Optimal Scheduling of Parallel Queues with Stochastic Flow Models: The $c\mu$-rule Revisited

A. Kebarighotbi $^*$ C. G. Cassandras $^*$

$^*$ Division of Systems Engineering, Center for Information and Systems Engineering, Boston University, Brookline, MA 02446

alik@bu.edu, cgc@bu.edu

Abstract: We consider a classic scheduling problem for optimally allocating a resource to multiple competing users and place it in the framework of Stochastic Flow Models (SFMs). We derive Infinitesimal Perturbation Analysis (IPA) gradient estimators for the average holding cost with respect to resource allocation parameters. These estimators are easily obtained from a sample path of the system without any knowledge of the underlying stochastic process characteristics. Exploiting monotonicity properties of these IPA estimators, we prove the optimality of the well-known $c\mu$-rule not only for the two-queue case (as in earlier work) but for an arbitrary finite number of queues and stochastic processes under non-idling policies.

Keywords: Hybrid Systems, Discrete-Event Systems, Stochastic Flow Models, Perturbation Analysis, Scheduling algorithms.

1. INTRODUCTION

The classic prototypical stochastic scheduling problem involves a single resource whose service capacity is to be optimally shared by $N$ competing users. In a queueing theory framework, this problem is modeled as a system of $N$ parallel queues, each with its own arrival process, connected to a single server. The server processes tasks from the $n$th queue with rate $\mu_n$, $n = 1, \ldots, N$, and uses a policy to select the next queue to serve from. Each task requires a random amount of time to be processed. The server may preempt a task by interrupting its processing to serve a new task from some other queue. This model applies to a large group of applications in communication networks, manufacturing, and computer processing.

The usual objective in this scheduling problem is to minimize the overall average holding cost of the tasks in the system with $c_n$ denoting the cost per unit waiting time in the $n$th queue. When the holding cost is a linear combination of the number of tasks in the competing queues, the well-known $c\mu$-rule has been shown, under certain conditions, to give the optimal allocation sequence. Following this rule, the queues are ordered according to the value of the product $c_n\mu_n$, from largest to smallest, and the server always selects a task from the first non-empty queue with the largest $c_n\mu_n$ value. The $c\mu$-rule is very attractive in that it is essentially static, except for the knowledge of whether a queue is empty or not. Thus, establishing its optimality in the most general possible setting is a goal that has been actively pursued through many years. The optimality of the $c\mu$-rule seems to have been first suggested in Smith (1956) under a deterministic and static setting, i.e., all tasks are present at time 0 with fixed service times. Relaxing these assumptions, Cox and Smith (1961) later proved the optimality of the $c\mu$-rule for a multi-class $M/G/1$ system. Using classical queueing models in a discrete time setting, the $c\mu$-rule was shown to be optimal for general arrival processes and geometrically distributed service times in Baras et al. (1985) and Buyukkoc et al. (1985). There have since been various attempts to extend these results. For example, it is shown in Hirayama et al. (1989) that for a discrete time $G/G/1$ model with a non-idling and non-preemptive server with decreasing failure rate the $c\mu$-rule is still optimal. Along a different direction, the scheduling problem above has been studied using a fluid flow abstraction in both a deterministic context Avram et al. (1995), Chen and Yao (1993) and a stochastic setting where the optimality of the $c\mu$-rule can be obtained using the heavy traffic (fluid limit) arguments Kingman (1961), Whitt (1968), Harrison (1986). A “generalized” $c\mu$-rule can then be shown to be asymptotically optimal Mieghem (1995) not only for the linear but also for convex cost objectives.

In Kebarighotbi and Cassandras (2009), we studied the basic stochastic scheduling problem above using a Stochastic Fluid (or Flow) Model (SFM). Using this model, it was shown that the $c\mu$-rule is optimal in the two-queue case, extending previous results in the literature to a broader class of stochastic processes and without any heavy traffic conditions. Further, when the cost is nonlinear in the queue contents, it was shown that the $c\mu$-rule may no longer be optimal. Nevertheless, the techniques used were based on easily obtainable gradient estimators that can be used to find an optimal allocation policy.
Unlike a deterministic fluid model or a stochastic model that makes use of heavy traffic assumptions, an SFM treats the arrival and service rates as stochastic processes of arbitrary generality (except for mild technical conditions), even under light traffic. The value of SFMs (introduced in Cassandras et al. (2002)) lies not in deriving approximations of performance measures of the underlying discrete event system, but rather in studying sample paths from which one can derive structural properties and optimal policies by making use of Infinitesimal Perturbation Analysis (IPA). IPA provides derivative estimates of performance measures (e.g., workload, loss) with respect to controllable parameters. These estimates are independent of the probability laws of the stochastic rate processes, therefore the actual values of these processes never enter the estimators, except (on occasion) for instantaneous rates at certain observable event times. Moreover, they require minimal information from the observed sample path. These properties, including the unbiasedness of the estimators, were established in Cassandras et al. (2002) for queueing systems with finite capacity and extended to serial networks (Sun et al. (2004a)), systems with feedback control mechanisms (Yu and Cassandras (2006)), and some multi-class models (Sun et al. (2004b); Cassandras et al. (2003); Panayiotou (2004)).

This paper extends the results in Kebarighotbi and Cassandras (2009) by establishing the optimality of the cp-rule to \( N > 2 \) queues (which has proved to be a somewhat surprising challenge.) Its contributions include the following. First, because our results are independent of the stochastic nature of the arrival and service rate processes, they provide evidence of the generality of the cp-rule as an optimal policy. Second, the analysis is based on explicit sample state and event time derivatives which can, therefore, be used to determine optimal schedules for different cost structures and be extended to more complex scheduling problems (e.g., systems with loss due to finite capacities) where the cp-rule no longer applies. We should point out, however, that our analysis relies on showing that perturbing away from the cp-rule policy increases the average linear holding cost, thus, it is possible that there exist other optimal policies. This is not surprising and it arises in other proofs of the cp-rule as well.

In Section 2, the basic scheduling problem is formulated in a SFM setting. Section 3 includes the proof of optimality of the cp-rule using a different approach than that of Kebarighotbi and Cassandras (2009) which allows the generalization to \( N > 2 \) queues. We conclude with Section 4 and discuss future research directions.

2. PROBLEM FORMULATION

Consider a SFM comprised of \( N \) queues competing for a shared resource as shown in Fig. 1. We will be studying this system over a finite time interval \([0, T]\). User requests from different classes \( n = 1, \ldots, N \) are abstracted into uncontrollable inflows \( \{\alpha_n(t)\} \) capturing the instantaneous rate of arriving tasks and treated as random processes. The associated fluid content random processes are denoted by \( \{x_n(t)\} \) with \( x_n(t) \geq 0 \). At each time \( t \), the rate at which the resource is processing the fluid from the queue \( n \) is denoted by \( u_n(t) \in [0, \mu_n] \). Here, \( \mu_n > 0 \) denotes the maximum processing rate for queue \( n \). The processing rates are subject to the capacity constraint

\[
\sum_{n=1}^{N} u_n(t) \leq \mu_n, \quad \forall t \in [0, T].
\]

The outflow rate from the resource is denoted by \( \{\beta(t)\} \) and is defined as \( \beta(t) = \sum_{n=1}^{N} u_n(t) \) for all \( t \). All random processes in the SFM are defined on a common probability space \((\Omega, \mathcal{F}, P)\). We define control functions \( \theta_n(t) \in [0, 1] \)

![Fig. 1. Stochastic Flow Model (SFM) for the scheduling problem.](image)

to represent the maximum fraction of \( \mu_n \) at which the resource processes the fluid from the \( n \)th queue. Therefore, we have \( u_n(t) \leq \mu_n \theta_n(t) \) for all \( t \).

Viewed as a stochastic hybrid system, each queue can only be in one of two discrete states: either \( x_n(t) = 0 \) over some Empty Period (EP) or \( x_n(t) > 0 \) over some Non-Empty Period (NEP). We assume that all the EPs and NEPs are close-left and open-right intervals. Given a sample path over \([0, T]\), we define \( \Sigma_n \) to be the set of all NEP start times and \( \Gamma_n \) to be the set of all NEP end times for queue \( n = 1, \ldots, N \). Regarding the controllability of event times in the SFM, they can be divided into two categories: (i) Exogenous events which are due to an uncontrollable discontinuity in some inflow rate and are assumed locally independent of \( \theta_n(t) \), and (ii) Endogenous events whose occurrence times can be controlled by \( \theta_n(t) \) for some \( n \). There are two further cases possible for an exogenous event: (a) an event that changes the discrete state of queue \( n \) by initiating a NEP \( \text{and} \) (b) one due to a possible discontinuity in \( \alpha_n(t) \) which leaves the discrete state intact. We denote the set of all event times in category (b) by \( \Lambda_n \). Also, there are two possibilities for an endogenous event: (a) one that ends a NEP at some queue \( n \) and (b) one that starts a NEP at some time \( t \) but is not caused by any discontinuity in the inflow rates. We make the following mild technical assumptions.

**Assumption 1.** With probability 1, no two events can occur at the same time unless the occurrence of one causes that of the other.

**Assumption 2.** The inflow processes \( \{\alpha_n(t)\} \) are piecewise continuous in \([0, T]\), with a finite number of discontinuities.

The queue contents \( x_n(t; \theta_n(t)) \), \( n = 1, \ldots, N \) evolve according to the one-sided differential equations

\[
\frac{dx_n(t; \theta_n(t))}{dt^+} = f_n(x_n(t; \theta_n(t))) = \alpha_n(t) - u_n(t; \theta_n(t))
\]

where, according to the definition of the rate processes, we can write for any \( n = 1, \ldots, N \) and \( t \in [0, T] \):

\[
u_n(t; \theta_n(t)) = \begin{cases} \alpha_n(t) & \text{if } x_n(t; \theta_n(t)) = 0 \\ \mu_n \theta_n(t) & \text{otherwise} \end{cases}
\]

8224
Below, we drop \( \theta_n(t) \) from the arguments of the functions \( u_n(t; \theta_n(t)), x_n(t; \theta_n(t)) \) and \( f_n(t; \theta_n(t)) \) to keep the notation manageable. Notice that by (2) and (3), we can write
\[
f_n(t) = \begin{cases} 
0 & \text{if } x_n(t; \theta_n(t)) = 0 \\
\alpha_n(t) - \mu_n \theta_n(t) & \text{otherwise}
\end{cases}
\]  
(4)

Let us consider a sample path \( \omega \in \Omega \) generated under some fixed functions \( \theta_n(t), n = 1, \ldots, N \). We define 0 = \( t_0 < t_1 < \ldots < t_M = T \) to be the occurrence times of all events that either start or end NEPs over all queues in the interval \((0, T)\) with the addition of the points 0 and T. Notice that \( M \) generally depends on the functions \( \theta_n(t) \). We further define:
\[
\theta_{n,m}(t) = \theta_n(t), \quad t \in [t_m, t_{m+1}), \quad m = 0, \ldots, M - 1.
\]  
(5)

Thus, in this case we can write \( \theta_{N,m}(t) = 1 - \sum_{n=1}^{N-1} \theta_{n,m}(t) \) and define the controllable vector:
\[
\theta(t) = [\theta_1(t), \ldots, \theta_{N-1}(t)]
\]
with \( v_m(t) = [\theta_1(t), \ldots, \theta_{N-1}(t)] \).

Henceforth, we omit \( \omega \) from the arguments of all processes understanding that our full analysis is carried out over some arbitrary but fixed sample path associated with it.

### 3. INFINITESIMAL PERTURBATION ANALYSIS

Let us rewrite (6) as
\[
Q(\theta) = \int_0^T \sum_{n=1}^{N} c_n x_n(t; \theta(t), \omega) \, dt. 
\]  
(6)

Differentiating with respect to \( \theta_{i,m}(t) \) for any pair of indices \( i = 1, \ldots, N - 1 \) and \( m = 0, \ldots, M - 1 \) gives
\[
\frac{\partial Q(\theta)}{\partial \theta_{i,m}} = \sum_{n=1}^{N} \int_{t_m}^{t_{m+1}} \frac{\partial x_n(t; \theta(t), \omega)}{\partial \theta_{i,m}} \, dt. 
\]  
(10)

By (4) and (5) the control \( \theta_{i,m}(t) \) can only affect the queue content evolutions for \( t \geq t_m \), so we conclude that
\[
\frac{\partial x_n(t; \theta(t), \omega)}{\partial \theta_{i,m}} = 0 \quad \forall t < t_m. 
\]  
(11)

However, the IPA derivative (10) still requires evaluating the derivatives \( \frac{\partial x_n(t; \theta(t), \omega)}{\partial \theta_{i,m}} \) for \( i = 1, \ldots, N \) for \( t \geq t_m \). Integrating (4) for any \( n \) and \( t \in [t_k, t_{k+1}) \) with \( k \geq m \) yields
\[
x_n(t) = x_n(t_k) + \int_{t_k}^{t} f_n(\tau) \, d\tau.
\]  

Note that within \( [t_k, t_{k+1}) \) a number of events corresponding to discontinuities in arrival rate processes may occur, i.e., let \( \tau_1, \ldots, \tau_{R_k} \) with \( \tau_0 \in \cup_{n=1}^{N} \Lambda_n \), be the associated event times and set \( \tau_{R_k} = t_k \) and \( \tau_{R_k+1} = t_{k+1} \) for convenience. Suppose \( t \in [\tau_{k,l}, \tau_{k,l+1}) \) for some \( l \in \{0, \ldots, R_k\} \), so that
\[
x_n(t) = x_n(t_k) + \sum_{l=0}^{l'} \int_{\tau_{k,l}}^{\tau_{k,l+1}} f_n(\tau) \, d\tau + \int_{\tau_{k,l+1}}^{t} f_n(\tau) \, d\tau.
\]

Differentiating the term above with respect to \( \theta_{i,m}(t) \) gives
\[
\frac{\partial x_n(t)}{\partial \theta_{i,m}} = \frac{\partial x_n(t_k)}{\partial \theta_{i,m}} + \sum_{l=0}^{l'} \frac{\partial x_n(t)}{\partial \theta_{i,m}} |_{\tau_{k,l}} \frac{\partial f_n(t_{k,l})}{\partial \theta_{i,m}} \, d\tau + \int_{\tau_{k,l+1}}^{t} \frac{\partial f_n(\tau)}{\partial \theta_{i,m}} \, d\tau.
\]  
(12)

All events occurring at \( \tau_{k,l} \), \( l = 1, \ldots, l' \) are exogenous and independent of \( \theta_{i,m}(t) \), so we have \( \frac{\partial x_n(t)}{\partial \theta_{i,m}} |_{\tau_{k,l}} = 0 \), leaving only the term corresponding to \( \tau_{k,0} = t_k \). Thus, we find
\[
\int_{\tau_{k,1}}^{t} \frac{\partial f_n(\tau)}{\partial \theta_{i,m}} \, d\tau = 0 \quad \text{if} \quad t_k \text{ is exogenous.}
\]  
(13)

However, if \( t_k \in \Gamma_n \) for some \( n \), we have \( x_n(t_k) = 0 \). Differentiating both sides with respect to \( \theta_{i,m}(t) \) gives
\[
\frac{\partial t_k}{\partial \theta_{i,m}} = -\frac{\partial x_n(t_k)}{\partial \theta_{i,m}} \frac{\partial f_n(t_k)}{\partial \theta_{i,m}} \quad \text{if} \quad t_k \in \Gamma_n.
\]  
(14)

If \( t_k \in \Sigma_n \), the following result can be established (all the lemma proofs are omitted due to space limitations but may be found in Kebargahtbi and Cassandras (2010)).

### Lemma 1

Let \( t_k \) be the start of a NEP for some queue \( n \). Then, \( \frac{\partial t_k}{\partial \theta_{i,m}} |_{f_i(t_k) - f_j(t_k)} = 0 \) for all pairs \( (i, m) \) and \( l = 1, 2, \ldots, N \).

### 3.1 IPA and the \( \mu \)-rule

The optimality of the \( \mu \)-rule for two queues is already established in Kebargahtbi and Cassandras (2009) and Kebargahtbi and Cassandras (2010) where the sample performance function \( Q(\theta) \) was expressed in terms of a single parameter \( \theta \), defined as the fraction of resource capacity allocated to queue 1. An explicit IPA derivative \( \frac{dQ}{d\theta} \) was derived and it was shown that if \( c_1 \mu_1 > c_2 \mu_2 \) then \( \frac{dQ}{d\theta} < 0 \). This is true for any sample path, leading to \( \theta^* = 1 \) and the \( \mu \)-rule is therefore optimal.

In the sequel, we use the optimality of the \( \mu \)-rule for 2 queues to prove its optimality for any \( N > 2 \) queues using an induction argument. To start let us define
\[
q(t) = \sum_{n=1}^{N} c_n x_n(t), \quad \forall t \in [0, T]
\]
and introduce the following recursion for the cost in (6):
\[ Q_M = 0 \]  
\[ Q_m = g_{m+1}(v_0, \ldots, v_m) + Q_{m+1}, \quad m = 1, \ldots, M - 1 \]
where for any \( m \in \{0, \ldots, M - 1\} \), the function \( g_{m+1}(\cdot) \) is dependent on the past controls \( v_0, \ldots, v_m \). Also, the recursion defined by (15a)-(15b) implies that the functions \( Q_m(\cdot) \) are generally dependent on \( v_0, \ldots, v_{m-1} \). In the sequel, we omit the control vectors from the arguments of functions \( g_m(\cdot) \), \( Q_m(\cdot) \) and \( Q(\cdot) \) for brevity. Noting that the control vector \( v_m \) can only affect the queue contents in the interval \([t_m, T]\), we use (15a)-(15b) to minimize \( Q(\cdot) \) through the (dynamic programming) recursion below.
\[ Q_M^* = 0 \]  
\[ Q_m^* = \min \{g_{m+1} + Q_{m+1}^*, \quad m = 1, \ldots, M - 1 \} \]
where \( Q_0^* = \min_{v_0} Q(v_0, \ldots, v_{M-1}) \). In simple terms, minimization of \( Q(\cdot) \) boils down to recursively finding the optimal controls \( v_0, \ldots, v_{M-1} \) for the interval \([t_m+1, T]\) and the best \( v_m \) for the interval \([t_m, t_m+1]\) assuming that the optimal controls for the intervals \([t_m+1, T]\) have already been implemented. Notice that when \( N = 2 \), \( v_m(t) = \theta_{1,m}(t) \) for any \( m = 0, \ldots, M - 1 \).

Considering the optimality of the \( c\mu \)-rule for \( N = 2 \) as the base step, we set up the induction hypothesis by assuming that the \( c\mu \)-rule is optimal for arbitrary \( K = 2, \ldots, N - 1 \) queues. In the inductive step, we combine (16a) and (16b) with the induction hypothesis to show that when \( N \) queues are present and the \( c\mu \)-rule is optimal and implemented in the interval \([t_{m+1}, T]\), it is also optimal for the interval \([t_m, T]\). We do this by showing that deviating from the \( c\mu \)-rule in the interval \([t_m, t_{m+1}]\) increases the cost \( Q(u_0, \ldots, u_{M-1}) \). Specifically, we perturb \( \theta_{1,m}(t) \) away from its value under the \( c\mu \)-rule by \( \theta_{n,m}(t) = \theta_{1,m}(t) + \varepsilon \) for \( n > 1 \) such that no event order changes results at and after \( t_{m+1} \). Moreover, this perturbation is such that (8) is preserved. Next, stepping backwards in time, the induction hypothesis applies to all intervals with fewer than \( N \) non-empty queues. Thus, we need only examine those intervals where all \( N \) queues are non-empty. To this end, let us consider an interval \([t_{m}, t_{m+1}]\) with \( N_m = N \) and let \( \eta_m \) be the first time at which the \( n \)th queue becomes empty after \( t_m \). We assume that the newly added queue has the lowest \( c\mu \) product, that is, \( cN_n \mu_N < c_{n+1} \mu_{n+1} \) for any \( n \neq N \). This is simply for convenience and does not restrict generality. To apply the above mentioned idea, we will introduce a series of lemmas all of which apply to the following setting: Consider an interval \([t_m, t_{m+1}]\) with \( N_m = N \). Also, assume that the \( c\mu \)-rule is applied for \( t \geq t_{m+1} \). Moreover, let \( \theta_{1,m}(t) \) be perturbed as explained in the outline of the analysis. Thus, as shown in Fig. 2, queues are served according to their \( c\mu \) value highest to lowest and NEPs end at \( \eta_1, \eta_2, \ldots, \eta_N \) and such that
\[ t_{m+1} = \eta_1 < \eta_2 < \ldots < \eta_N. \]
Now let us define for any event time \( t_k, k = 1, \ldots, M - 1 \):
\[ \Delta \theta_n(t_k) = \theta_n(t_k^+) - \theta_n(t_k^-) = \theta_{n,k}(t_k^+) - \theta_{n,k-1}(t_k^-) \]
representing the amount by which a control changes when this event occurs and affecting the queue content deriv-

---

Fig. 2. Time line of events in and after the interval \([t_m, t_{m+1}]\).
We now introduce the main theorem of this section.

**Theorem 1.** The \( c^\mu \)-rule minimizes the cost (6) for a system of \( N \) parallel queues with dynamics (4), control parameters (5) and the constraint (8).

**Proof:** Due to page limit, we omitted some details. A complete proof may be found in Kebabirbottib and Cassandas (2010).

*Last interval \([t_{M-1}, t_M] \):* If \( N_{M-1} < N \), by the induction hypothesis the \( c^\mu \)-rule is already optimal. If \( N_m = N \),

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} = \sum_{n=1}^{N} c_n \frac{\partial x_n(t)}{\partial \theta_{1,m}}, \quad i = 1, \ldots, N - 1.
\]

Using Lemma 3, for each \( i = 1, \ldots, N - 1 \), we get

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} = -(c_{1^i} - c_{N^N} N) (t - \Pi_{m-1})
\]

Since \( c_{1^i} > c_{N^N} N \), for all \( i = 1, \ldots, N - 1 \), we get

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} < 0, \quad i = 1, \ldots, N - 1.
\]  

Moreover, according to the \( c^\mu \) order in (7), we can write \( (c_{1^i} - c_{N^N} N) \leq 1 - \frac{1}{2} \Pi_{m-1} (t - \Pi_{m-1}) \) for all \( i = 1, \ldots, N - 1 \), hence, (34) is proven.

Next, using part (b) to find \( \Delta \theta_k(t_{k-1}, m) \) for \( k = i+1, \ldots, j \).

If there is an interval \([t_i, t_j] \subset [\eta_j, \eta_{j+1}, m] \) over which \( x_n(t) > 0 \) with \( n < j \), it is shown in Kebabirbottib and Cassandas (2010) that \( \frac{\partial x_n(t)}{\partial \theta_n,m} = 0 \) and (28) is still valid.

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} = \sum_{n=j+1}^{N} c_n \frac{\partial x_n(t)}{\partial \theta_{1,m}} + \sum_{k=i}^{j} \frac{\partial x_{k,m}}{\partial \theta_{1,m}} \mu_n \Delta \theta_n(t_{k-1}, m).
\]  

(29)

There are now two cases regarding the range of \( i \):

- \( i > j \): In this case, the inner sum in (29) vanishes.
- \( i < j \): We consider two subcases:
  - (a) \( j \geq 2 \): In this case, since the sum in (29) is such that \( n = j \), the condition \( n = i \) cannot be true. Using Lemma 3, we find that the first sum in (29) becomes \( c_N \frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}} \). Concerning the inner sum in (29) we can use part (b) of Lemma 2 to see that \( \frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}} = 0 \) except for the case \( n = j \) and \( k = j \). Using these results in (29) we conclude that, for \( t \in [\eta_j, \eta_{j+1}, m] \),

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} < \frac{\partial q(t)}{\partial \theta_{j+1,m}} < \cdots < \frac{\partial q(t)}{\partial \theta_{N-1,m}} < 0.
\]  

(30)

- (b) \( i = 1 \): We have

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} = \frac{\partial x_1(t)}{\partial \theta_{1,m}} = \frac{\partial x_1(t)}{\partial \theta_{1,m}}, \quad t \in [\eta_1, \eta_{1+1}, m] \]

(31)

By (21), we have

\[
\frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}} = -\frac{\mu_N}{\mu_1} \frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}}.
\]

The derivative \( \frac{\partial q(t)}{\partial \theta_{1,m}} \) can be calculated by Lemma 7 for \( j \geq 1 \). For brevity, we also define:

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} = \frac{1}{\mu_1} \frac{\partial x_1(t_{j-1}, m)}{\partial \theta_{1,m}} \Delta \theta_{j-1}(t_{j-1}, 1) + c_N \frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}}
\]

(32)

By Lemma 3 we have \( \frac{\partial x_N(t_{j-1}, m)}{\partial \theta_{1,m}} < 0 \). We shall now show that \( r_i,m A_i,j \Delta \theta_j(\eta_j) = 1 \) for \( i < j \) and \( r_i,m A_i,j \Delta \theta_j(\eta_j) \geq 1 \) for \( i = j \) which yields

\[
\frac{\partial q(t)}{\partial \theta_{1,m}} < \frac{\partial q(t)}{\partial \theta_{j+1,m}} < \cdots < \frac{\partial q(t)}{\partial \theta_{N-1,m}} < 0
\]

(33)

for all \( t \in [\eta_j, \eta_{j+1}, m] \), \( j = 2, \ldots, N - 1 \). We show that,

\[
r_i,m \Delta \theta(t_{j-1}, 1) \geq 1, \quad r_k,m \Delta \theta(k_{k-1}, 1) = 1, \quad k = 3, \ldots, j.
\]

(34)

Using part (a) of Lemma 2 and (24) with \( k = 1 \) we get

\[
\Delta \theta(t_{j-1}, 1) = 1 - \frac{\theta_2(t_{j-1}, 1) - \alpha(t_{j-1}, 1)}{\theta_1(t_{j-1}) - \alpha(t_{j-1}, 1)}.
\]

(35)

As discussed in the outline of the analysis, in \([\eta_j, \eta_{j+1}, m] \) we deviate from the \( c^\mu \)-rule by perturbing \( \theta_n,m(t) \) by \( \delta_n,m(t) \) while preserving (8). It follows that \( \theta(t_{j-1}, 1) \leq 1 - \theta_2(t_{j-1}, 1) \), hence, (34) is proven.

Next, using part (b) to find \( \Delta \theta(t_{k-1}, m) = 1 + i + \ldots, j \)
and using $r_{k-1,m} = \frac{-\mu_{k-1}}{f_{k-1}(\eta_{k-1,m})}$, after omitting details we get $r_{k-1,m} \Delta \theta_k(\eta_{k-1,m}) = 1$ which verifies (35). Applying this to (32) for $i > 1$, gives

$$
\frac{\partial q(t)}{\partial \theta_{1,m}} = \frac{\partial q_1(\eta_{1,m})}{\mu_1 \Delta \theta_k(\eta_{k-1,m})} = c_{\eta_{1,m}} \mu N, \forall t \in \{t_j,m, t_{j+1,m+1}\}
$$

which is independent of $i$. This proves the equalities in (33). In a similar way we find that when $i = 1$ (34) yields $r_{1,m} \Delta \theta_k(\eta_{1,m}) \geq 1$ and (33). Combining (30) and (33), for any $t \in [\eta_{1,m}, \eta_{1,m+1}]$ with $j \geq 2$ we have

$$
\frac{\partial q(t)}{\partial \theta_{1,m}} \leq \frac{\partial q(t)}{\partial \theta_{2,m}} = \cdots = \frac{\partial q(t)}{\partial \theta_{j-1,m+1}} < \cdots < \frac{\partial q(t)}{\partial \theta_{N-1,m}} < 0.
$$

(b) $i = j = 1$: Applying Lemma 3 to (29) and with the help of (21) with $t = \eta_{1,m}$ and (14) with $t_k = \eta_{1,m}$ we get

$$
\frac{\partial q(t)}{\partial \theta_{1,m}} = 1 \frac{\partial q_1(\eta_{1,m})}{\mu_1 \Delta \theta_k(\eta_{1,m})} \left[ \sum_{n=2}^{N} \frac{c_{\eta_{1,m}} \Delta \theta_n(\eta_{1,m}) - c_{\eta_{1,m}} \mu N}{\alpha_i(\eta_{1,m})} \right].
$$

By Lemma 3, we have $\frac{\partial q_1(\eta_{1,m})}{\partial \theta_{1,m}} < 0$, therefore it remains to establish the negativity of the above derivative by showing that the bracketed term is positive. We accomplish this by finding a lower bound $L_m < \sum_{n=2}^{N} \frac{c_{\eta_{1,m}} \Delta \theta_n(\eta_{1,m})}{\alpha_i(\eta_{1,m})}$ such that

$$
L_m - c_{\eta_{1,m}} \mu N > 0.
$$

Using part (a) of Lemma 2, we have $\Delta \theta_k(\eta_{1,m}) > 0$ and $\Delta \theta_k(\eta_{1,m}) < 0$ for $n > 2$. Using the fact that $c_{\eta_{1,m}} \mu N < c_{\eta_{1,m}} \mu N$ for $n > 2$, we find a lower bound

$$
L_m = c_{\mu N} \sum_{n=2}^{N} \frac{\Delta \theta_n(\eta_{1,m})}{\alpha_i(\eta_{1,m})} \theta_i(\eta_{1,m}) - \frac{\alpha_{i+1}(\eta_{1,m})}{\alpha_i(\eta_{1,m})}.
$$

According to (8) we have $\theta_i(\eta_{1,m}) = 1 - \sum_{n=2}^{N} \theta_n(\eta_{1,m})$, thus, we get $L_m = c_{\mu N} \theta_i(\eta_{1,m})$. Next, noting that $c_{\mu N} \theta_i(\eta_{1,m})$ reveals that $\frac{\partial q(t)}{\partial \theta_{1,m}} < 0$. The proof is complete noting that $\frac{\partial q(t)}{\partial \theta_{1,m}} < \frac{\partial q(t)}{\partial \theta_{1,m}}$ for all $t \in [\eta_{1,m}, \eta_{2,m}]$ and for all $i = 2, \ldots, N - 1$ using (29).

4. CONCLUSIONS

We have considered a classic scheduling problem with a single resource shared by $N$ competing queues in the context of SFMs. By means of the IPA methodology, we have derived explicit sample derivatives of the cost function with respect to a controllable set of parameters in the scheduling policy. In case of the linear total holding cost objective, exploiting the monotonicity of these sample derivatives we have proved the optimality of the well-known $c\mu$-rule not only for the two-queue case (as in earlier work Kebarighotbi and Cassandras (2009)) but for an arbitrary finite number of queues and stochastic processes under non-idling policies. The generality of our results confirms the validity of the $c\mu$-rule without having to make explicit distributional assumptions on the random processes involved or resort to heavy traffic analysis.

REFERENCES


