, $\pi \in \Pi$.

Proof. Let π be an arbitrary policy in Π.

Let $\{t_n\}_1^{\infty}$, $0 = t_1 < t_2 < \cdots$, be the sequence resulting from the superposition of both sequences $\{T_n^{\pi}\}_1^{\infty}$ and $\{T_n^{\gamma}\}_1^{\infty}$. The proof is by induction on the times of events $t_1 < t_2 < \cdots < t_n < t_{n+1} < \cdots$.

Basis step. Trivially true for $t = 0$ (since by definition of the model $V_i^{\gamma}(0) = V_i^{\pi}(0)$ for all $i = 1, 2, \ldots, K$).

Induction step. Assume that the (3.3) holds for $0 < t \leq t_n$ and let us show that it is still true for $t_n < t \leq t_{n+1}$. There are two steps.

Step 1: $t_n < t < t_{n+1}$.

If $\sum_{i=1}^{K} V_i^\gamma(t_n) = 0$ then (3.3) clearly holds for $t_n < t < t_{n+1}$. Consider the case that $\sum_{i=1}^{K} U_i^\gamma(t_n) > 0$. By the definition of γ there exists an $l \in \{1, 2, \ldots, K\}$ such that

$$(V_1^\gamma(t), \ldots, V_K^\gamma(t)) = \left(0, \ldots, 0, V_l^\gamma(t) - (t - t_n), V_{l+1}^\gamma(t_n), \ldots, V_K^\gamma(t_n)\right). \tag{3.4}$$

For $1 \le k \le l - 1$, it is seen from (3.4) that

$$0 = \sum_{i=1}^{k} V_i^\gamma(t) \le \sum_{i=1}^{k} V_i^\pi(t).$$

On the other hand, we have for $l \le k \le K$, cf. (3.4),

$$\begin{aligned} \sum_{i=1}^{k} V_i^\gamma(t) &= \sum_{i=1}^{k} V_i^\gamma(t_n) - (t - t_n), \\ &\le \sum_{i=1}^{k} V_i^\pi(t_n) - (t - t_n), & (3.5) \\ &\le \sum_{i=1}^{k} V_i^\pi(t). & (3.6) \end{aligned}$$

Inequality (3.5) follows from the induction hypothesis. Equality takes place in (3.6) if and only if the server does not idle in (t_n, t_{n+1}) under π and is allocated to a customer from one of the classes $1, 2, \ldots, k$ during this period of time.

Step 2: $t = t_{n+1}$.

Consider different events. If t_{n+1} is not an arrival epoch, then $V_i^\gamma(t_{n+1}) = V_i^\gamma(t_{n+1}^-)$ and $V_i^\pi(t_{n+1}) = V_i^\pi(t_{n+1}^-)$ for all $i = 1, 2, \ldots, K$. Inequality (3.3) at time t_{n+1} then follows from step 1.

If t_{n+1} is an arrival epoch, then clearly

$$\begin{aligned} V_i^\gamma(t_{n+1}) &= V_i^\gamma(t_{n+1}^-) + \sum_{m \ge 1} S_m \, 1(A_m = t_{n+1}, C_m = i); \\ V_i^\pi(t_{n+1}) &= V_i^\pi(t_{n+1}^-) + \sum_{m \ge 1} S_m \, 1(A_m = t_{n+1}, C_m = i), \end{aligned}$$

for $i = 1, 2, \ldots, K$. Again, inequality (3.3) at time t_{n+1} follows from step 1, which concludes the proof. ∎

4 Optimality of the μc-Rule

In this section we establish the optimality of the μc rule for the G/M/1 queue as a simple consequence of Corollary 3.1.

Let S_n^i denote the service requirement of the n-th customer of class i, $n \geq 1$, $i = 1, 2, \ldots, K$. Observe that $S_n^i = \sum_{k \geq 1} S_k \mathbf{1}\left(C_k = i, \sum_{l=1}^{k-1} 1(C_l = i) = n - 1\right)$. We shall assume throughout this section that

A1 The sequences $\{S_n^1\}_1^\infty, \ldots, \{S_n^K\}_1^\infty$ form K mutually independent renewal sequences, further independent of the arrival sequence $\{A_n, C_n\}_1^\infty$;

A2 $P(S_n^i \leq x) = 1 - e^{-\mu_i x}$ for all $x \geq 0$, $n \geq 1$, $i = 1, 2, \ldots, K$.

Without loss of generality, we shall also assume that the system is empty at time 0 (i.e., $Q = 0$ and $W = 0$ a.s.).

Let $\Pi^* \subset \Pi$ be the set of all scheduling policies that do not know future service times of the customers. Formally speaking, this means that for any policy $\pi \in \Pi^*$ there exist two collections of mappings $\{f_n^1\}_1^\infty$ and $\{f_n^2\}_1^\infty$

$$\begin{aligned} f_n^1 : \quad & \mathbf{H}_n^* \to \{0, 1, \ldots, K\}; \\ f_n^2 : \quad & \mathbf{K}_n^* \to \overline{\mathbb{R}}_+, \end{aligned}$$

where $\Omega^* := \mathbb{N}^K \times \{\mathbb{R}_+ \times \{1, 2, \ldots, K\}\}^{\mathbb{N}} \times [0,1]^{\mathbb{N}} \times [0,1]^{\mathbb{N}}$, $\mathbf{H}_1^* := \Omega^*$, $\mathbf{H}_{n+1}^* := \mathbf{H}_n^* \times \{0, 1, \ldots, K\} \times \overline{\mathbb{R}}_+^2 \times \mathbb{N}^K$, $\mathbf{K}_1^* := \Omega^* \times \{0, 1, \ldots, K\}$, $\mathbf{K}_{n+1}^* := \mathbf{K}_n^* \times \overline{\mathbb{R}}_+^2 \times \mathbb{N}^K \times \{0, 1, \ldots, K\}$, such that

$$\begin{aligned} \pi_1^1(h_1) &= f_1^1(\omega^*); \\ \pi_n^1(h_n) &= f_n^1(\omega^*; u_1, i_1, t_1, q_2, u_2, i_2, t_2, \ldots, q_n), \quad n \geq 2; \\ \pi_1^2(k_1) &= f_1^1(\omega^*, u_1); \\ \pi_n^2(k_n) &= f_n^2(\omega^*; u_1, i_1, t_1, q_2, u_2, i_2, t_2, \ldots, q_n, u_n), \quad n \geq 2, \end{aligned}$$

for all $h_n \in \mathbf{H}_n$ (cf. (2.3), (2.4)), $k_n \in \mathbf{K}_n$ (cf. (2.5), (2.6)), $\omega \in \Omega$ (cf. (2.2)), where

$$\omega^* := \left(\omega^1, \left\{\omega_{n,1}^3, \omega_{n,3}^3\right\}_1^\infty, \left\{\omega_n^4\right\}_1^\infty, \left\{\omega_n^5\right\}_1^\infty\right).$$

The following lemma holds:

Lemma 4.1 *Assume that* **A1** *and* **A2** *holds. Then, for every* $t \geq 0$, $i = 1, 2, \ldots, K$, $\pi \in \Pi^*$,

$$E[Q_i^\pi(t)] = \mu_i E[V_i^\pi(t)]. \tag{4.1}$$

Proof. Fix $t \geq 0$, $i \in \{1, 2, \ldots, K\}$ and $\pi \in \Pi^*$.

Let $N_i := \{N_i(t), t \geq 0\}$ be a Poisson process with intensity μ_i, where $N_i(t)$ denotes the number of jumps in $[0, t]$. We assume that N_i is independent of the RV's $\{A_n, C_n, S_n, \alpha_n, \beta_n\}_n$. Because of assumptions **A1** and **A2** and because the policy π does not know future service times, it is seen that

$$Q_i^\pi(t) = A_i(t) - \int_0^t \mathbf{1}(S^\pi(s) = i)\, dN_i(s) \quad \text{a.s.,} \tag{4.2}$$

where $A_i(s) := \sum_{n\geq 1} \mathbf{1}(A_n \leq s, C_n = i)$ gives the number of class i arrivals in $[0, s]$, and where

$$S^\pi(s) := \sum_{n\geq 1} U_n^\pi\, \mathbf{1}(T_n^\pi \leq s < T_{n+1}^\pi) \tag{4.3}$$

reports the state of the server at time s. In other words, the Poisson process N_i may be seen as the virtual departure process of queue i in the sense that if a jump occurs in N_i (say at time t) while the server is serving queue i, then a departure will occur in queue i at time t; otherwise, no departure will occur in queue i.

Define $\mathcal{F}_i^\pi(t)$ to be the σ-field generated by the RV's $\{N_i(s), S^\pi(s)\, 0 \leq s \leq t\}$. Let us assume that the Poisson process $N_i(t)$ has the $\mathcal{F}_i^\pi(t)$-intensity μ_i for all $t \geq 0$, that is (Brémaud, [1])

$$E[N_i(t) - N_i(s) \,|\, \mathcal{F}_i^\pi(s)] = \mu_i(t - s), \tag{4.4}$$

for all $0 \leq s \leq t$.

Then, since $S^\pi(t)$ is $\mathcal{F}_i^\pi(t)$-adapted and left-continuous (cf. (4.3)), it follows from Brémaud [1, T5, Chapter 1]) that $S^\pi(t)$ is $\mathcal{F}_i^\pi(t)$- predictable, which, in turn, implies that formula (2.3) in Brémaud [1, p. 24] applies to yield

$$E\left[\int_0^t \mathbf{1}(S^\pi(s) = i)\, dN_i(s)\right] = \mu_i E\left[\int_0^t \mathbf{1}(S^\pi(s) = i)\, ds\right]. \tag{4.5}$$

Combining (4.2) and (4.5) gives

$$E\left[Q_i^\pi(t)\right] = E[A_i(t)] - \mu_i E\left[\int_0^t \mathbf{1}(S^\pi(s) = i)\, ds\right]. \tag{4.6}$$

On the other hand, we have

$$\begin{aligned} E[V_i^\pi(t)] &= E\left[\sum_{n=1}^{A_i(t)} S_n^i\right] - E\left[\int_0^t \mathbf{1}(S^\pi(s) = i)\, ds\right], \\ &= \mu_i^{-1}\, E[A_i(t)] - E\left[\int_0^t \mathbf{1}(S^\pi(s) = i)\, ds\right], \end{aligned} \tag{4.7}$$

where (4.7) follows from Wald's identity (which applies here since the arrival process and the service time process for customers of class i are independent). Combining (4.6) and (4.7) yields formula (4.1).

It remains to show that (4.4) holds for all $0 \leq s \leq t$. Because the service times are mutually independent, exponential and independent of the RV's $\{A_n, C_n, \alpha_n, \beta_n\}_n$ and because the policy π

does not depend on future service times, it follows from (2.7)-(2.8) and (4.3) that $N_i(t) - N_i(s)$ is independent of $S^\pi(u)$ for all $0 \leq u \leq s \leq t$. Therefore,

$$\begin{aligned} E[N_i(t) - N_i(s) \,|\, \mathcal{F}_i^\pi(s)] &= E[N_i(t) - N_i(s) \,|\, \sigma(N_i(u)\,, u \leq s)], \\ &= \mu_i(t-s), \end{aligned}$$

for all $0 \leq s \leq t$, which completes the proof. ∎

We now turn to the main result of this section. Let $\{c_i\}_1^K$ be nonnegative constants. Up to a renumbering of the classes, we may assume that $\mu_i c_i \geq \mu_{i+1} c_{i+1}$ for $i = 1, 2, \ldots, K-1$. Define $\delta \in \Pi^*$ to be the nonidling policy that gives preemptive priority to class i customers over class j customers if $i < j$, $i, j = 1, 2, \ldots, K$. In other words, policy $\delta := \{\delta_n^1, \delta_n^2\}_1^\infty$ is such that $\delta_n^1(h_n^*) = i$ for all $h_n^* \in \mathbf{H}_n^*$ such that $q_{n,j} = 0$ for $j = 1, 2, \ldots, i-1$ and $q_{n,i} > 0$, $i = 1, 2, \ldots, K$, $n \geq 1$. As long as (2.15) holds, the mappings δ_n^2, $n \geq 1$, are arbitrary since δ is not allowed to idle.

The following proposition holds:

Proposition 4.1 *Assume* **A1** *and* **A2** *hold. Then, for every $t \geq 0$,*

$$\sum_{i=1}^K c_i\, E[Q_i^\delta(t)] \leq \sum_{i=1}^K c_i\, E[Q_i^\pi(t)],$$

for all $\pi \in \Pi^*$ such that (2.15) holds.

Proof. The proof follows from Corollary 3.1 by letting $r_i := \mu_i c_i$ for $i = 1, 2, \ldots, K$ and by using Lemma 4.1. ∎

Proposition 4.1 says that the μc-rule is optimal out of the policies that may know future arrival times but not future service times. This result can be seen as the continuous-time analog of the result in Baras et al. [4] and in Buyukkoc et al. [6] (see remark (4.2)).

Remark 4.1 Because the service times are exponentially distributed, it is seen that condition (2.15) is satisfied, in particular, if there is a finite number of arrivals in any finite interval of time (i.e., the arrival process is non-explosive, see Brémaud [1]) and if $\sum_{n\geq 1} I_n^\pi = \infty$ a.s. for all $\pi \in \Pi^*$.

Remark 4.2 The discrete-time version of the problem (see Baras et al. [4], Buyukkoc et al. [6]) can be addressed using the same approach. In the discrete-time setting, we assume that the service times are geometrically distributed with queue dependent parameter μ_i, $1 \leq i \leq K$. Given that a decision is made at every time $t \in \mathbb{N}$, the objective is to find a policy $\pi \in \Pi^*$ that minimizes $E[\sum_{i=1}^k c_i\, Q_i^\pi(t)]$ for all $t \in \mathbb{N}$, $k = 1, 2, \ldots, K$. Fix $\pi \in \Pi^*$, $t \in \mathbb{N}$, $1 \leq i \leq K$. It is seen that

$$E[Q_i^\pi(t)] = E[A_i(t)] - \sum_{s=1}^t E[S^\pi(s-1) = i, B_i(s) = 1], \tag{4.8}$$

where $\{B_i(s)\}_1^\infty$ is a Bernoulli sequence of RV's with parameter μ_j, independent of the RV's $\{A_n, C_n, S_n, \alpha_n, \beta_n\}_1^\infty$. The sequence $\{B_i(s)\}_1^\infty$ characterizes the virtual departure process of queue i and is the continuous-time analog of the Poisson process N_i introduced in the proof of Lemma 4.1. Because the policy π does not know future service times, we observe that the RV's $S^\pi(s-1)$ and $B_i(s)$ are independent for all $s = 1, 2, \ldots, t$. Therefore, cf. (4.8),

$$\begin{aligned} E[Q_i^\pi(t)] &= E[A_i(t)] - \mu_i \sum_{s=1}^{t} E[S^\pi(s-1) = i], \\ &= \mu_i \, E[V_i^\pi(t)]. \end{aligned}$$

The proof that the μc-rule is optimal again follows from Corollary 3.1.

References

[1] P. Brémaud, *Point Processes and Queues. Martingale Dynamics*, Springer Verlag, New York, 1980.

[2] J. M. Harrison, "Dynamic scheduling of a multiclass queue: Discount optimality," *Operations Res.* **23** (1975) 270-282.

[3] T. Hirayama, M. Kijima and S. Nishimura, "Further results for dynamic scheduling of multiclass G/G/1 queues," *J. Appl. Prob.* **26** (1989) 595-603.

[4] J. S. Baras, D.-J. Ma and A. M. Makowski, "K competing queues with geometric requirements and linear costs: the μc-rule is always optimal," *Systems Control Lett.* **6** (1985) 173-180.

[5] P. Nain, "Interchange arguments for classical scheduling problems in queues," *Systems Control Lett.* **12** (1989) 177-184.

[6] C. Buyukkoc, P. Varaiya and J. Walrand, "The μc-rule revisited," *Adv. Appl. Prob.* **17** (1985) 237-238.