Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

Revisiting the Optimality of the $c\mu$-rule with Stochastic Flow Models

Ali Kebarighotbi and **Christos G. Cassandras**
Division of Systems Engineering
and Center for Information and Systems Engineering
Boston University
Brookline, MA 02446
alik@bu.edu,cgc@bu.edu

***Abstract*— We revisit, in the context of Stochastic Flow Models (SFMs), a classic scheduling problem for optimally allocating a resource to multiple competing users. For the two-user case, we establish the optimality of the well-known $c\mu$-rule for arbitrary stochastic processes using calculus of variations arguments as well as an Infinitesimal Perturbation Analysis (IPA) approach. The latter allows us to derive an explicit sensitivity estimate of the cost function with respect to a controllable parameter and to further study the problem when the cost function is nonlinear, deriving simple distribution-free cost sensitivity estimates and analyzing why the $c\mu$-rule may fail in this case.**

I. Introduction

The classic prototypical stochastic scheduling problem involves a single resource whose service capacity is to be optimally shared by N competing users. Each user submits tasks which may have to wait for service in the user's queue, normally on a First Come First Served (FCFS) basis. In a queueing theory framework, this problem is modelled as a system of N parallel queues, each with its own arrival process, connected to a single server. The server processes tasks from the nth queue with rate μ_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, but the server may preempt a task by interrupting its processing to serve a new task from some other queue. This basic model applies to a large spectrum of applications in communication networks, manufacturing, and computer processing.

The usual objective in the scheduling problem is to minimize the overall average holding cost of tasks in the system with c_n denoting the cost per unit waiting time in the nth 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 queue (the one with largest $c_n\mu_n$ value) unless it is empty; in that case, the server selects the second queue and so on. The optimality of the $c\mu$-rule seems to have been first suggested in [1] 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 [2] later proved the optimality of the $c\mu$-rule for a multi-class $M/G/1$ system. 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.

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 [3] and [4]. There have since been various attempts to extend these results. For example, it is shown in [5] that for a discrete time $G/G/1$ model with a non-idling and non-preemptive server 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 [6], [7] and a stochastic setting where the optimality of the $c\mu$-rule can be obtained using heavy traffic (fluid limit) arguments [8],[9],[10]. A "generalized" $c\mu$-rule can then be shown to be asymptotically optimal [11] not only for the linear but also for convex cost objectives.

In this paper, we revisit the basic stochastic scheduling problem using a *Stochastic Fluid Model* (SFM). 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. Clearly, finding "appropriate" rate processes to approximate the behavior of the system to any arbitrary degree of accuracy is far from trivial. However, the emphasis in using SFMs is not in deriving approximations of performance measures of the underlying discrete event system, but rather studying sample paths from which one can derive structural properties and optimal policies. SFMs were introduced in [12] to carry out *Infinitesimal Perturbation Analysis* (IPA) for a queueing system with finite capacity to estimate derivatives of performance measures such as workload and loss with respect to controllable parameters and, therefore, solve performance optimization problems using stochastic gradient-based algorithms. In this case, the derivative estimates are independent of the probability laws of the stochastic rate processes and require minimal information from the observed sample path as shown in [12]. Extensions to serial networks [13], systems with feedback control mechanisms [14], and

The authors' work is supported in part by NSF under Grants DMI-0330171 and EFRI-0735974, by AFOSR under grants FA9550-07-1-0361 and FA9550-09-1-0095, and by DOE under grant DE-FG52-06NA27490.

some multi-class models [15], [16], [17] have also been obtained.

Our purpose in this paper is to take a first step in studying general scheduling problems using SFMs. For the specific scheduling problem described above, we view the arrival processes as flows with arbitrary time-varying rates that behave as random processes. On the processing side, we associate a maximal rate μ_n with each flow class and control its actual service rate $u_n(t)$ so that $\sum_n u_n(t)/\mu_n \leq 1$. In this paper, we restrict ourselves to the case of two classes. Using simple calculus of variations arguments on an arbitrary sample path of the SFM, we show that, if $c_1\mu_1 > c_2\mu_2$, the optimal solution is the $c\mu$-rule. We then obtain the same result using IPA by deriving a sample derivatives of the total holding cost metric with respect to a fixed parameter θ such that $u_1(t) = \mu_1\theta$ as long as queue 1 is not empty; we show that if $c_1\mu_1 > c_2\mu_2$, this derivative is always negative, thereby proving the optimality of the $c\mu$-rule independent of stochastic characteristics or traffic load.

Although this result is pleasing, it comes as no major surprise since it merely formalizes the known fact that the $c\mu$-rule is a property of the underlying system dynamics and not its stochastic characteristics (e.g., [6]). However, this raises some interesting questions, such as "at what exact point does the $c\mu$-rule break down?", "if not optimal, when can it provide a good approximation of the optimal policy?", and "how do we proceed to solve problems where it no longer applies and an optimal solution might depend on the *unknown* stochastic characteristics of the model?" Thus,the contributions of this paper are to (i) provide insight through IPA into the properties of a sample path that enable the $c\mu$-rule to hold and obtain an explicit derivative estimator which can be readily extended to more general scheduling problems, and (ii) consider an extension of the basic scheduling problem where the cost function is nonlinear in the queue contents.

In Section II of the paper, the basic scheduling problem is formulated in a SFM setting and calculus of variations methods are used to derive the $c\mu$-rule on a sample path basis. In Section III, IPA is carried out to derive the sample derivative of the cost function with respect to a scheduling policy parameter. In Section III, the monotonicity of the IPA derivative is proved and, therefore, the optimality of the $c\mu$-rule. In Section IV, we study the case of nonlinear costs and conclude with Section V where we outline ongoing work.

II. Problem Formulation

In the context of SFM, the basic scheduling problem we consider is depicted in Fig.1 which consists of a single resource and two parallel queues competing for service. The queues hold different types of fluids and are indexed by $n = 1, 2$. The maximum rate at which the resource processes fluid from the nth queue is denoted by μ_n. The system involves a number of stochastic processes which all are defined on a common probability space $(\Omega, \mathcal{F}, P)$. The external *inflow* rate process to the nth queue is denoted by $\{\alpha_n(t)\}$ capturing the instantaneous rate of arriving tasks. The process describing the *service flow* rate from the nth queue to the resource at any time t is denoted by $\{u_n(t)\}$. These flow rates are subject to the capacity constraint:

$$\frac{u_1(t)}{\mu_1} + \frac{u_2(t)}{\mu_2} \leq 1 \tag{1}$$

where the rate $u_n(t)$ is controllable provided it also satisfies $0 \leq u_n(t) \leq \mu_n$. The *outflow* rate from the resource is denoted by $\{\beta(t)\}$. Note that $\beta(t) = u_1(t) + u_2(t)$ for all t. Finally, the process describing the content of the nth queue is denoted by $\{x_n(t)\}$, where $x_n(t) \geq 0$. We will be studying this SFM over a finite time interval $[0, T]$.

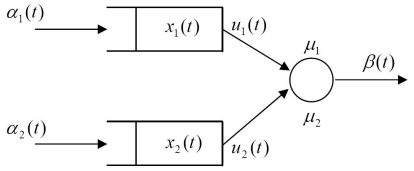

Fig. 1. Stochastic Fluid Model (SFM) for a scheduling problem.

The inflow rate processes $\{\alpha_n(t)\}$, $n = 1, 2$, are allowed to be arbitrary except for the following condition. However, all subsequent results are readily extendable to piecewise continuously differentiable and bounded-variation inflow rates.

Assumption 1. W.p.1, the inflow processes $\{\alpha_n(t)\}$ are continuously differentiable in $[0, T]$.

The queue content dynamics follow the one-sided differential equations, for $n = 1, 2$,

$$\frac{dx_n(t)}{dt^+} = \begin{cases} 0 & x_n(t) = 0, u_n(t) \geq \alpha_n(t) \\ \alpha_n(t) - u_n(t) & \text{otherwise.} \end{cases} \tag{2}$$

A typical sample path of the nth queue content is shown in Fig. 2. There are two types of events associated with this SFM, one which initiates a *non-empty period* (NEP) at queue n and one that terminates it and initiates an *empty period* (EP). Let $\xi_{n,k}$ denote the kth time in the sample path of queue n when this queue becomes non-empty. Similarly, $\eta_{n,k}$ denotes the kth transition time of queue n into an EP. Accordingly, an EP is a maximal interval $[\eta_{n,k}, \xi_{n,k+1}]$, over which $x_n(t) = 0$ and an NEP is a supremal interval $(\xi_{n,k}, \eta_{n,k})$ with $x_n(t) > 0$. Let $L_{n,k} = \eta_{n,k} - \xi_{n,k}$ be the length of the kth NEP of queue n in $[0, T]$.

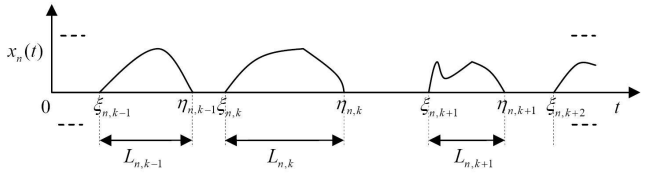

Fig. 2. A typical sample path of the system

It follows from (2) that $u_n(t) = \alpha_n(t)$ during an EP of the nth queue, since the case $u_n(t) > \alpha_n(t)$ would correspond to the resource devoting more of its capacity than is needed to meet the demand rate $\alpha_n(t)$. Conversely, an NEP starts as

soon as $\alpha_n(t) > u_n(t)$; this may arise because the external inflow rate exceeds the currently allocated service flow $u_n(t)$ or because the controllable flow $u_n(t)$ is selected to be below the current inflow rate.

The controllable service flow rate $u_n(t)$ defines the scheduling policy adopted in the system. We set $u_n(t) = \mu_n\theta_n(t)$ with $\theta_n(t) \in [0,1]$ to allocate a fraction of the maximal allowable service rate μ_n to the nth queue. This fraction may depend on $x_1(t)$, $x_2(t)$, assuming they are observable (this may not always be the case, as in a wireless network where a channel is allocated to two upstream links whose queues may not be known instantaneously). In the sequel, we let $\theta_n(t)$ be time-varying but show that for the specific class of optimal scheduling problems considered, it is constant or at most switches between its feasible limits 0 and 1. Formally, we set:

$$u_1(t) = \begin{cases} \min\{\alpha_1(t), \mu_1\theta(t)\}, & x_1(t) = 0, \\ \mu_1\theta(t), & x_1(t) > 0. \end{cases} \quad (3)$$

Thus, by (2), an event at time t such that $x_1(t) = 0$ and $\alpha_1(t) > \mu_1\theta(t)$ defines the start of an NEP at queue 1. Using (1), $u_2(t) \leq \mu_2(1 - \frac{u_1(t)}{\mu_1})$. Therefore, if at time t queue 2 is empty and $\alpha_2(t) \leq \mu_2(1 - \frac{u_1(t)}{\mu_1})$, then $u_2(t) = \alpha_2(t)$, otherwise $u_2(t) = \mu_2(1 - \frac{u_1(t)}{\mu_1})$, i.e.,

$$u_2(t) = \begin{cases} \min\{\alpha_2(t), \mu_2(1 - \frac{u_1(t)}{\mu_1})\} & x_2(t) = 0, \\ \mu_2(1 - \frac{u_1(t)}{\mu_1}) & x_2(t) > 0. \end{cases} \quad (4)$$

We consider a *total holding cost* performance objective defined for any sample path denoted by $\omega \in \Omega$ as

$$Q(\omega) = \frac{1}{T}\int_0^T \sum_{n=1}^{2} c_n x_n(t,\omega)dt \quad (5)$$

where c_n is a cost rate associated with queue n. Moreover, for each queue $n = 1, 2$ we define a sample function $Q_n(\omega) = \int_0^T c_n x_n(t,\omega)dt$.
We can immediately observe that since $x_n(t,\omega) = 0$ during EPs at queue n, we can rewrite this in the form

$$Q_n(\omega) = \sum_{k=1}^{M_n} \int_{\xi_{n,k}}^{\eta_{n,k}} c_n x_n(t,\omega)dt \quad (6)$$

where $M_n \geq 0$ is the number of NEPs for the nth queue, $n = 1, 2$, including a possibly incomplete NEP at T.

The optimization problem we aim to solve is:

$$\min_{\theta(t)\in[0,1]} \mathrm{E}\left[Q(\omega)\right], \quad (7)$$
$$\text{subject to } (1), (2), (3) \text{ and } (4).$$

Let us first consider a specific sample path ω so that the problem above involves the minimization of (5). Let the right-hand-side of (2) be $f_n(x_1, x_2, \theta)$, $n = 1, 2$. Then, viewed as an optimal control problem, we may write the Hamiltonian function after some regrouping as

$$H(\theta, x_1, x_2, \lambda_1, \lambda_2) = c_1x_1 + c_2x_2 + \lambda_1\alpha_1(t) + \lambda_2\alpha_2(t) - \lambda_2\mu_2 - (\mu_1\lambda_1 - \mu_2\lambda_2)\theta(t)$$

where $\lambda_n(t)$, $n = 1, 2$, are the costate variables. We have focused on the case where $u_1(t) = \mu_1\theta(t)$ and $x_2(t) > 0$; otherwise the Hamiltonian is independent of one or both costate functions. The remaining cases do not add any insight and are omitted. By simple application of Pontryagin's principle, we can see that

$$\theta^*(t) = \mathrm{sgn}[\mu_1\lambda_1(t) - \mu_2\lambda_2(t)] \quad (8)$$

where $\mathrm{sgn}[x] = 1$ if $x \geq 0$ and 0 otherwise. Thus, aside from the case $\mu_1\lambda_1(t) - \mu_2\lambda_2(t) = 0$, when $\mu_1\lambda_1(t) - \mu_2\lambda_2(t) > 0$ the Hamiltonian is minimized by $\theta(t) = 1$, otherwise it is minimized by $\theta(t) = 0$. Observe that this result holds regardless of $\alpha_1(t)$, $\alpha_2(t)$ or the form of the integrand in (6) as long as it does not depend on $\theta(t)$. This fact is consistent with the analysis of a similar deterministic scheduling problem in [6], where it is aptly pointed out that the nature of an optimal policy in such problems is determined by the underlying dynamics and not the stochastic characteristics.

One can subsequently proceed to study the costate equations given by

$$\frac{d\lambda_n}{dt} = -\frac{\partial H}{\partial x_n} = -c_n, \quad \lambda_n(T) = 0 \quad (9)$$

to determine the behavior of $\mathrm{sgn}[\mu_1\lambda_1(t) - \mu_2\lambda_2(t)]$ or $\mathrm{sgn}[\mu_1\lambda_1(t)]$. It is easy to see that $\lambda_1(t) = c_1(T - t) > 0$ and $\lambda_2(t) = c_2(T - t) > 0$ for at least some interval $(t, T]$, therefore, if $c_1\mu_1 > c_2\mu_2$ we get $\mu_1\lambda_1(t) - \mu_2\lambda_2(t) > 0$ for all $t < T$. In other words, as long as $u_1(t) = \mu_1\theta(t)$ according to (3), the queue 1 flow is served at its maximal feasible rate μ_1 and is, therefore, prioritized (unless it is empty and $\alpha_1(t) < \mu_1$); this is precisely the $c\mu$-rule.

Next, we proceed with an IPA approach that recovers the same result. We set $\theta(t) = \theta$ to be a fixed parameter and analyze the sample derivative $dQ/d\theta$. The behavior of this derivative will show us whether $\theta^* = 0$ or 1 or whether it switches under certain conditions. This approach has the benefit of providing us with an explicit form of this derivative which allows us to study solutions of problem (7) over the class of policies parameterized by θ. While for this problem we can show that the optimal solution is the simple $c\mu$-rule, it paves the way for considering more general scheduling problems where one often resorts to such parametric families of policies. We henceforth omit ω in the expressions to simplify the notation. We also use θ as argument in (4) through (6) and use $x_n(t;\theta)$, $n = 1, 2$ to stress their dependence on it; we sometimes keep this implicit in $\xi_{n,k}$ and $\eta_{n,k}$ to save space.

III. Infinitesimal Perturbation Analysis (IPA)

In the sequel, we denote the derivative of any function $g(t,\theta)$ with respect to θ by $g'(t;\theta)$. Using this notation for (6) we obtain

$$Q_n'(\theta) = \sum_{k=1}^{M_n} \Big\{ [\eta_{n,k}' c_n x_n(\eta_{n,k};\theta) - \xi_{n,k}' c_n x_n(\xi_{n,k};\theta)] + \int_{\xi_{n,k}}^{\eta_{n,k}} c_n x_n'(t;\theta)dt \Big\}.$$

At the start and end of NEPs, $x_n(\xi_{n,k};\theta) = x_n(\eta_{n,k};\theta) = 0$. The only possible exception is if $\eta_{n,M_n} = T$, in which case obviously $\eta'_{n,M_n} = 0$. Therefore, for $n = 1, 2$ we get

$$Q'_n(\theta) = \sum_{k=1}^{M_n} \int_{\xi_{n,k}}^{\eta_{n,k}} c_n x'_n(t;\theta)dt. \tag{10}$$

Thus, the first step in derivation of $Q'_n(\theta)$ is to find $x'_n(t;\theta)$ for $t \in [\xi_{n,k}, \eta_{n,k})$ over the index range. Let $f_n(t;\theta) = \alpha_n(t) - u_n(t;\theta)$ be the *net flow function* for each queue. Then, for $t \in [\xi_{n,k}, \eta_{n,k})$ we have $x_n(t;\theta) = \int_{\xi_{n,k}}^{t} f_n(\tau;\theta)d\tau$. Consequently, we have $x'_n(t;\theta) = -\xi'_{n,k}(\theta)f_n(\xi_{n,k}) + \int_{\xi_{n,k}}^{t} f'_n(\tau;\theta)d\tau$. The first term in this expression is nil thanks to the lemma bellow. In the sequel, proofs are omitted due to space limitations.

Lemma 1. If $\xi_{n,k}(\theta), n = 1, 2$ is the start time of an NEP in a sample path of $x_n(t;\theta)$, then $\xi'_{n,k}(\theta)f_n(\xi_{n,k};\theta) = 0$.

Using Lemma 1, we can write

$$x'_n(t;\theta) = \int_{\xi_{n,k}}^{t} f'_n(\tau;\theta)d\tau, \quad t \in [\xi_{n,k}\ \eta_{n,k}) \tag{11}$$

and $\xi'_{n,k}(\theta)$ is not needed in our analysis; however, we will need $\eta'_{n,k}(\theta)$. For the kth NEP of queue n, we have $x_n(\eta_{n,k};\theta) = \int_{\xi_{n,k}}^{\eta_{n,k}} f_n(\tau;\theta)d\tau = 0$. Using (11) and Lemma 1, we obtain $\eta'_{n,k}(\theta)f_n(\eta^-_{n,k};\theta) + x'_n(\eta^-_{n,k};\theta) = 0$ where $f_n(\eta^-_{n,k};\theta) = \lim_{t\uparrow\eta_{n,k}} f_n(t;\theta)$ and $x'_n(\eta^-_{n,k};\theta)$ is similarly defined. Note that, by (2), $f_n(\eta^-_{n,k};\theta) \neq 0$ (in fact, $f_n(\eta^-_{n,k};\theta) < 0$ since an NEP ends at $\eta_{n,k}$) and we conclude that

$$\eta'_{n,k}(\theta) = \frac{-x'_n(\eta^-_{n,k};\theta)}{f_n(\eta^-_{n,k};\theta)}, \quad n = 1, 2,\ k = 1, \ldots, M_n \tag{12}$$

In what follows we consider a typical NEP (ξ_n, η_n), $n = 1, 2$, dropping the index k for simplicity. Regarding the relative positioning of NEPs on the time line, there are six possible cases that can arise as shown in Figs. 3(a) through 3(f) in which d is the length of an *overlapping interval* (if one exists) between NEPs of two queues. The case where $\xi_1 = \xi_2$ can be accommodated within Cases 3 through 6. Moreover, multiple NBPs of one queue in the NBP of another can be constructed using superposition of these 6 cases. We consider each of the cases in Figs. (3(a)) through (3(f)) and derive the associated derivatives in (10).

A. Determining $Q'_1(\theta)$

Looking at (3), note that for all $t \in [\xi_1, \eta_1)$ we have $u_1(t;\theta) = \mu_1\theta$, therefore $f_1(t;\theta) = \alpha_1(t) - \mu_1\theta$ and $f'_1(t;\theta) = -\mu_1$. Thus, using (11),

$$x'_1(t;\theta) = -\mu_1(t - \xi_1), \quad t \in [\xi_1, \eta_1). \tag{13}$$

It follows that

$$\int_{\xi_1}^{\eta_1} c_1 x'_1(t;\theta)dt = -c_1\mu_1 \int_{\xi_1}^{\eta_1} (t-\xi_1)dt = -\frac{c_1\mu_1}{2}(\eta_1-\xi_1)^2$$

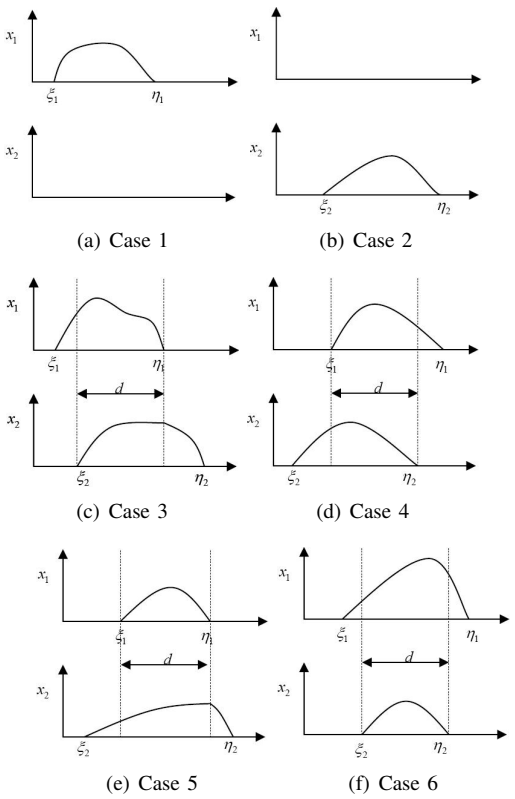

Fig. 3. Relative positioning of NEPs for the two SFM queues.

and, using (10) and recalling that $L_{1,k} = \eta_{1,k} - \xi_{n,k}$, gives

$$Q'_1(\theta) = -\frac{c_1\mu_1}{2} \sum_{k=1}^{M_1} L^2_{1,k}. \tag{14}$$

It is worth noting that $Q'_1(\theta) \leq 0$, as expected.

B. Determining $Q'_2(\theta)$

Based on the SFM dynamics (2) and service flow rates (3),(4), $f_2(t;\theta)$ can be expressed as

$$f_2(t;\theta) = \begin{cases} \alpha_2(t) - \mu_2(1-\theta) & \text{if } x_1(t;\theta) > 0, \\ & \quad x_2(t;\theta) > 0, \\ \alpha_2(t) - \mu_2(1 - \frac{\alpha_1(t)}{\mu_1}) & \text{if } x_1(t;\theta) = 0, \\ & \quad x_2(t;\theta) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where the first row has used the fact that when $x_1(t) > 0$, $u_1(t) = \mu_1\theta$. Upon differentiation with respect to θ, we get

$$f'_2(t;\theta) = \begin{cases} \mu_2 & \text{if } x_1(t;\theta) > 0,\ x_2(t;\theta) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Referring to Figs. (3(a)) through (3(f)), note that in Case 1, $Q_2(\theta) = 0$, hence $Q'_2(\theta) = 0$, and in Case 2, (10) and (16) imply also that $Q'_2(\theta) = 0$. For other cases $Q'_2(\theta) \neq 0$.

Let us consider case 3, in particular, each of the intervals $[\xi_1, \xi_2)$, $[\xi_2, \eta_1)$, and $[\eta_1, \eta_2)$, respectively.

a) $t \in [\xi_1, \xi_2)$: In this case $x_2(t) = 0$, therefore,

$$x'_2(t) = 0, \quad t \in [\xi_1, \xi_2). \tag{17}$$

b) $t \in [\xi_2, \eta_1)$: Using (11) and (16) we obtain

$$x_2'(t) = \int_{\xi_2}^{t} \mu_2 d\tau = \mu_2(t - \xi_2), \quad t \in [\xi_2, \eta_1). \tag{18}$$

c) $t \in [\eta_1, \eta_2)$: We have

$$x_2(t) = \int_{\xi_2}^{\eta_1} f_2(\tau;\theta)d\tau + \int_{\eta_1}^{t} f_2(\tau;\theta)d\tau.$$

By (15), $f_2(\tau;\theta) = \alpha_2(\tau) - \mu_2(1-\theta)$ for $\tau \in [\xi_2, \eta_1)$ and $f_2(\tau;\theta) = \alpha_2(\tau) - \mu_2(1 - \frac{\alpha_1(\tau)}{\mu_1})$ for $\tau \in [\eta_1, t)$. Therefore, differentiating with respect to θ and using Lemma 1 we get:

$$x_2'(t) = \mu_2(\eta_1 - \xi_2) + \eta_1' \left[\frac{\mu_2}{\mu_1}(\mu_1\theta - \alpha_1(\eta_1))\right].$$

Using (12), we have $\eta_1' = \frac{\mu_1(\eta_1 - \xi_1)}{\alpha_1(\eta_1) - \mu_1\theta}$ Following some algebraic manipulations, we obtain:

$$x_2'(t;\theta) = -\mu_2(\xi_2 - \xi_1), \quad t \in [\eta_1, \eta_2) \tag{19}$$

Combining the results (17), (18), and (19) and using (10), $Q_2'(\theta) = \frac{c_2\mu_2}{2}d^2 - c_2\mu_2\Delta$ where $\Delta = (\xi_2 - \xi_1)(\eta_2 - \eta_1) > 0$ is called *drift* and $d = (\eta_1 - \xi_2)$.

The analysis in the remaining three cases is similar. Therefore, omitting details, in all cases: $Q_2'(\theta) = \frac{c_2\mu_2}{2}d^2$ with $d = \eta_2 - \xi_1$, $d = \eta_1 - \xi_1$ and $d = \eta_2 - \xi_2$ for cases 4, 5 and 6, respectively. In summary, $Q_2'(\theta) = \frac{c_2\mu_2}{2}d^2 > 0$ with the exception of Case 3 in which the final result has an extra negative term. It is also easy to see that every sample path of the SFM over $[0, T]$ can be partitioned into intervals that are either EPs or NEPs that fall into one of the six cases in Figs. 3(a) through 3(f). To obtain a general expression for $Q_2'(\theta)$, let the jth NEP of queue 2 include D_j overlapping intervals with lengths $d_{j,k}, k = 1, \ldots, D_j$. Let

$$j^* = \arg \max_{i=1,2,\ldots} \{i : \xi_{1,i} \le \xi_{2,j}\}$$

if it exists, i.e., $(\xi_{1,j^*}, \eta_{1,j^*})$ is the last NEP of queue 1 which starts before $(\xi_{2,j}, \eta_{2,j})$. Define

$$\Delta_j = (\xi_{2,j} - \xi_{1,j^*})(\eta_{2,j} - \eta_{1,j^*}) \tag{20}$$

where $\xi_{2,j} - \xi_{1,j^*} \ge 0$ by the definition of j^* and observe that Δ_j^+ (where $a^+ = \max\{0, a\}$) is precisely the term associated with $d_{j,k}^2$ whenever Case 3 arises. Then, collecting all the results above for the six cases, we get

$$Q_2'(\theta) = \frac{c_2\mu_2}{2} \sum_{j=1}^{M_2} \left\{ \sum_{k=1}^{D_j} d_{j,k}^2 - 2\Delta_j^+ \right\}.$$

Alternatively, all overlapping intervals can also be indexed according to NEPs of queue 1. Thus, let the ith NEP of queue 1 include D_i overlapping intervals with lengths $d_{i,k}, k = 1, \ldots, D_i$, and define $\Delta_i = (\xi_{2,i^*} - \xi_{1,i})(\eta_{2,i^*} - \eta_{1,i})$ to be the obvious analog of Δ_j above, with i^* being the index of the last NEP of queue 2 which starts within $(\xi_{1,i}, \eta_{1,i})$. Then, the last equation can also be written in the form

$$Q_2'(\theta) = \frac{c_2\mu_2}{2} \sum_{i=1}^{M_1} \left\{ \sum_{k=1}^{D_i} d_{i,k}^2 - 2\Delta_i^+ \right\}. \tag{21}$$

Combining (14), (21) and letting D_i be the number of overlapping intervals in queue 1's ith NBP, we obtain:

$$Q'(\theta) = \frac{1}{T} \sum_{i=1}^{M_1} \frac{-c_1\mu_1 L_{1,i}^2 + \sum_{k=1}^{D_i} c_2\mu_2 d_{i,k}^2}{2} - c_2\mu_2\Delta_i^+. \tag{22}$$

We can now establish our main result as follows.

Theorem 1. If $c_1\mu_1 \ge c_2\mu_2$, then $Q'(\theta) \le 0$. Moreover, if $c_1\mu_1 > c_2\mu_2$, then $Q'(\theta) < 0$.

The optimality of the $c\mu$-rule in this case is a direct implication of the theorem. If $c_1\mu_1 > c_2\mu_2$, then $Q'(\theta) < 0$ and the minimum of $Q(\theta)$ is attained at $\theta^* = 1$, the maximum feasible value of the parameter θ.

IV. Extension to nonlinear costs

In this section, we replace $\sum_{n=1}^{2} c_n x_n(t,\omega)$ in (5) by $q(x_1(t;\omega), x_2(t;\omega)) = c_1 g_1(x_1(t;\omega)) + c_2 g_2(x_2(t;\omega))$ where $g_1(\cdot)$ and $g_2(\cdot)$ are nonlinear functions such that, $g_n(0) = 0$ and $\frac{dg_n(x_n)}{dx_n}$ exists and is positive for $0 \le x_n < \infty$ and $n = 1, 2$. Determining the optimal switching structure and times requires explicitly solving a multipoint boundary value problem which is notoriously hard to solve. To use the IPA approach instead, consider the cost function $Q(\theta) = \frac{1}{T}\int_0^T [c_1 g_1(x_1(t;\theta)) + c_2 g_2(x_2(t;\theta))]dt$. We can interpret θ as the average amount of time during which $\theta(t) = 1$ in a schedule which switches between 1 and 0, or, in the case of the actual underlying queueing system, as the probability of allocating the resource to queue 1. If, for example, we find that θ^* is close to 1 when $c_1\mu_1 > c_2\mu_2$, we can conclude that the the $c\mu$-rule is near-optimal. Now let $Q_n(\theta) = \int_0^T c_n g_n(x_n(t;\theta))dt$ for $n = 1, 2$ and $h_n(x_n(t;\theta)) = \frac{dg_n(x_n(t;\theta))}{dx_n}$ where we assume this derivative is a known function. The sample derivative, then is

$$Q_n'(\theta) = \sum_{i=1}^{M_n} \int_{\xi_{n,i}}^{\eta_{n,i}} c_n x_n'(t;\theta) h_n(x_n(t;\theta))dt. \tag{23}$$

Starting with $n = 1$, we have $x_1'(t;\theta) = -\mu_1(t - \xi_1)$ by (13). Recall that the actual sample paths we can observe are those of the underlying queueing system so that an NEP $[\xi_{1,i}, \eta_{1,i})$ can be partitioned into intervals $[e_{i,p-1}, e_{i,p})$, with $p = 1, \ldots, N_i$, $e_{i,0} = \xi_{1,i}$ and $e_{i,N_i} = \eta_{1,i}$, defined by all queue 1 task arrival and departure events. In other words, in the pth interval $[e_{i,p-1}, e_{i,p})$ the queue content is fixed and given by $x_{i,p} \in \{1, 2, \ldots\}$ and the associated values of $h_1(x_{i,p})$ can be pre-computed for them. Using this information in (23) and after simplifying terms, we get

$$Q_1'(\theta) = -c_1\mu_1 \sum_{i=1}^{M_1} \sum_{p=1}^{N_i} h_1(x_{i,p}) a_{i,p} (b_{i,p} - \xi_{1,i}) \tag{24}$$

where $a_{i,p} = e_{i,p+1} - e_{i,p}$ and $b_{i,p} = \frac{e_{i,p+1} + e_{i,p}}{2}$.
Observe that this IPA derivative, with pre-computed values $h_1(1), h_1(2), \ldots$, is evaluated with minimal computation. It depends only on the event times $e_{i,0}, \ldots, e_{i,N_i}$ within each of the M_1 NEPs of queue 1. Moreover, $Q_1'(\theta)$ *does not depend* on the inflow rates or any probabilistic parameter

of the model and provides an extremely simple sensitivity estimate to be used in standard gradient-based schemes.

The derivation of $Q_2'(\theta)$ is similar, but we first need to partition the NEP of queue 2 into overlapping intervals $[\nu_{j,k}, \upsilon_{j,k}), k = 1, \ldots, D_j$ (if any exist) and then partition the kth such interval into intervals $[e_{k,p-1}, e_{k,p})$, based on queue 2 task arrival and departure events at times $e_{k,p}$, $p = 1, \ldots, N_{j,k}$. Omitting details, we get

$$Q_2'(\theta) = c_2\mu_2 \sum_{j=1}^{M_2} \sum_{k=1}^{D_j} \sum_{p=1}^{N_{j,k}} h_2(x_{k,p}) a_{k,p}(b_{k,p} - \nu_{j,k}) - \Delta_j^+$$

where $a_{k,p} = e_{k,p+1} - e_{k,p}$, $b_{k,p} = \frac{e_{k,p+1}+e_{k,p}}{2}$, and Δ_j was defined in (20).

Finally, let us take a closer look at why and how the $c\mu$-rule may fail when the holding cost function is nonlinear. For $n = 1$, using (13) and integration by parts, (23) gives

$$Q_1'(\theta) = -c_1\mu_1 \sum_{i=1}^{M_1} \left\{ \left[\left(\frac{t^2}{2} - \xi_{1,i} t \right) h_1(x_1(t;\theta)) \right]_{\xi_{1,i}}^{\eta_{1,i}} - \int_{\xi_{1,i}}^{\eta_{1,i}} \left(\frac{t^2}{2} - \xi_{1,i} t \right) \frac{dh_1(x_1(t;\theta))}{dt} dt \right\}. \quad (25)$$

Applying the mean value theorem for integration [18] and doing some algebraic manipulations we get

$$Q_1'(\theta) = \frac{-c_1\mu_1}{2} \sum_{i=1}^{M_1} L_{1,i}^2 h_1(x_1(\sigma_i;\theta))$$

for some $\sigma_i \in [\xi_{1,i}, \eta_{1,i}]$. The same method can be applied to find $Q_2'(\theta)$. We state here the final result for $Q'(\theta)$:

$$Q'(\theta) = \frac{1}{2T} \sum_{i=1}^{M_1} \left\{ - c_1\mu_1 L_{1,i}^2 h_1(x_1(\sigma_i;\theta)) + \sum_{k=1}^{D_i} c_2\mu_2 d_{i,k}^2 h_2(x_2(\tau_{i,k};\theta)) - 2c_2\mu_2\Delta_i^+ \right\} \quad (26)$$

for some $\tau_{i,k}$ in an overlapping interval of length $d_{i,k}$ and

$$\Delta_i = (\xi_{2,i^*} - \xi_{1,i}) \int_{\eta_{1,i}}^{\eta_{2,i^*}} h_2(x_2(t;\theta)) dt$$

where i^* is defined as before. One can easily see that when $g_n(x_n(t;\theta)) = x_n(t;\theta)$ for $n = 1, 2$, (26) reduces to (22) since $h_n(t;\theta) = 1$. A closer look at (26) suggests that when $\sum_{i=1}^{M_1} \left[-L_{1,i}^2 h_1(x_1(\sigma_i;\theta)) + \sum_{k=1}^{D_i} d_{i,k}^2 h_2(x_2(\tau_{i,k};\theta)) \right]$ is large enough, even having $c_1\mu_1 > c_2\mu_2$ cannot guarantee the negativity of $Q'(\theta)$ thereby violating the $c\mu$-rule. Such a situation may arise when $h_2(\tau_{i,k};\theta)$ becomes very large for some non-overlapping interval. This typically may occur when $\theta = 1$ since it may cause queue 2 to build-up a large content before queue 1 becomes non-empty (consider Fig. 3 for $\theta = 1$), thereby having $x_2(\tau_{i,k};\theta) \gg x_1(\sigma_i;\theta)$ which may lead to having $h_1(x_1(\sigma_i;\theta)) \ll h_2(\tau_{i,k};\theta)$ for some choices of $h_1(.)$ and $h_2(.)$. Aside from this, it is worth noting that operating at $\theta = 1$ makes the possibility of having a drift smaller which can even make the probability of $Q'(\theta) > 0$ larger.

V. Conclusions

We have considered a classic scheduling problem with a single resource shared by two competing queues in the context of SFMs and shown that the $c\mu$-rule is optimal using simple calculus of variations arguments on a sample path basis as well as through IPA, which also provides explicit sample derivatives of the cost function with respect to a controllable parameter in the scheduling policy. When the cost function is nonlinear in the queue contents, IPA provides a simple, distribution-free estimate of the cost function with respect to a controllable parameter. Further, it provides insights to why the $c\mu$-rule no longer applies. The use of SFMs and IPA opens up a spectrum of possibilities for studying complex stochastic scheduling problems without having to resort to explicit probabilistic models.

References

[1] W. E. Smith, "Various optimizers for single-stage production," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 59–66, 1956.

[2] D. R. Cox and W. L. Smith, *Queues*. London: Methuen, 1961.

[3] J. S. Baras, A. J. Dorsey, and A. M. Makowski, "Two competing queues with linear costs and geometric service requirements: The μc-rule is often optimal," *Adv. Appl. Prob.*, no. 17, pp. 186–209, 1985.

[4] C. Buyukkoc, C. Varaiya, and J. Walrand, "The $c\mu$ rule revisited," *Adv. Appl. Probability*, vol. 17, pp. 237–238, 1985.

[5] T. Hirayama, M. Kijima, and S. Nishimura, "Further results for dynamic scheduling of multiclass G/G/1 queues," *J. Appl. Probability*, vol. 26, pp. 595–603, 1989.

[6] F. Avram, D. Bertsimas, M. Ricard, F. Kelly, and R. Williams, "Fluid models of sequencing problems in open queueing networks: an optimal control approach," in *Stochastic Networks, Proceedings of the IMA*, vol. 71. New York: Springer-Verlag, 1995, pp. 199–234.

[7] H. Chen and D. D. Yao, "Dynamic scheduling of a multiclass fluid network," *Oper. Res.*, vol. 41, no. 6, pp. 1104–1115, 1993.

[8] J. F. C. Kingman, "The single server queue in heavy traffic," in *Proc. Camb. Phil. Soc.*, vol. 57, 1961, pp. 902–904.

[9] W. Whitt, "Weak convergence theorems for queues in heavy traffic," Ph.D. dissertation, Cornell University (Technical Report No. 2, Department of Operations Research, Stanford University.), 1968.

[10] J. M. Harrison, "Brownian models of queueing networks with heterogeneous customers," in *Proc. IMA Workshop on Stochastic Differential Systems*, 1986.

[11] J. A. V. Mieghem, "Dynamic scheduling with convex delay costs: The generalized cμ rule," *Ann. Appl. Probability*, vol. 5, no. 3, pp. 809–833, 1995.

[12] C. G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C. G. Panayiotou, "Perturbation analysis for on-line control and optimization of stochastic fluid models," *IEEE Trans. on Automatic Control*, vol. AC-47, no. 8, pp. 1234–1248, 2002.

[13] G. Sun, C. G. Cassandras, and C. G. Panayiotou, "Perturbation analysis and optimizatin of stochastic flow networks," *IEEE Trans. on Automatic Control*, vol. AC-49, no. 12, pp. 2113–2128, 2004.

[14] H. Yu and C. G. Cassandras, "Perturbation analysis and feedback control of communication networks using stochastic hybrid models," *J. of Nonlinear Analysis*, vol. 65, no. 6, pp. 1251–1280, 2006.

[15] G. Sun, C. G. Cassandras, and C. G. Panayiotou, "Perturbation analysis of multiclass stochastic fluid models," *J. of Discrete Event Dynamic Systems*, vol. 14, no. 3, pp. 267–307, 2004.

[16] C. G. Cassandras, G. Sun, C. G. Panayiotou, and Y. Wardi, "Perturbation analysis and control of two-class stochastic fluid models for communication networks," *IEEE Trans. on Automatic Control*, vol. 48, no. 5, pp. 770–782, 2003.

[17] C. G. Panayiotou, "On-line resource sharing in communication networks using infinitesimal perturbation analysis of stochastic fluid models," *43rd IEEE Conf. on Decision and Control*, vol. 1, pp. 563–568, 2004.

[18] R. Courant, *Principles of real analysis*. Wiley-IEEE, 1988.