Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule

Jan A. Van Mieghem
Graduate School of Business, Stanford University
October 25, 1994 (Final version of MS #0494-007 Rev. 1)

Abstract

We consider a general single-server multiclass queueing system that incurs a delay cost $C_k(\tau_k)$ for each class k job that resides τ_k units of time in the system. This paper derives a scheduling policy that minimizes the total cumulative delay cost when the system operates during a finite time horizon.

Denote the marginal delay cost and (instantaneous) service rate functions of class k by $c_k = C_k'$ and μ_k, and let $a_k(t)$ be the "age" or time that the oldest class k job has been waiting at time t. We call the scheduling policy that at time t serves the oldest waiting job of that class k with the highest index $\mu_k(t)c_k(a_k(t))$, the *Generalized $c\mu$ Rule*. As a dynamic priority rule that depends on very little data, the Generalized $c\mu$ Rule is attractive to implement. We show that with non-decreasing convex delay costs, the Generalized $c\mu$ Rule is asymptotically optimal if the system operates in heavy traffic, and give explicit expressions for the associated performance characteristics: the delay (throughput time) process and the minimum cumulative delay cost. The optimality result is robust in that it holds for a countable number of classes and several homogeneous servers in a non-stationary, deterministic or stochastic environment where arrival and service processes can be general and interdependent.

1 Introduction

We consider a general single-server multiclass queueing system that incurs a delay cost $C_k(\tau_k)$ for each class k job that resides τ_k units of time in the system. Since queueing theory is the natural paradigm to study dynamic competition for scarce resources, it is interesting to

[0] *AMS 1991 subject classifications:* Primary: 90B35, 90B22. Secondary: 60K25, 60J70, 93E20.

Key words and phrases: Scheduling, production control, queueing systems, dynamic priorities, $c\mu$ rule, heavy traffic limit, asymptotic optimality.

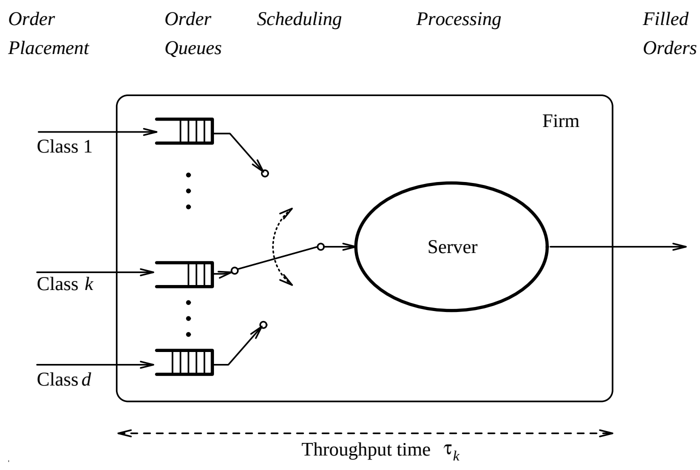

Figure 1: The Scheduling Problem

think of our system as modeling order fulfillment at a firm which dynamically receives orders ("jobs") from customers for several different types or classes of goods and services it provides as shown in Figure 1. In addition to the usual revenue and operating cost associated with filling an order, the firm incurs a delay cost $C_k(\tau_k)$ for each class k order that takes τ_k units of time to fill. (The order fulfillment time τ is also called throughput time, response time or cycle time.) The purpose of the paper is to show how the firm should sequence the different orders that are competing for its scarce resources in order to minimize the total cumulative delay cost during a finite time horizon.

Many providers of goods and services are experiencing an increase in the variety and degree of customization in their customer orders. At the same time, service quality metrics such as order fulfillment time are increasingly important in environments where time performance provides a source of competitive advantage. When facing a delay-sensitive economic environment characterized by a high degree of uncertainty, decisions about allocation of scarce resources to orders can be important to the performance of the firm.

Denote the marginal delay cost function for class k by $c_k = C_k'$. If the functions C_k are linear (and the marginal delay costs constant), the well known $c\mu$ Rule gives the optimal sequence under mild additional assumptions. Denoting by $1/\mu_k$ the average processing time for class k, we associate with each waiting class k job the index $c_k\mu_k$ and at each decision point serve the class with the highest index. (With linear delay costs, it does not matter how jobs are sequenced *within* a class.) Thus, small jobs that are costly to delay are given

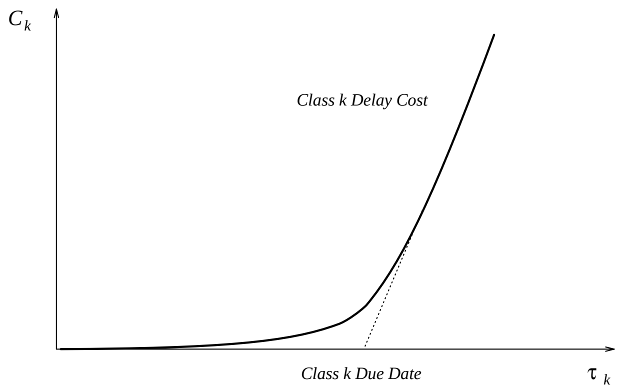

Figure 2: Non-linear Convex Delay Costs.

priority. This static priority rule is robust in that it is optimal in many settings where delay costs are linear. It appears that the optimality of the $c\mu$ Rule was first suggested by Smith [31] for a deterministic, static (that is, all jobs are present at time 0 and no dynamic arrivals are allowed) environment. Cox and Smith [5] seem to be the first to have shown optimality for a stochastic, dynamic (multiclass M/G/1) environment with arbitrary time horizon. The $c\mu$ Rule was also shown to be optimal in stochastic, static settings (e.g., see [24, 25]). Many extensions have been developed. For example, Klimov [18] extended the $c\mu$ Rule to multiclass M/G/1 systems with feedback, Harrison [12] showed optimality of a more complex static priority rule when delay costs are discounted in multiclass M/G/1 systems, and Tcha and Pliska [32] studied the combination of discounting and feedback (again a static priority rule is optimal). More recently, Buyukkoc, Varaiya and Walrand [1] and Hirayama, Kijima and Nishimura [15] have shown that the $c\mu$ Rule also extends to discrete time systems with general arrival patterns and DFR service times, and Nain [23] generalized to continuous time, discounting and partial feedback. De Serres [6] has shown that a $c\mu$ Rule can also arise when scheduling and flow control are optimized simultaneously.

In practice however, delay cost functions are usually non-linear. This non-linearity may stem from physical phenomena (e.g., processing perishable goods or landing fuel-limited aircraft) or, more frequently, from customer expectations. A customer often expects a certain delay or is quoted one in the form of a promised delivery date. The marginal cost to the firm of not meeting the expected delay or due-date is usually much higher than the marginal cost when the customer's expectations are realized, as shown in Figure 2. This cost includes not only traditional holding costs, but also the opportunity cost of future lost sales and other strategic effects such as a decrease of customer good will, market reputation and credibility. Shycon and Sprague [30] show empirical data that out-of-pocket delay costs in the food industry are strongly convex increasing even without taking opportunity costs into account. Chardaire and Lesk [2] argue that packet-switched computer networks are severely constrained in the delay that can be incurred in each node, giving rise to non-linear delay costs. Thadhani [33] presents empirical data that productivity in interactive computing is a non-linear function of computer response time. Other domains where timeliness is important are software development, securities trading, airline reservation systems, banking and communication systems, as discussed by Dewan and Mendelson [7]. Finally, the common practice in manufacturing environments of expediting orders that have been waiting too long—and thus violating the static priority rule—gives empirical evidence that marginal delay costs increase when the delay increases.

Denote the "age" or the time that the oldest class k job has been waiting at time t by $a_k(t)$, and let μ_k be the instantaneous service rate function of class k. We will refer to the scheduling policy that at time t serves the oldest waiting job of that class k with the highest index $\mu_k(t)c_k(a_k(t))$ as the *Generalized $c\mu$ Rule.* This paper will show that with non-decreasing convex delay costs, the Generalized $c\mu$ Rule is "approximately optimal" if the system is "operating near full capacity" and will give explicit expressions for its associated performance characteristics: the delay (throughput time) process and the minimum cumulative delay cost. (These statements will be spelled out and proved in precise mathematical terms in the following sections). The optimality result is robust in that a countable number of classes and several homogeneous servers are allowed in a non-stationary, deterministic or stochastic environment where arrival and service processes can be general and interdependent. The Generalized $c\mu$ Rule is a dynamic or time-dependent priority rule that depends on very little data (service rate and age) and is thus inexpensive and simple to implement. In the presence of due-dates, it shows that the practice of scheduling late orders according to both their lateness penalty and expected processing time is sound.

Among the scheduling research that does address non-linear problems, most studies consider static environments (e.g., see [25, 26, 27, 29, 35]). Veklerov [34] shows that results for static scheduling problems do not necessarily generalize to a dynamic setting. Haji and Newell [9] study the related problem of scheduling two classes with convex delay cost during a "rush hour" in which the arrival rate exceeds the service rate. By "neglecting stochastic effects and justifications of approximations" [9, p. 228], they solve a calculus of variations problem with a two-dimension specific method and arrive also at the policy which we call the Generalized $c\mu$ Rule. Our work generalizes the latter and is different in that it employs a method that is independent of the dimension (the number of classes), incorporates stochastic effects, provides expressions for the delay process and for the lower bound on cumulative delay cost, and shows the optimality of the Generalized $c\mu$ Rule while being explicit about the necessary assumptions.

This paper uses the framework introduced by Harrison [10] that endows a processing network model with dynamic control capability and then takes a "heavy traffic limit". Harrison's paper has started a whole body of research. Like Harrison and Wein [14, 13], Wein [36, 37, 38, 39, 40], Kelly and Laws [17], we have a special structure which is amenable to analysis and yields an explicit dynamic scheduling policy. Like Krichagina *et al.* [19], Kushner *et al.* [20, 21], and Martins *et al.* [22], we give a rigorous proof of optimality (without requiring the same degree of mathematical sophistication for our setting). Our approach differs slightly from this stream of research in that it starts with a deterministic or pathwise analysis and considers a broader class of scheduling control policies.

The paper is organized as follows. In the next section we present our model and discuss our methodology. Section 3 analyzes the model and Section 4 shows the main optimality results of the Generalized $c\mu$ Rule. We conclude in Section 5 with extensions and discussion.

2 Model and Methodology

Consider a general single-server multiclass queueing system that operates during the finite time horizon, $t \in [0, n]$. Jobs arrive at the system and require a service. Jobs are categorized into d (for dimension) different classes depending on their specific arrival patterns, service requests, and time delay sensitivity. A class k job resides in the system for an amount of time τ_k (which consists of actual processing time and waiting delays), inflicts a delay cost $C_k(\tau_k)$ onto the system, and then departs.

The model has three primitives: a d-dimensional arrival process A, a d-dimensional service process S, and a d-dimensional delay cost function C where each component $C_k : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing convex. $A_k(t)$ represents the number of class k jobs that have arrived during

$[0,t]$ and $S_k(t)$ is the number of class k jobs that are served during the first t time units that the server devotes to class k. Construct d sequences of interarrival times $\{u_{k,i} : i \in \mathbb{N}\}$ for $k = 1, \cdots, d$ and a corresponding partial sums process U such that

$$
\begin{aligned}
U_k(j) &= \sum_{i=1}^{\lfloor j \rfloor} u_{k,i} \qquad \text{with} \qquad U_k(0) = 0, && (1)\\
A_k(t) &= \max\{j \in \mathbb{N} : U_k(j) \le t\}. && (2)
\end{aligned}
$$

$U_k(j)$ is the arrival time of the j^{th} class k job. Similarly, one can construct d sequences of service times $\{v_{k,i} : i \in \mathbb{N}\}$ for $k = 1, \cdots, d$ and a corresponding cumulative service process V, where $V_k(j)$ is the total service requirement of the first j class k jobs. For ease of exposition, we assume that the system is empty at time $t = 0$. (Section 5 discusses how to incorporate different initial conditions.)

The objective is to determine a scheduling policy that minimizes the cumulative delay cost function J, possibly at every point in time. Denoting by $\tau_{k,i}$ the time that the i^{th} class k job spends in the system, the cumulative delay cost up to time $t \in [0,n]$ is

$$J(t) = \sum_{k=1}^{d} \sum_{i=1}^{A_k(t)} C_k(\tau_{k,i}). \qquad (3)$$

Introduce any continuous-time process $\tau_k(\cdot)$ with $\tau_k(U_k(i)) = \tau_{k,i}$ so that $\tau_k(t)$ represents the delay of the job that arrived at time t. Then J can be written as

$$J = \sum_k \int C_k(\tau_k(t)) \, \mathrm{d}A_k(t). \qquad (4)$$

In order to proceed we need a representation of a scheduling policy (the decision variable) and a relation that expresses the delay process τ in terms of the primitives. We adapt the processing network model with dynamic control capability introduced by Harrison [10] as follows. A scheduling policy is expressed as a vector allocation process T, where $T_k(t)$ represents the total amount of time during $[0,t]$ that the server allocates to class k. Let $N_k(t)$ denote the total number of class k jobs present in the system at time t, and define the vector headcount process N in the obvious way. We have the fundamental flow identity

$$N_k(t) = A_k(t) - S_k(T_k(t)). \qquad (5)$$

The server may not have enough work to keep him busy at all times, and may conceivably be idle when there is work to do. However, if preemption is allowed, it is optimal to enable the server whenever there is work waiting and not to insert scheduled idleness. Such a scheduling policy is called work conserving[1]. Due to the rather crude nature of the asymptotic analysis of sections 3 and 4, the assumptions made regarding preemption do not affect the scheduling policy that will emerge from the analysis. Define $I(t)$ as the cumulative server idleness up to time t:

$$I(t) = t - \sum_k T_k(t). \tag{6}$$

We assume that the arrival times and the queues are observable, as is usually the case in practice. Thus, the decision maker can base the allocation decision at time t only on the observed evolution of (A, N) up to t. According to (5), this means that only the service times of the *processed* jobs are observable, not those of the waiting jobs. The requirement that T be non-anticipating with respect to (A, N) and its interpretation as a cumulative time allocation translate into the following conditions. Formally, a policy T is *feasible* if

F1. $T(0) = 0$ and $\{T(t), t \in (0, n]\}$ is adapted to the filtration $\{\mathcal{F}_t; t \in (0, n]\}$ where $\mathcal{F}_t = \sigma\{(A(s), N(s)), 0 \leq s < t\}$.

F2. T is continuous and non-decreasing.

F3. I is non-decreasing.

F4. $N \geq 0$.

Define the *workload input* process L and the *workload* process W via

$$\begin{aligned} L_k(t) &= V_k(A_k(t)), && (7) \\ W_k(t) &= L_k(t) - T_k(t). && (8) \end{aligned}$$

$L_k(t)$ represents the total amount of work (expressed in units of time) requested by all the class k jobs that have arrived by time t, and $W_k(t)$ is the amount of work requested by those class k jobs that are in the system at time t. It follows directly that the total work input $L_+ = \sum_k L_k$ and total workload $W_+ = \sum_k W_k$ are independent of the work conserving scheduling policies. Because L is exogenous, one could also use W instead of T to express the scheduling policy.

[1]In general, a policy is said to be work conserving if it does not affect the arrival or service process, and if service is provided whenever the system is not empty.

In order to derive the system equation for the delay process τ we first show that serving each class in first-in first-out (FIFO) order is optimal.

Proposition 1 *FIFO sequencing within a class is optimal in the expected*[2] *sense,* $EJ_{\mathrm{FIFO}} \leq EJ_{\mathrm{not-FIFO}}$*, if class service times are homogeneous and not observable and if the class delay cost function is non-decreasing convex.*

(All proofs are given in the appendix.) Notice that for strictly convex delay functions, FIFO is the *unique* optimal service order. It follows from the definition of T and W that, if each class is FIFO sequenced, the delay process is given by

$$\tau_k(t) = \inf\{s \in \mathbb{R}_+ : W_k(t) \leq T_k(t+s) - T_k(t)\}, \tag{9}$$

or

$$W_k(t) = T_k(t + \tau_k(t)) - T_k(t). \tag{10}$$

Given the generality of the model that does not make any assumptions regarding the arrival and service processes one cannot possibly hope for an exact solution to this problem. Therefore, we will focus on policies that are asymptotically optimal as "the time horizon n becomes large compared to the job delays and the system operates near full capacity". Before we can rephrase this loosely stated condition in precise terms we will need some more analysis. Considering heavily loaded systems is not very restrictive given that the impact of scheduling is greatest when a system is operating close to its capacity constraint.

The methodology that we use to study processing systems operating near full capacity is *heavy traffic* analysis. One considers a *sequence* of systems similar to the one described in this section. The n^{th} system has a time horizon of n, and as n gets large, utilization approaches 1 and the system is operating near full capacity. Because in the limit the jumps of the arrival and service process become negligible, a considerable simplification occurs and the problem becomes analytically tractable.

Nowhere have we made an assumption regarding uncertainty in the arrival and service process primitives. The method is applicable to both deterministic and stochastic settings. In the next section we will analyze our system under heavy traffic without needing any reference to a stochastic environment which allows a more accessible, less technical exposition. We call this the deterministic system, but one could equally well describe it as a sample path

[2] Section 4 shows that FIFO is also asymptotically optimal in the stochastic sense.

analysis or a study of a specific realization of the stochastic system. Section 4 shows how this analysis ties into a stochastic setting.

We will use the following notation. $\mathcal{C}$ denotes the space of continuous real functions on $[0,1]$, $\mathcal{C}^1$ is the space of real functions on $[0,1]$ that have continuous first derivatives, and $\mathcal{D}$ is the space of simply discontinuous functions on $[0,1]$. The functions may be scalar or d-dimensional vector functions, which will be clear from the context. Subscripts denote components of a vector. We write $x^n \to x$ and say that "x^n converges" to mean that the functions $x^n \in \mathcal{D}$ converge to some function $x \in \mathcal{D}$ under the uniform norm

$$\|x\| = \sup_{0 \le t \le 1} |x(t)|, \tag{11}$$

which is interpreted as $\sup_t \max_k |x_k(t)|$ for a vector function. Slightly abusing the notation, we denote a vector function with components $x_k(t)y_k(t), x_k(t)/y_k(t)$, and $x_k(y_k(t))$ at time t by $xy, x/y$, and $x \circ y$ respectively. Finally, the identity function is denoted by $e, e(t) = t$.

3 Deterministic Analysis

This section describes the heavy traffic analysis of our problem. Consider a sequence of systems, indexed by $n \in \mathbb{N}$, similar to the one described in the previous section. The n-th system has an arrival process A^n, service process S^n, and delay cost function C^n as its primitives, and operates during $[0,n]$. The purpose is to derive insight into the effect of a policy on the dynamics of a system that is operating near full capacity. The primitives and policies (T^n) can be different from system to system, but to yield meaningful insights, they cannot be completely unrelated. The requirement to operate near full capacity relates the arrival process and service process within one system and among systems. We will also relate the cost functions of different systems. Finally, we make the problem analytically tractable by imposing convergence assumptions on the arrival and service processes. The analysis makes no reference to a stochastic environment and simplifies to an exercise in real analysis. However, the results will be applicable to both deterministic and stochastic settings.

3.1 Analysis

Convergence assumptions on the arrival and service processes are conveniently stated after a time transform to the common domain $t \in [0,1]$, similar to familiar functional central limit theorems (FCLTs) of stochastic systems. All interarrival and service times are assumed finite in all systems so that the arrival and service processes of the n-th system are of order n. We will show in Proposition 2 that the decision variable T^n is (asymptotically) determined to a first order by the primitives A^n and S^n. Thus a second order analysis is necessary to study a specific control policy. The FCLT for renewal processes states that the second order term is of order $n^{1/2}$. And since the unit-size discontinuities of A^n and S^n are of order $1 = o(n^{1/2})$, it is natural to decompose A^n and S^n into a sum of continuous functions $\bar{A}^n, \bar{S}^n, \tilde{A}^n, \tilde{S}^n$ in $\mathcal{C}$ so that for $t \in [0,1]$:

$$
\begin{aligned}
A^n(nt) &= n\bar{A}^n(t) + n^{1/2}\tilde{A}^n(t) + o(n^{1/2}), && (12)\\
S^n(nt) &= n\bar{S}^n(t) + n^{1/2}\tilde{S}^n(t) + o(n^{1/2}). && (13)
\end{aligned}
$$

One may think of the first and second order terms as the long-term trend and the variation around this trend, respectively. Because A^n and S^n are non-decreasing, we can always require same of their continuous first order terms $\bar{A}^n$ and $\bar{S}^n$ so that the inverse functions $\bar{A}^{n^{-1}}$ and $\bar{S}^{n^{-1}}$ exist. Introduce the following functions

$$R_k^n = \bar{S}_k^{n^{-1}} \circ \bar{A}_k^n \qquad \text{and} \qquad R_+^n = \sum_k R_k^n. \tag{14}$$

We will show that R_k^n is the first order approximation of the work input process L^n, so that the n-th system operates near full capacity if R_+^n is close to the identity function.

Assumption 1 (Main Convergence) *There exist functions $\tilde{A}^*, \tilde{S}^*, \tilde{c}^* \in \mathcal{C}$ and increasing functions $\bar{A}^*, \bar{S}^* \in \mathcal{C}^1$, such that:*

$$
\begin{aligned}
\bar{A}^n &\to \bar{A}^* & \bar{S}^n &\to \bar{S}^*, && (15)\\
\tilde{A}^n &\to \tilde{A}^* & \tilde{S}^n &\to \tilde{S}^*, && (16)
\end{aligned}
$$

$$n^{1/2}(R_+^n - e) \to \tilde{c}^*. \tag{17}$$

Equation (17) is the heavy traffic condition stating that for large n, the system is operating near full capacity. Denote the positive first derivatives by

$$
\begin{aligned}
\bar{A}^{*'} &= \lambda, && (18)\\
\bar{S}^{*'} &= \mu, && (19)\\
R_k^{*'} &= \rho_k = \lambda_k/\mu_k, && (20)
\end{aligned}
$$

which are all bounded on $[0,1]$ because they are continuous. λ, μ and ρ represent the (asymptotic) scaled instantaneous arrival rate, service rate, and traffic intensity of class k. The main assumptions imply convergence relations for all other system variables:

Proposition 2 *Given Assumption 1, we have that for any scheduling policy*

$$\begin{aligned} N^n(nt) &= n^{1/2}\tilde{N}^n(t) + o(n^{1/2}), && (21)\\ T^n(nt) &= n\bar{T}^n(t) + n^{1/2}\tilde{T}^n(t) + o(n^{1/2}), && (22)\\ U^n(nt) &= n\bar{U}^n(t) + n^{1/2}\tilde{U}^n(t) + o(n^{1/2}), && (23)\\ V^n(nt) &= n\bar{V}^n(t) + n^{1/2}\tilde{V}^n(t) + o(n^{1/2}), && (24)\\ W^n(nt) &= n^{1/2}\tilde{W}^n(t) + o(n^{1/2}), && (25) \end{aligned}$$

and for FIFO sequencing in each class

$$\tau^n(nt) = n^{1/2}\tilde{\tau}^n(t) + o(n^{1/2}), \qquad (26)$$

with the following convergence relationships

$$\begin{aligned} \bar{T}^n &\to R^* \in \mathcal{C}^1, && (27)\\ \bar{U}^n &\to \bar{U}^* \in \mathcal{C}^1, && (28)\\ \bar{V}^n &\to \bar{V}^* \in \mathcal{C}^1, && (29)\\ \tilde{U}^n &\to \tilde{U}^* \in \mathcal{C}, && (30)\\ \tilde{V}^n &\to \tilde{V}^* \in \mathcal{C}, && (31)\\ \tilde{W}^n_+ &\to \tilde{W}^*_+ \in \mathcal{C}, && (32)\\ n^{-1/2} \sup_{1\le i\le A^n(n\cdot)} u^n_{k,i} &\to 0, && (33)\\ n^{-1/2} \sup_{1\le i\le S^n(n\cdot)} v^n_{k,i} &\to 0, && (34) \end{aligned}$$

$$\tilde{W}^n \text{ converges} \iff \tilde{T}^n \text{ converges} \iff \tilde{N}^n \text{ converges} \iff \tilde{\tau}^n \text{ converges}, \qquad (35)$$

and where $\limsup_n \|\tilde{N}^n\|, \limsup_n \|\tilde{T}^n\|, \limsup_n \|\tilde{W}^n\|, \limsup_n \|\tilde{\tau}^n\|$ *are all bounded.*

Since counting processes and partial sums processes are almost inverse processes, the convergence relationships for U^n and V^n are not surprising. Equation (27) shows that the decision variable T^n is asymptotically known to a first order as argued intuitively by Harrison [10].

Also, the scaled *total* workload $\tilde{W}_+^n$ converges, but that need not be true for the class workload process $\tilde{W}^n$. Moreover, the convergence of the class workload processes implies the convergence of the second order policy process $\tilde{T}^n$, the headcount process $\tilde{N}^n$, and the delay process $\tilde{\tau}^n$. Non-converging policies are not an esoteric mathematical artifact; they can represent scheduling policies that are widely used in practice. For example, polling systems where the different classes are served until exhaustion in a specific order have non-converging class workload processes as discussed by Coffman, Reiman and Puhalskii [4]. The underlying reason is that, in heavy traffic, the class workload process lives on a smaller time scale than the total workload process. Unlike other researchers [13, 14, 17, 19, 20, 21, 22, 36, 37, 38, 39, 40] who define the asymptotic policy *a priori* as an RCLL function (that is, an element of $\mathcal{D}$), we study a broader class of control policies that includes nonconvergent policies.

The Law of Large Numbers applied to the workload process yields that class workloads are well approximated by the product of class headcount and asymptotic service requirement:

Proposition 3 (LLN) *Given Assumption 1,*

$$\mu_k \tilde{W}_k^n - \tilde{N}_k^n \to 0. \tag{36}$$

Little's Law, relating time-averages of the delay, arrival and headcount process, generalizes:

Proposition 4 (Little's Law) *Given Assumption 1, and $a, b \in [0,1]$ where $a < b$,*

$$\frac{n^{-3/2}}{\bar{A}_k^n(b) - \bar{A}_k^n(a)} \int_{na}^{nb} \tau_k^n \mathrm{d}A_k^n - \frac{1}{\bar{A}_k^n(b) - \bar{A}_k^n(a)} \int_a^b \tilde{N}_k^n(t)\mathrm{d}t \to 0. \tag{37}$$

Delay cost functions in a system are defined in terms of the natural time scale of throughput times. Because the n-th system has throughput times of order $n^{1/2}$, its delay cost functions will assign a moderate cost to delays of this order. To investigate the asymptotic behavior of costs, we therefore assume the following:

Assumption 2 (Cost Convergence) *The (vector) cost functions C^n in different systems scale to a non-decreasing convex function C^* as*

$$C^n(n^{1/2}\cdot) \to C^*(\cdot). \tag{38}$$

Therefore, the total cumulative cost $J^n(nt)$ is of order n, and we define the scaled cumulative cost $\tilde{J}^n$ as

$$\tilde{J}^n(t) = n^{-1} J^n(nt) = \sum_k \int_0^{nt} C_k^n(\tau_k^n) n^{-1}\, \mathrm{d}A_k^n \qquad \text{for } t \in [0,1]. \tag{39}$$

3.2 Converging Policies

If the sequence of policies $\tilde{T}^n$ is convergent, then so are $\tilde{W}^n$, $\tilde{N}^n$ and $\tilde{\tau}^n$, according to Proposition 2, and we denote their corresponding limiting functions by $\tilde{W}^*$, $\tilde{N}^*$ and $\tilde{\tau}^*$. Propositions 3 and 4 directly yield the following:

Proposition 5 *Given Assumptions 1 & 2, if the sequence of policies $\tilde{T}^n$ converges, then*

$$\lambda\tilde{\tau}^* = \tilde{N}^*, \tag{40}$$

$$\mu\tilde{W}^* = \tilde{N}^*, \tag{41}$$

and the corresponding sequence of cumulative cost functions converges:

$$\tilde{J}^n \to \sum_k \int \lambda_k(t) C_k^*(\tilde{\tau}_k^*(t))\mathrm{d}t. \tag{42}$$

The convergence in (42) follows directly from the Generalized Lebesgue Convergence Theorem [28, p. 270] because $\tilde{\tau}^*$ is bounded. Thus, if the policies converge, there exists a *limiting system* in which, according to Proposition 5, throughput times are proportional to workloads (*i.e.*, Little's Law holds at each *point* in time). However, an exclusive analysis of the limiting system precludes the consideration of non-converging control sequences which may have a superior performance and can be important in practice.

4 Asymptotic Optimality

In this section, we first present a closed-form, asymptotic lower bound on the scaled cumulative cost function of any feasible policy, converging or not. Then we introduce a family of policies whose asymptotic cumulative cost function attains the lower bound for all times t simultaneously. These policies, which include the Generalized $c\mu$ Rule, are called "asymptotically optimal" and we give an expression for their associated delay process. Finally, we show how these results extent to stochastic systems.

4.1 An Asymptotic Lower Bound on the Cost $\tilde{J}^n$

Define the mapping $g : \mathcal{D} \to \mathcal{D}^d$ such that $y \to g \circ y$ where $g \circ y(t)$ is the solution of the minimization problem

$$g \circ y(t) = \arg\min_{x \in \Omega} \sum_{k=1}^{d} \lambda_k(t) C_k^* \left(\frac{x_k}{\rho_k(t)} \right), \tag{43}$$

where $\Omega = \{x \in R^d_+ : \sum_k x_k = y(t)\}$. It will be shown later that the mapping g applied to the total workload process yields the optimal class workloads, $\tilde{W}^* = g \circ \tilde{W}^*_+$. Because the objective function is convex on the convex set Ω, the solution set is also convex. If C^* is convex increasing, the solution is unique and g is continuous at any continuous y. (If C^* is non-decreasing convex, there can be an uncountable set of solutions, but we can pick a particular solution such that g is continuous). We can now show the following lower bound.

Proposition 6 *Given Assumptions 1 & 2, the asymptotic cost is bounded from below, that is, for any sequence of feasible policies, the associated sequence of cumulative costs $\{\tilde{J}^n : n \in \mathbb{N}\}$ satisfies for each $t \in [0,1]$*

$$\liminf_{n\to\infty} \tilde{J}^n(t) \geq \tilde{J}^*(t), \tag{44}$$

where

$$\tilde{J}^* = \sum_k \int \lambda_k(t) C^*_k \left(\frac{[g \circ \tilde{W}^*_+]_k(t)}{\rho_k(t)} \right) \mathrm{d}t. \tag{45}$$

Notice that the lower bound depends only on the instantaneous rates λ and μ and variability, reflected by the second order "tilde processes", affects $\tilde{J}^*$ only through the total workload process. The following section will show that the bound is tight.

4.2 Asymptotically Optimal Scheduling Rules

From the expression of the lower bound $\tilde{J}^*$ and Proposition 5, it follows that any sequence of policies that controls the class workloads such that $\tilde{W}^n \to g \circ \tilde{W}^*_+$ is asymptotically optimal. Thus, if we approximate W_k by $\mu_k^{-1} N_k$ (Proposition 3), then "serving to hug the optimal workload curve" as shown in Figure 3 is a feasible and asymptotically optimal policy.

Another way to attain the lower bound is to control according to the first order optimality conditions of the minimization problem (43) if C^* is smooth, *i.e.*, $C^* \in \mathcal{C}^1$. Denoting the derivative (gradient) of C^* by c^*, the marginal cost function of C^*, the Kuhn-Tucker optimality conditions are sufficient because the objective function is convex, and for each fixed $t \in [0,1]$, the solution $x^* = g \circ y(t)$ solves

$$\mu_k(t) c^*_k \left(\frac{x^*_k}{\rho_k(t)} \right) - \alpha_k = \alpha_0 \tag{46}$$

$$\alpha_k x^*_k = 0 \tag{47}$$

$$\sum_k x^*_k = y(t) \tag{48}$$

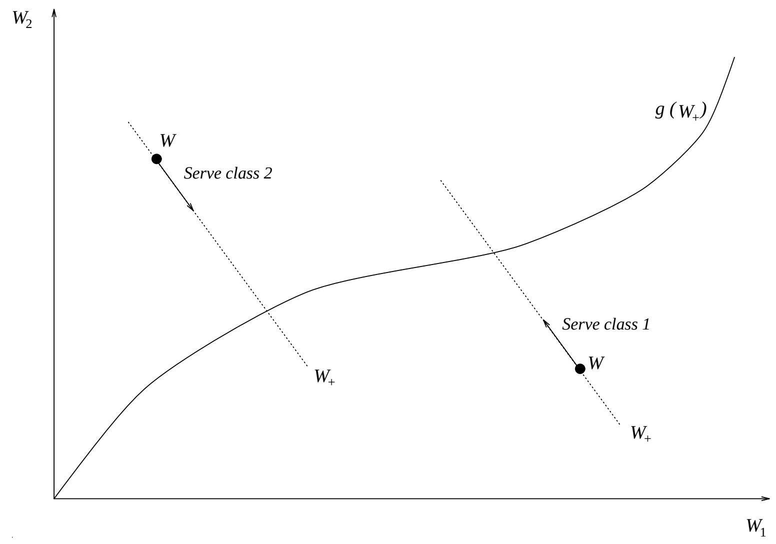

Figure 3: “Hug-the-curve” Scheduling

where the Lagrange multipliers $\alpha_k \geq 0$ and $\alpha_0 \in \mathbb{R}$, and $1 \leq k \leq d$. In the general case, there can be boundary solutions, *i.e.*, x^* belongs to the boundary $\bigcup_{k=1}^{d}\{x_k = 0\}$ of the set Ω, or an uncountable set of solutions. To keep the exposition simple, we will assume that the solution x^* is *unique* and *interior* for all $t \in [0,1]$, which is guaranteed by the following regularity assumption:

Assumption 3 (Cost Regularity) *The (vector) cost function C^* is strictly convex, smooth ($C^* \in \mathcal{C}^1$), and has an interior solution to the minimization problem (43).*

Under Assumption 3, the sufficient conditions reduce to

$$\mu_k(t)c_k^*\left(\frac{x_k^*}{\rho_k(t)}\right) = \alpha_0 \tag{49}$$

$$\sum_k x_k^* = y(t). \tag{50}$$

Proposition 7 *Given Assumptions 1, 2 & 3, the sequence of feasible policies $\{\tilde{T}^n : n \in \mathbb{N}\}$ such that*

$$\max_{1\leq k,l\leq d}\left\|\mu_k c_k^*\left(\frac{\tilde{W}_k^n}{\rho_k}\right) - \mu_l c_l^*\left(\frac{\tilde{W}_l^n}{\rho_l}\right)\right\| \to 0, \tag{51}$$

where $\tilde{W}^n = \tilde{L}^n - \tilde{T}^n$, is asymptotically optimal, that is, the associated sequence of cumulative costs $\{\tilde{J}^n : n \in \mathbb{N}\}$ attains the lower bound

$$\tilde{J}^n \to \tilde{J}^*, \tag{52}$$

and the associated sequence of delay processes $\{\tilde{\tau}^n : n \in \mathbb{N}\}$ *satisfies*

$$\tilde{\tau}^n \to \tilde{\tau}^* = \frac{g \circ \tilde{W}^*_+}{\rho}. \tag{53}$$

The proof of the proposition shows that this sequence of policies is necessarily convergent so that according to proposition 5 the asymptotic optimal scheduling rule implies that

$$\max_{1 \le k,l \le d} \|\mu_k c_k^*(\tilde{\tau}_k^n) - \mu_l c_l^*(\tilde{\tau}_l^n)\| \to 0. \tag{54}$$

Serving the class k with highest $\mu_k c_k^*(\tilde{\tau}_k^n)$ increases the $\mu_l c_l^*(\tilde{\tau}_l^n)$ thereby lowering the maximum difference among the classes. Because the difference between the age of the oldest job and its delay becomes negligible for large n, the Generalized $c\mu$ Rule implements precisely an asymptotic optimal scheduling policy. Since we have shown that both the Generalized $c\mu$ Rule and "hug-the-curve" scheduling are asymptotically optimal, they are essentially equivalent. The former provides a concise mathematical representation for any number d of classes, while the latter has an attractive pictorial form, especially if $d = 2$ (although it carries over to higher dimensions).

Recall that we have assumed that the minimization problem (43) has a unique interior solution for all t. In general, there can be boundary solutions, such that for some $i \in \{1, \cdots, d\}$, the solution $x^* = g \circ y(t)$ has $x_i^* = 0$. This means that we should schedule these classes such that $\tilde{W}_i^n \to 0$, and the remaining classes k, l according to (51). $\tilde{W}_i^n \to 0$ implies that "boundary" classes i should be given priority above "interior" classes k, l. Under heavy traffic conditions, it is irrelevant how the ranking is done among the boundary classes because their queue lengths will be negligible compared to those of the interior classes. Therefore, serving them according to the Generalized $c\mu$ Rule is also asymptotically optimal and scheduling the highest cost generating class first remains intuitively attractive. Finally, because $\mu_i c_i^*(0) > \mu_k c_k^*(x_k^*/\rho_k)$ for any boundary class i and interior class k, serving *all* classes according to the Generalized $c\mu$ Rule is an asymptotic optimal strategy (regardless whether the optimal point is interior or on the boundary).

4.3 Stochastic Systems

Now embed the analysis in a probabilistic structure. We are given a sequence of stochastic systems defined on a corresponding sequence of probability spaces $\{(\Omega^n, \mathcal{F}^n, P^n) : n \in \mathbb{N}\}$.

We write $X^n \Longrightarrow X$ to denote weak convergence of random elements $X^n \in \mathcal{D}$ to $X \in \mathcal{D}$ in the space $\mathcal{D}$ under the Skorohod topology. All limiting functions X in this paper will be continuous, in which case convergence under the Skorohod metric is equivalent to uniform convergence (*i.e.*, convergence under the norm $\|\cdot\|$ of (11)) according to Glynn [8, Proposition 4, p. 149]. In that case, invoking the Skorohod Representation theorem [8], the above analysis holds for almost all sample paths in the Skorohod space. Relating these results to the original system sequence immediately yields the following proposition.

Proposition 8 *Given Assumptions 2 & 3, if there exist processes $\tilde{A}^*, \tilde{S}^*, \tilde{c}^*$ with a.s. continuous sample paths on $[0,1]$ and processes $\bar{A}^*, \bar{S}^*$ with a.s. continuously differentiable increasing sample paths on $[0,1]$, such that:*

$$\left(\bar{A}^n, \tilde{A}^n, \bar{S}^n, \tilde{S}^n, n^{1/2}(R_+^n - e)\right) \Longrightarrow \left(\bar{A}^*, \tilde{A}^*, \bar{S}^*, \tilde{S}^*, \tilde{c}^*\right), \tag{55}$$

then the asymptotic cost is stochastically bounded from below, that is, for any feasible policy, the associated sequence of cumulative costs $\{\tilde{J}^n : n \in \mathbb{N}\}$ satisfies for each $t \in [0,1]$

$$\liminf_{n\to\infty} \tilde{J}^n(t) \geq_{st} \tilde{J}^*(t), \tag{56}$$

and the sequence of feasible policies $\{\tilde{T}^n : n \in \mathbb{N}\}$ such that

$$\max_{1\leq k,l\leq d} \left\| \mu_k c_k^* \left(\frac{\tilde{W}_k^n}{\rho_k}\right) - \mu_l c_l^* \left(\frac{\tilde{W}_l^n}{\rho_l}\right) \right\| \Longrightarrow 0, \tag{57}$$

where $\tilde{W}^n = \tilde{L}^n - \tilde{T}^n$, is asymptotically optimal in the stochastic sense, that is, the associated sequence of cumulative costs $\{\tilde{J}^n : n \in \mathbb{N}\}$ attains the lower bound

$$\tilde{J}^n \Longrightarrow \tilde{J}^*, \tag{58}$$

and the associated sequence of delay process $\{\tilde{\tau}^n : n \in \mathbb{N}\}$ satisfies

$$\tilde{\tau}^n \Longrightarrow \tilde{\tau}^* = \frac{g \circ \tilde{W}_+^*}{\rho}. \tag{59}$$

Proposition 8 applies directly to multiclass GI/G/1 systems with independent renewal arrival and service processes. Similar to "classical" heavy traffic scaling, set $\bar{A}^n(t) = \lambda t + n^{-1/2}\tilde{c}^*(t)$, $\tilde{A}^n(t) = n^{-1/2}(A^n(nt) - \lambda nt) - \tilde{c}^*(t)$, and require $\tilde{c}^*(t) = \gamma t$ for a real constant vector γ. The functional strong law and central limit theorem for renewal processes state that $\bar{A}^*$ is the deterministic linear function λe and $\tilde{A}^*$ is a Brownian motion with drift γ, likewise for the service process, so that the assumptions of Proposition 8 are satisfied. However, the proposition is much more general and also applies to non-stationary systems with dependent arrival and service processes (that satisfy the joint FCLT in (55)).

5 Extensions and Discussion

If the system is not empty at time $t = 0$, the analysis needs to be extended. The initial data add a fourth primitive to the model: for each class k, the number of jobs present at time 0 together with their age and service times: $\{-U_k(i), v_{k,i} : i = -1, \cdots, -N_k(0)\}$. The jobs present at time $t = 0$ represent an additional delay cost J_{ini}:

$$J_{ini} = \sum_{k=1}^{d} \sum_{i=1}^{N_k(0)} C_k(\tau_{k,-i}). \tag{60}$$

Assumption 1 is extended with: *There exist a vector $\tilde{W}^*(0) \in \mathbb{R}^d_+$ such that for the initial data and $k = 1, \cdots, d$:*

$$n^{-1/2} \sum_{i=1}^{N_k^n(0)} v_{k,-i}^n \;\rightarrow\; \tilde{W}_k^*(0), \tag{61}$$

$$n^{-1} \sum_{i=1}^{N_k^n(0)} C_k^n(n^{1/2} + U_k^n(-i)) \;\rightarrow\; 0. \tag{62}$$

The last assumption guarantees that the additional delay cost J_{ini}^n becomes negligible compared to the cumulative delay cost J^n for large n. The only impact of the initial conditions is in providing an initial workload condition $\tilde{W}^*(0)$ which influences the lower bound $\tilde{J}^*$ through the initial total workload $\tilde{W}_+^*(0)$.

The Generalized $c\mu$ Rule also extends to mildly time-dependent delay functions. As long as the delay functions do not vary substantially over a delay period (that is, a time period of order $n^{1/2}$ for a system with time horizon n), the analysis still applies.

In addition, the Generalized $c\mu$ Rule is asymptotically optimal for a multiclass system with multiple parallel servers with equal capabilities. In heavy traffic, the multiserver simplifies to a single server with service capacity equal to the sum of the parallel servers, and the analysis still applies.

The Generalized $c\mu$ Rule is a myopic or greedy rule. Assume the system has Poisson arrival and service processes. If one serves a class k job with age a during $[t, t + \Delta]$, the probability of its service completion during that interval is $\mu_k \Delta + o(\Delta)$. The reduction in cost would be $C_k(a + \Delta) - C_k(a) = c_k(a)\Delta + o(\Delta)$, so that serving class k would decrease the total expected delay cost by $\mu_k c_k(a)\Delta^2 + o(\Delta^2)$. A greedy minimization approach is to serve the job with highest index $\mu_k c_k(a)$.

Similar to the $c\mu$ Rule, the Generalized $c\mu$ Rule requires very little input data: only service rates and age. Interestingly, no arrival data nor higher moments of the service distribution are needed. In this sense it resembles scheduling rules derived from fluid models such as discussed by Chen and Yao [3]. On the other hand, as in diffusion models, variability influence shows up in the expression of the optimal total cumulative cost and associated throughput time process.

The fact that the Generalized $c\mu$ Rule applies to non-stationary and finite horizon settings makes the model particularly relevant to current economic environments where notions such as infinite time horizon, stationarity, and long-run average costs become almost irrelevant. Chen and Yao [3] argue that it is only natural (as well as practical) in that case to follow a policy generated by a myopic procedure, which is reminiscent of a rolling horizon method.

The Generalized $c\mu$ Rule is also pertinent in the presence of due-dates where typically the marginal delay cost strongly increases past the quoted due-date (Figure 2). Our model could be used for instance to study the effects of quoting different due-dates for different "grades of service" where one would offer a product at multiple prices representing a promised faster due-date.

Another factor that should be considered in relation to the Generalized $c\mu$ Rule is the empirical estimation/quantification of the delay costs. Also, the Generalized $c\mu$ Rule is shown to be asymptotically optimal. From a theoretical point of view, it would be interesting to investigate how the rule performs when operating with plenty of excess capacity, although in practice scheduling matters most when resources are scarce and constrained. Therefore, the fact that the Generalized $c\mu$ Rule is "only" asymptotically optimal should not diminish its applicability.

Acknowledgement: I gratefully acknowledge J. Michael Harrison for suggesting this research problem and for his helpful advice on both the technical and expository aspects of this paper. I also thank Elizabeth Schwerer and the two anonymous referees for their thoughtful comments on an earlier version.

A Proofs

A.1 Proposition 1

Proof: Assume class k is not ordered FIFO at time t. Then there is at least one class k job (say the j-th) that arrived at time $U_k(j) \leq t$ which is scheduled before the i-th class-k job which arrived at $U_k(i) < U_k(j)$. We will show that interchanging these two jobs cannot increase the expected cumulative cost $\mathrm{E}(J(n) - J(t))$.

Interchanging the two jobs can only affect the delays of the two jobs and of those jobs currently scheduled between them. In addition, if class k service times are homogeneous and not observable, we cannot distinguish a priori between the service times of i and j, and interchanging the two jobs can therefore not change the a priori estimate of (and thus the expected) delay cost of those jobs currently scheduled between them. Denote the change in cumulative cost due to the interchange by $\Delta J(n) = J_{\text{changed}}(n) - J_{\text{original}}(n)$. Also, denote the total service requirements of all jobs originally scheduled before job j and all jobs in between j and i by v^{before} and v^{between} respectively. Then

$$\begin{aligned}
\mathrm{E}\Delta J(n) &= \mathrm{E}\Big[C_k(v^{\text{before}} + t - U_k(i) + v_{k,i}) + C_k(v^{\text{before}} + v_{k,i} + v^{\text{between}} + t - U_k(j) + v_{k,j}) \\
&\quad - C_k(v^{\text{before}} + t - U_k(j) + v_{k,j}) - C_k(v^{\text{before}} + v_{k,j} + v^{\text{between}} + t - U_k(i) + v_{k,i})\Big] \\
&= \mathrm{E}\Big[C_k(v^{\text{before}} + t - U_k(i) + v_{k,l}) + C_k(v^{\text{before}} + v_{k,m} + v^{\text{between}} + t - U_k(j) + v_{k,l}) \\
&\quad - C_k(v^{\text{before}} + t - U_k(j) + v_{k,l}) - C_k(v^{\text{before}} + v_{k,m} + v^{\text{between}} + t - U_k(i) + v_{k,l})\Big].
\end{aligned}$$

Because C_k is non-decreasing convex on $\mathbb{R}_+$, we have that for any $x, y, z \in \mathbb{R}_+$ with $x \neq y$:

$$\frac{C_k(x) - C_k(y)}{x - y} \leq \frac{C_k(x+z) - C_k(y+z)}{x - y}.$$

Set $x = v^{\text{before}} + t - U_k(i) + v_{k,l}$, $y = v^{\text{before}} + t - U_k(j) + v_{k,l}$ and $z = v^{\text{between}} + v_{k,m}$. By assumption, $x - y > 0$ such that $\mathrm{E}\Delta J(n) \leq 0$. This remains true for any other time t while i and j are in the system. Therefore, $\mathrm{E}J_{\text{FIFO}} \leq \mathrm{E}J_{\text{not-FIFO}}$. ∎

A.2 Proposition 2

Proof: From Assumption 1 that $\bar{A}^*$ ($\bar{S}^*$) is increasing, we can infer that the associated time-scaled arrival (service) epochs $\{n^{-1}U^n(i) : i \in \mathbb{N}\}$ become dense in $[0, 1]$, and thus

$$n^{-1} \sup_{1 \leq i \leq A^n(n)} u^n_{k,i} \rightarrow 0, \tag{63}$$

$$n^{-1} \sup_{1 \leq i \leq S^n(n)} v^n_{k,i} \rightarrow 0. \tag{64}$$

Therefore, using $t - \sup_{1\le i\le A^n(t)} u^n_{k,i} < U^n_k(A^n_k(t)) \le t$ (and likewise for V and S), we have that a counting process and its associated partial sums process are (asymptotically) inverse processes:

$$n^{-1}U^n \circ A^n \circ n\,e \quad\to\quad e, \tag{65}$$

$$n^{-1}V^n \circ S^n \circ n\,e \quad\to\quad e. \tag{66}$$

Thus, the convergence main assumption imply[3] the expansions (23) and (24) where the limits can be expressed in terms of the limits of the associated counting processes

$$\begin{aligned}
\bar U^n &\to \bar U^* = \bar A^{*^{-1}},\\
\bar V^n &\to \bar V^* = \bar S^{*^{-1}},\\
\tilde U^n &\to \tilde U^* = -\frac{\tilde A^* \circ \bar A^{*^{-1}}}{\lambda \circ \bar A^{*^{-1}}},\\
\tilde V^n &\to \tilde V^* = -\frac{\tilde S^* \circ \bar S^{*^{-1}}}{\mu \circ \bar S^{*^{-1}}}.
\end{aligned}$$

Notice that $\bar U^*, \bar V^* \in \mathcal{C}^1$ and $\tilde U^*, \tilde V^* \in \mathcal{C}$. Moreover, we can choose

$$\bar U^n = \bar A^{n^{-1}} \qquad \text{and} \qquad \bar V^n = \bar S^{n^{-1}}. \tag{67}$$

From (7), it follows that L^n has the following expansion:

$$L^n(nt) = n\bar L^n(t) + n^{1/2}\tilde L^n(t) + o(n^{1/2}), \tag{68}$$

where (using $\bar V^{*'}(\cdot) = 1/\mu(\bar S^{*^{-1}}(\cdot) \in \mathcal{C})$

$$\bar L^n = R^n \quad\to\quad R^*, \tag{69}$$

$$\tilde L^n \quad\to\quad \tilde L^* = \left(\bar V^{*'} \circ \bar A^*\right) \tilde A^* + \tilde V^* \circ \bar A^*. \tag{70}$$

From (17) and (69) it follows that the total workload netflow process $X^n = L^n_+ - e$ is of order $n^{1/2}$:

$$X^n(nt) = n^{1/2}\tilde X^n(t) + o(n^{1/2}), \tag{71}$$

where

$$\tilde X^n = \tilde L^n_+ + n^{1/2}(R^n_+ - e) \to \tilde X^* = \tilde L^*_+ + \tilde c^*. \tag{72}$$

From the continuity of the reflection mapping ϕ (cf., Harrison [11]) it follows that the total workload process $W^n_+ = \phi(X^n)$ has expansion

$$W^n_+ = n^{1/2}\tilde W^n_+ + o(n^{1/2}), \tag{73}$$

[3] It follows directly that $n^{-1}U^n \circ nA^* \circ e \to e$ and thus $U^n(n\cdot) = n\bar U^n(\cdot) + o(n)$, where $\bar U^n \to \bar U^* = \bar A^{*^{-1}}$. Because $\bar A^{*^{-1}} \in \mathcal{C}^1$, we can *choose* a continuously differentiable function for $\bar U^n$ (the expansions are only unique in the limit), from which the bounded second order term follows directly by Taylor expansion.

where

$$\tilde{W}_+^n = \phi(\tilde{X}^n) \to \tilde{W}_+^* = \phi\left(\tilde{L}_+^* + \tilde{c}^*\right). \tag{74}$$

The latter implies that the (class) workload process W^n is also of order $n^{1/2}$ such that we have (25) and $\limsup_n \|\tilde{W}^n\|$ is bounded by $\|\tilde{W}_+^*\|$ which is finite (because $\tilde{W}_+^* \in \mathcal{C}$) but $\tilde{W}^n$ need not converge! From (8), (68) and (25), it follows that T^n has expansion (22) where

$$\bar{T}^n \to R^*, \tag{75}$$

$$\tilde{T}^n + \tilde{W}^n \to \tilde{L}^*. \tag{76}$$

Thus, like $\limsup \tilde{W}^n$, the $\limsup \tilde{T}^n$ is bounded, but $\tilde{T}^n$ need not necessarily converge. Moreover, convergence of $\tilde{T}^n$ is equivalent to convergence of $\tilde{W}^n$. Given the expansions of A^n, S^n, W^n and T^n, using (5) and (10), we have that both N^n and τ^n (under class FIFO) are of order $n^{1/2}$ as stated in (21) and (26). Using the convergence and differentiability assumption of $\bar{S}^*$ together with the boundedness of $\limsup \tilde{T}^n$,

$$\tilde{N}^n = \tilde{A}^n - \tilde{S}^n \circ R^n - (\mu \circ R^n)\tilde{T}^n. \tag{77}$$

Also, from (27), (26) and Assumption 1s, it follows that

$$\bar{T}_k^n(t + n^{-1}\tau_k^n(nt)) - \bar{T}_k^n(t) = \rho_k(t) n^{-1}\tau_k^n(nt) + o(n^{-1/2}), \tag{78}$$

so that (10) yields

$$\tilde{W}^n = \rho\,\tilde{\tau}^n + \tilde{T}_k^n(t + n^{-1}\tau_k^n(nt)) - \tilde{T}_k^n(t). \tag{79}$$

Again, $\limsup \tilde{N}^n$ and $\limsup \tilde{\tau}^n$ are bounded but $\tilde{N}^n$ and $\tilde{\tau}^n$ need not necessarily converge. However, their convergence is linked, yielding (35).

To show (33) and (34) we need the following lemma (which is a generalization of Lemma 3.3.c. in the seminal work of Iglehart and Whitt [16]).

Lemma 1 *If $\bar{U}_k^*, \tilde{U}_k^* \in \mathcal{C}$, then $n^{-1/2} \sup_{1 \le i \le A^n(n\cdot)} u_{k,i}^n \to 0$.*

Proof: From $\bar{U}_k^n \to \bar{U}_k^*$, it follows that $n^{-1/2}(U_k^n \circ n - n\bar{U}_k^*) \to \tilde{U}_k^*$. Define the maximum jump function $h : \mathcal{D} \to \mathbb{R}_+ : x \to h(x) = \sup_{t\in[0,1]} \mid x(t) - x(t-) \mid$. Because h is continuous at any $x \in \mathcal{C}$, $h(n^{-1/2}(U_k^n \circ n - n\bar{U}_k^*)) \to h(\tilde{U}_k^*)$. Because $\bar{U}_k^*$ and $\tilde{U}_k^*$ are continuous, this yields $n^{-1/2} h(U_k^n \circ n) \to 0$ which finishes the proof. ∎

The assumptions of this lemma are satisfied for both U^n and V^n, which concludes the proof of the proposition. ∎

A.3 Proposition 3

Proof: Denote by $v_k^{n(0)}$ the amount of service, if any, already given by time t to the oldest class k job. We have that

$$\begin{aligned} W_k^n(t) &= V_k^n(A_k^n(t)) - V_k^n(A_k^n(t) - N_k^n(t)) - v_k^{n(0)} && (80)\\ &= n\bar{V}_k^n(n^{-1}A_k^n(t)) - n\bar{V}_k^n(n^{-1}A_k^n(t) - n^{-1}N_k^n(t)) + o(N_k^n(t)) - v_k^{n(0)} && (81)\\ &= \bar{V}_k^{*\prime}(n^{-1}A_k^n(t))N_k^n(t) + o(N_k^n(t)) - v_k^{n(0)}, && (82)\end{aligned}$$

and thus

$$\tilde{W}_k^n(t) = \bar{V}_k^{*\prime}(\bar{A}_k^n(t))\tilde{N}_k^n(t) + o(\tilde{N}_k^n(t)) - n^{-1/2}v_k^{n(0)}. \qquad (83)$$

Thus, because $\lim \|\bar{V}_k^{*\prime} \circ \bar{A}_k^n\| < \infty$,

$$\|\tilde{W}_k^n - (\bar{V}_k^{*\prime} \circ \bar{A}_k^n)\,\tilde{N}_k^n\| \leq o(\|\tilde{N}_k^n\|) + n^{-1/2} \sup_{1\leq i\leq A_k^n(n)} v_{k,i}^n. \qquad (84)$$

Using Proposition 2 and noting that equation (17) implies that $\bar{V}_k^{*\prime} \circ \bar{A}^n - \mu_k^{-1} \to 0$ ends the proof. ∎

A.4 Proposition 4

Proof: Define C_a^n, C_c^n, C_d^n as follows (recall that $\tau_{k,i}^n$ represents the throughput time of the i-th class k job in the n-th system)

$$\begin{aligned} n^{3/2}(\bar{A}_k^n(b) - \bar{A}_k^n(a))C_a^n &= \sum_{i=A_k^n(na)}^{A_k^n(nb)} \tau_{k,i}^n \\ n^{3/2}(\bar{A}_k^n(b) - \bar{A}_k^n(a))C_c^n &= \int_{na}^{nb} N_k^n(t)\mathrm{d}t \\ n^{3/2}(\bar{A}_k^n(b) - \bar{A}_k^n(a))C_d^n &= \sum_{i=A_k^n(na)}^{A_k^n(nb)-N_k^n(nb)} \tau_{k,i}^n. \end{aligned}$$

The quantities on the right hand side may be thought of as three different charging schemes where jobs pay one dollar per unit time spent in the system. C_a^n charges the entire job cost at the job's arrival, C_d^n at the job's departure, and C_c^n charges continuously. It is clear that

$$C_d^n \leq C_c^n \leq C_a^n. \qquad (85)$$

Further,

$$
\begin{aligned}
C_a^n - C_d^n &= \frac{n^{-3/2}}{\bar{A}_k^n(b) - \bar{A}_k^n(a)} \sum_{i=A_k^n(nb)-N_k^n(nb)+1}^{A_k^n(nb)} \tau_{k,i}^n \\
&\leq \frac{n^{-3/2}}{\bar{A}_k^n(b) - \bar{A}_k^n(a)} N_k^n(b) \max_{A_k^n(nb)-N_k^n(nb)+1 \leq i \leq A_k^n(nb)} \tau_{k,i}^n \\
&\leq \frac{n^{-1/2}}{\bar{A}_k^n(b) - \bar{A}_k^n(a)} \tilde{N}_k^n(b) \|\tilde{\tau}_k^n\| + o(n^{-1/2}),
\end{aligned}
$$

and using Proposition 2,

$$\lim_{n\to\infty} C_a^n - C_d^n = 0. \tag{86}$$

Since

$$n^{-3/2} \int_{na}^{nb} N_k^n(t)\mathrm{d}t = \int_a^b \tilde{N}_k^n(t)\mathrm{d}t + o(1), \tag{87}$$

taking lim in (85) ends the proof. ∎

A.5 Proposition 6, the Lower Bound

Proof: Fix $\epsilon > 0$ and, for any $n \in \mathbb{N}$, consider the sequence of stopping times of $\tilde{W}_+^*$, $\{t_i : i \in \mathbb{N}\}$, defined as follows:

$$
\begin{aligned}
t_1 &= \min\left\{1, \inf\left\{0 < t \leq 1 : \left|\tilde{W}_+^*(t) - \lfloor \tilde{W}_+^*(0)/\epsilon \rfloor \epsilon\right| \geq \epsilon\right\}\right\} && (88) \\
t_{i+1} &= \min\left\{1, \inf\left\{t_i < t \leq 1 : \left|\tilde{W}_+^*(t) - \tilde{W}_+^*(t_i)\right| \geq \epsilon\right\}\right\}. && (89)
\end{aligned}
$$

Thus t_{i+1} is the first time $\tilde{W}_+^*$ changes by ϵ starting from $\tilde{W}_+^*(t_i)$ at time t_i. Because $\tilde{W}_+^*$ is continuous, $\sup_i(t_{i+1} - t_i) \to 0$ as $\epsilon \to 0$, so that $\sup_i(t_{i+1} - t_i) = O(\epsilon)$. Using Jensen's Inequality for convex functions, we have that

$$
\begin{aligned}
\tilde{J}^n &= \sum_k \sum_i \int_{nt_i}^{nt_{i+1}} C_k^n(\tau_k^n)\, n^{-1} \mathrm{d}A_k^n && (90) \\
&\geq \sum_k \sum_i n^{-1} \left[A_k^n(t_{i+1}) - A_k^n(t_i)\right] C_k^n \left(\left[A_k^n(t_{i+1}) - A_k^n(t_i)\right]^{-1} \int_{nt_i}^{nt_{i+1}} \tau_k^n \mathrm{d}A_k^n \right). && (91)
\end{aligned}
$$

Using Assumption 1 that $n^{-1}A^n(n\cdot) \to \bar{A}^*$, with continuous first derivative $\bar{A}^{*'} = \lambda$, we have that

$$
\begin{aligned}
n^{-1}\left[A^n(t_{i+1}) - A^n(t_i)\right] &= \bar{A}^*(t_{i+1}) - \bar{A}^*(t_i) + o_n(1) && (92) \\
&= \lambda(t_i)(t_{i+1} - t_i) + o(\epsilon) + o_n(1), && (93)
\end{aligned}
$$

where $o_n(1) \to 0$ as $n \to \infty$, and both bounds $o_n(1)$ and $o(\epsilon)$ are uniform over $[0,1]$. Thus,

$$\tilde{J}^n \geq \sum_k \sum_i [\lambda(t_i)(t_{i+1}-t_i) + o(\epsilon) + o_n(1)]\, C_k^n \left(n^{-1} [\lambda(t_i)(t_{i+1}-t_i) + o(\epsilon) + o_n(1)]^{-1} \int_{nt_i}^{nt_{i+1}} \tau_k^n \mathrm{d}A_k^n \right).$$

Evaluate the argument of C_k^n as follows,

$$\begin{aligned}
& n^{-1} [\lambda(t_i)(t_{i+1}-t_i) + o(\epsilon) + o_n(1)]^{-1} \textstyle\int_{nt_i}^{nt_{i+1}} \tau_k^n \mathrm{d}A_k^n \\
= \;& n^{-1} \left[(\lambda(t_i)(t_{i+1}-t_i))^{-1} + o(\epsilon) + o_n(1) \right] \textstyle\int_{nt_i}^{nt_{i+1}} \tau_k^n \mathrm{d}A_k^n && [(x+\Delta x)^{-1} = x^{-1} - \Delta x + o(\Delta x)] \\
= \;& n^{1/2} \left([\lambda(t_i)(t_{i+1}-t_i)]^{-1} \textstyle\int_{t_i}^{t_{i+1}} \tilde{N}_k^n \mathrm{d}t + o(\epsilon) + o_n(1) \right) && [\text{Proposition 4 + } \tilde{N}^n \text{ is bounded}] \\
= \;& n^{1/2} \left([\lambda(t_i)(t_{i+1}-t_i)]^{-1} \textstyle\int_{t_i}^{t_{i+1}} \mu_k \tilde{W}_k^n \mathrm{d}t + o(\epsilon) + o_n(1) \right) && [\text{Proposition 3}] \\
= \;& n^{1/2} \left([\mu_k(t_i) + O(\epsilon)]\, [\lambda(t_i)(t_{i+1}-t_i)]^{-1} \textstyle\int_{t_i}^{t_{i+1}} \tilde{W}_k^n \mathrm{d}t + o(\epsilon) + o_n(1) \right) && [\mu \text{ is continuous}] \\
= \;& n^{1/2} \left([\rho(t_i)(t_{i+1}-t_i)]^{-1} \textstyle\int_{t_i}^{t_{i+1}} \tilde{W}_k^n \mathrm{d}t + O(\epsilon) + o_n(1) \right) && [\tilde{W}_k^n \text{ is bounded}].
\end{aligned}$$

Assumption 2 and the continuity of C_k^* on $[0, \|\tilde{W}_+^*\|]$ together with the bound $\limsup \tilde{W}^n \leq \|\tilde{W}_+^*\|$ yield

$$\begin{aligned}
\tilde{J}^n \;&\geq\; \sum_k \sum_i \lambda_k(t_i)(t_{i+1}-t_i) C_k^* \left(\rho_k(t_i)^{-1} (t_{i+1}-t_i)^{-1} \int_{t_i}^{t_{i+1}} \tilde{W}_k^n(t) \mathrm{d}t \right) + o_n(1) + O(\epsilon) \\
&\geq\; \sum_k \sum_i \lambda_k(t_i)(t_{i+1}-t_i) C_k^* \left(\rho_k(t_i)^{-1} \left[g \circ (t_{i+1}-t)^{-1} \int_t^{t_{i+1}} \mathrm{d}t \circ \tilde{W}_+^n \right]_k (t_i) \right) + o_n(1) + O(\epsilon),
\end{aligned}$$

where we invoked the mapping g. Using the fact that $\tilde{W}_+^n \to \tilde{W}_+^*$ and the construction of the stopping times t_i, we have that

$$(t_{i+1}-t_i)^{-1} \int_{t_i}^{t_{i+1}} \tilde{W}_+^n(t) \mathrm{d}t = \tilde{W}_+^*(t_i) + O(\epsilon) + o_n(1). \tag{94}$$

The continuity of C_k^* and g on the bounded interval $[0, \|\tilde{W}_+^*\|]$, give a uniform bound

$$\tilde{J}^n \geq \sum_k \sum_i \lambda_k(t_i)(t_{i+1}-t_i) C_k^* \left(\rho_k(t_i)^{-1} \left[g \circ \tilde{W}_+^* \right]_k (t_i) \right) + O(\epsilon) + o_n(1), \tag{95}$$

and thus,

$$\liminf_{n\to\infty} \tilde{J}^n \geq \sum_k \sum_i \lambda_k(t_i)(t_{i+1}-t_i) C_k^* \left(\rho_k(t_i)^{-1} \left[g \circ \tilde{W}_+^* \right]_k (t_i) \right) + O(\epsilon). \tag{96}$$

The left hand side is independent of ϵ. Therefore, since ϵ is arbitrary, letting $\epsilon \to 0$ (which implies $\sup_i (t_{i+1}-t_i) \to 0$) and invoking the definition of the Riemann integral[4] completes the proof. ∎

[4]Since the function $C_k(g_k(\tilde{W}_+^*(\cdot)))/\rho_k(\cdot)$ is continuous on $[0,1]$, it is Riemann integrable.

A.6 Proposition 7

Proof: We will first show that $\tilde{T}^n$ converges. Fix a class, say j, and define the sequence of scalar functions $\{h^n : n \in \mathbb{N}\}$ where $h^n(t) = \mu_j(t) c_j^*(\rho_j(t)^{-1} \tilde{W}_j^n(t))$. The policy shows that for all $\epsilon > 0$ there exist an integer N such that for all $n > N$:

$$\left| c_k^*(\rho_k(t)^{-1} \tilde{W}_k^n(t)) - \mu_k^{-1}(t) h^n(t) \right| < \epsilon,$$

for all $t \in [0,1]$ (because $\mu_k > 0$ is bounded on $[0,1]$). According to Assumption 3, c_k^* is increasing and continuous. Therefore, its inverse function $c_k^{*^{-1}}$ is also continuous on $[0,1]$. Thus, for all $\epsilon' > 0$ there exist a (uniform) $\delta > 0$ such that if $\epsilon < \delta$ then

$$\left| \rho_k(t)^{-1} \tilde{W}_k^n(t) - c_k^{*^{-1}} \left(\mu_k^{-1}(t) h^n(t) \right) \right| < \epsilon'.$$

Summing over the (maximally d) classes k:

$$\left\| \tilde{W}_+^n(\cdot) - \sum_k \rho_k(\cdot) c_k^{*^{-1}} \left(\frac{h^n(\cdot)}{\mu_k(\cdot)} \right) \right\| < \|\rho\| \epsilon' d.$$

Because the marginal costs are increasing and $\tilde{W}_+^n$ converges, h^n and thus also $\tilde{W}^n$ converge.
The policy controls the workloads such that $\tilde{W}^*$ is the solution to the sufficient first order conditions of the minimization problem (43). Thus, $\tilde{W}^* = g \circ \tilde{W}_+^*$ and proposition 5 shows that $\tilde{J}^n \to \tilde{J}^*$. ∎

References

[1] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\,\mu$ rule revisited. *Adv. Appl. Prob.*, 17:237–238, 1985.

[2] Pierre Chardaire and Marc Lesk. Grade of service and optimization of distributed packet-switched networks. *Computer Networks and ISDN Systems*, 11:139–146, 1986.

[3] Hong Chen and David D. Yao. Dynamic scheduling of a multiclass fluid network. *Operations Research*, 41(6):1104–1115, Nov-Dec 1993.

[4] E. G. Coffman, Jr., M. I. Reiman, and A. A. Puhalskii. Polling systems with zero switchover times: a heavy-traffic averaging principle. *submitted*, 1994.

[5] D. R. Cox and Walter L. Smith. *Queues*. Methuen (London) and Wiley (New York), 1961.

[6] Yves De Serres. Simultaneous optimization of flow control and scheduling in a single server queue with two job classes. *Operations Research Letters*, 10:103–112, 1991.

[7] Sanjeev Dewan and Haim Mendelson. User delay costs and internal pricing for a service facility. *Management Science*, 36(12):1502–1517, December 1990.

[8] Peter W. Glynn. Diffusion approximations. In D. P. Heyman and M. J. Sobel, editors, *Handbooks on OR & MS, Vol. 2*, pages 145–198, Elsevier Science Publishers, B.V., North-Holland, 1990.

[9] Rasoul Haji and G. F. Newell. Optimal strategies for priority queues with nonlinear costs of delay. *SIAM J. Appl. Math.*, 20(2):224–240, March 1971.

[10] J. Michael Harrison. Brownian models of queueing networks with heterogeneous customer populations. In W. Fleming and P. L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 147–186, Springer-Verlag, New York, 1988.

[11] J. Michael Harrison. *Brownian Motion and Stochastic Flow Systems.* John Wiley and Sons, New York, 1985.

[12] J. Michael Harrison. Dynamic scheduling of a multiclass queue: discount optimality. *Operations Research*, 23(2):270–282, March-April 1975.

[13] J. Michael Harrison and Lawrence M. Wein. Scheduling networks of queues: heavy traffic analysis of a two-station closed network. *Operations Research*, 38:1052–1064, 1990.

[14] J. Michael Harrison and Lawrence M. Wein. Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Systems*, 5:265–280, 1989.

[15] Tetsuji Hirayama, Masaaki Kijima, and Shoichi Nishimura. Further results for dynamic scheduling of multiclass $G/G/1$ queues. *J. of App. Prob.*, 26:595–603, 1989.

[16] Donald L. Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. *I* and *II*. *Adv. Appl. Prob.*, 2:150–177, 1970.

[17] F. P. Kelly and C. N. Laws. Dynamic routing in open queueing networks. *J. of App. Prob.*, 12:542–554, 1992.

[18] G. P. Klimov. Time-sharing service systems *I*. *Theory of Probability and its Applications*, 19(3):558–576, 1974.

[19] Elena V. Krichagina, Sheldon X. C. Lou, Suresh P. Sethi, and Michael I. Taksar. Production control in a failure-prone manufacturing system: diffusion approximation and asymptotic optimality. *Ann. Appl. Prob.*, 3:421–453, 1993.

[20] H. J. Kushner and L. F. Martins. Routing and singular control for queueing networks in heavy traffic. *SIAM J. Control Optimization*, 28:1209–1233, 1990.

[21] H. J. Kushner and K. M. Ramachandran. Optimal and approximately optimal control policies for queues in heavy traffic. *SIAM J. Control Optimization*, 27:1293–1318, 1989.

[22] L. F. Martins, S. E. Shreve, and H. M. Soner. Heavy traffic convergence of a controlled, multi-class queueing system. *submitted*, 1994.

[23] Philippe Nain. Interchange arguments for classical scheduling problems in queues. *Systems & Control Letters*, 12:177–184, 1989.

[24] Michael Pinedo. Stochastic scheduling with release dates and due dates. *Operations Research*, 31(3):559–572, May-Jun 1983.

[25] Rhonda Righter. Scheduling. In M. Shaked and J. G. Shanthikumar, editors, *Stochastic Orders*, Academic Press, 1994.

[26] Rhonda Righter and Susan H. Xu. Scheduling jobs on non-identical IFR processors to minimize general cost functions. *Adv. Appl. Prob.*, 23:909–924, 1991.

[27] Michael H. Rothkopf and Stephen A. Smith. There are no undiscovered priority index sequencing rules for minimizing total delay costs. *Operations Research*, 32(2):451–456, March-April 1984.

[28] H. L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[29] J. George Shanthikumar and David D. Yao. Multiclass queueing systems: polymatroidal structure and optimal scheduling control. *Operations Research*, 40(Supp. No. 2):S293–S299, May-Jun 1992.

[30] Harvey N. Shycon and Christopher R. Sprague. Put a price tag on your customer servicing levels. *Harvard Business Review*, 71–78, July-August 1975.

[31] Wayne E. Smith. Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3:59–66, 1956.

[32] Dong-Wan Tcha and Stanley R. Pliska. Optimal control of single-server queueing networks and multi-class $M/G/1$ queues with feedback. *Operations Research*, 25(2):248–258, March-April 1977.

[33] A. J. Thadhani. Interactive user productivity. *IBM Systems Journal*, 20(4):407–423, 1981.

[34] Eugene Veklerov. On Rothkopf and Smith's statement regarding optimal priority assignment. *Operations Research*, 37(3):498–500, May-June 1989.

[35] Richard R. Weber. Stochastic scheduling on parallel processors and minimization of concave functions of completion times. In W. Fleming and P. L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 601–609, Springer-Verlag, New York, 1988.

[36] Lawrence M. Wein. Due-date setting and priority sequencing in a multiclass $M/G/1$ queue. *Management Science*, 37(7):834–850, July 1991.

[37] Lawrence M. Wein. The impact of processing time knowledge on dynamic job-shop scheduling. *Management Science*, 37(8):1002–1014, August 1991.

[38] Lawrence M. Wein. Optimal control of a two-station brownian network. *Mathematics of Operations Research*, 15(2):215–242, May 1990.

[39] Lawrence M. Wein. Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs. *Operations Research*, 38(6):1065–1078, November-December 1990.

[40] Lawrence M. Wein and Philippe B. Chevalier. A broader view of the job-shop scheduling problem. *Management Science*, 38(7):1018–1033, July 1992.

Graduate School of Business, Stanford University, Stanford, CA 94305-5015. (janvm@leland.stanford.edu)