On the Conservation Law and the Performance Space of Single Server Systems

Leonidas Georgiadis
High Performance Computing and Communications
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Ioannis Viniotis [1]
Department of Electrical and Computer Engineering
Center for Communications and Signal Processing
North Carolina State University, Raleigh, NC 27695

[1]Work supported by IBM, while the author was a visiting scientist at the IBM T. J. Watson Research Center.

Abstract

We consider a multiclass $GI|G|1$ queueing system, operating under an arbitrary work-conserving scheduling policy π. We derive an invariance relation for the Cesaro sums of waiting times under π, which does not require the existence of limits of the Cesaro sums. This allows us to include in the set of admissible policies important classes, such as time-dependent and adaptive policies. For these classes of policies, ergodicity is not known a priori and may not even exist. Therefore, the classical invariance relations, involving statistical averages do not hold. For an $M|G|1$ system, we derive inequalities involving the Cesaro sums of waiting times, that further characterize the achievable performance region of the system.

1 Introduction.

Conservation laws (i.e., invariance relations,) regarding average waiting times and average number of customers in the system have been known for a long time. Most of the conservation laws presented so far [13, 12], deal with single server, *work conserving* systems, or multiple server, *strongly work conserving* systems [12, 6]. Roughly speaking, in such systems, no work (service time) is created or destroyed within the system (e.g., customers do not balk, the server maintains always the same speed, etc.) For such systems, *under ergodicity assumptions* (that is, under the assumption that limits of sample averages of waiting times exist under a scheduling policy π,) one can prove invariance relations of the form (see e.g., [12],)

$$\sum_{i=1}^{N} \rho_i W_i(\pi) = F, \tag{1}$$

where N is the number of classes, $W_i(\pi)$ is the limit of the sample average waiting time for customers of class i, ρ_i is the utilization of class i and F is a constant, independent of the scheduling policy. Relatively little can be said about multiple server systems [6].

The equation in (1) constrains the average waiting times under an ergodic scheduling policy to lie on an N-dimensional plane. A number of inequalities [8] further constrain the set of achievable average waiting times (i.e., the performance region) to be a convex polygon. These inequalities take typically the form

$$\sum_{i \in R} \rho_i W_i(\pi) \geq F(R), \tag{2}$$

where $R \subset \{1, 2, \cdots, N\}$ and $F(R)$ is a function independent of the policy π. For further details, the reader may consult [8].

In practice, work can be created and/or destroyed, within the system, in a scheduling policy-dependent fashion. For example, when preemptive priority rules are considered, interruption of a low priority customer introduces an overhead. Since this overhead depends on the priority rule in hand, an invariance relation is no longer possible. Policy-dependent overhead is also generated in polling systems, whenever the server spends a nonzero time to switch from a station

to another (nonzero server walking time.) For such $non\ work\text{-}conserving$ systems, a number of $pseudoconservation$ laws have appeared (see for example [2]). Such laws express a relation between waiting times, akin to that of eq (1), but with policy-dependent right hand sides.

Ergodicity of the scheduling policy used is a key assumption for all the above studies. In fact, many authors (an exception being [12, Theorem 11-13]) constrain further the class of work-conserving policies to policies that have regenerative structure [4],[8],[6]. However, for a large (and rather important for applications,) class of policies, namely adaptive [1], and in general, policies that base their decisions on the entire history of the system, convenient regeneration points may not exist. Moreover it may not be a priori clear that limits of sample averages exist under such policies.

The motivation to study existence of invariance relations without ergodicity assumptions came from the application studied in [1]. Briefly, the problem considered in [1] is the following: requests belonging to N classes arrive for service at a single server queue. With each class, there is an associated response time objective g_i. The goal is to determine a scheduling policy π such that $\bar{W}_k(\pi) \stackrel{\Delta}{=} \limsup_{k\to\infty} \sum_{m=1}^{k} W_{im}(\pi)/k$, where $W_{im}(\pi)$ is the response time of the mth served customer from class i, is kept below the objective g_i. Note that for an ergodic policy, $\bar{W}_k(\pi)$ is the limit of the sample average of the waiting times of customers from class i. No a priori knowledge of system statistics was assumed. This restriction excluded randomized policies [8] and lead to the design of an adaptive policy. In addition, in this framework, the question arose whether by using nonergodic policies one can improve the performance of the system. This question raises the issue of what equalities/inequalities the sample averages of waiting times under general nonpreemptive work-conserving policies may satisfy. Using the results of the current paper, it was shown that, within the whole class of (both ergodic and nonergodic) nonpreemptive, work-conserving policies, the policy proposed in [1] achieved the above mentioned goal.

The paper is organized as follows: in Section 2 we define the $GI|G|1$ model for which the conservation law holds and specify the class of admissible policies. In Section 3 we prove a conservation law for the $GI|G|1$ system and provide a counterexample to the claim that limits

of sample averages of waiting times always exist when the interarrival times are exponentially distributed. In Section 4, we derive inequalities that impose further constraints on the sample averages of waiting times for the $M|G|1$ system.

2 System model and admissible policies

In the sequel, we use the following model. (This model has been used in [15] as well.)

System Model: We consider a single server queue with N classes of customers. The time interval between the kth and $k+1$st arriving customer ($k \geq 1$) is denoted by T_k. We assume that the first arrival occurs at time $t = 0$. With each arrival there is an associated random variable, C_k, which denotes the class to which the arrival belongs; that is, if $C_k = i, \quad i = 1, \cdots, N$, then the kth customer belongs to class i. If the kth arriving customer belongs to class i, its service time is a random variable S_{ik}. Therefore, S_k, the service time of the kth arriving customer is given by

$$S_k = \sum_{i=1}^{N} I_{\{C_k=i\}} S_{ik}, \tag{3}$$

where $I_{\{A\}}$ denotes the indicator function of the event A. We assume that $\{T_k,\ k \geq 1\}$, $\{C_k,\ k \geq 1\}$ and $\{S_{ik},\ k \geq 1\}$, for each $i = 1, \cdots, N$, are independent sequences. Each sequence contains i.i.d. random variables. It follows from (3) that $\{S_k,\ k \geq 1\}$ is a sequence of i.i.d. random variables independent of $\{T_k,\ k \geq 1\}$. We define the total arrival rate as $\lambda \stackrel{\Delta}{=} 1/ET_1$ and the arrival rate of class i as $\lambda_i \stackrel{\Delta}{=} P(C_1 = i)/ET_1$. To avoid unnecessary complications, we assume that $\lambda > 0$, $P(T_1 = 0) = 0$, $P(S_1 = 0) = 0$ and $P(C_1 = i) > 0, \quad i = 1, \cdots, N$.

For the conservation law of Theorem 1 in Section 3, the distribution of $\{T_k,\ k \geq 1\}$ is arbitrary. However, for the inequality relations of Theorem 2 in Section 4 we need to assume that the random variables $\{T_k,\ k \geq 1\}$ are exponentially distributed. Then this is equivalent to the assumption that the arrival instants of each class constitute a Poisson process with rate $\lambda_i \stackrel{\Delta}{=} P(C_1 = i)/ET_1$, independent of the arrival processes of the other classes.

We specify next the set of $admissible$ scheduling policies (i.e., all policies under which the

system may operate.) This is the class of $nonpreemptive,\ work\text{-}conserving$ policies. We place no ergodicity restriction on an admissible policy. Thus, any adaptive or time-dependent policy is a member of this class. A rigorous definition of this class is given in [9]. Roughly speaking, a work-conserving policy does not idle the server while customers are waiting in the queue, does not affect the amount of service time given to a customer or the arrival time of a customer and is $nonanticipative$ (that is, the scheduling decisions do not depend on future interarrival times, future service times or the service times of customers that are present in the system at the time when a scheduling decision is made.) Under a nonpreemptive policy, a customer in the queue may not replace a customer who is being served before its service requirements are completed.

Under a nonpreemptive, work-conserving policy, the service times of the $departing$ customers from class i are i.i.d. r.v., identically distributed to $\{S_{ik},\ k \geq 1\}$. Moreover, the service time of the kth departing customer from class i is independent of its waiting time (time the customer was waiting in queue before its service started) and of the waiting times of the customers from class i that were served before the kth customer. These properties are crucial for the results in the next sections and are stated formally in Lemma 1 below. Although the assertions of Lemma 1 are fairly intuitive, the proof requires a more precise definition of the class of admissible policies as well as some technical arguments which are presented in [9].

For the kth departing customer from class i, let W_{ik} denote its waiting time and $\bar{S}_{ik}$ its service time. For $i = 1, 2, \cdots, N$, and $k \geq 2$, let

$$\begin{aligned} \mathcal{F}_{i1} &\stackrel{\Delta}{=} \sigma(W_{i1}) \\ \mathcal{F}_{ik} &\stackrel{\Delta}{=} \sigma(W_{in}, 1 \leq n \leq k;\ \ \bar{S}_{in}, 1 \leq n \leq k-1), \end{aligned}$$

where $\sigma(\mathcal{X})$ denotes the σ-field generated by the set of random variables $\mathcal{X}$.

Lemma 1 *Under any nonpreemptive, work-conserving policy the following statements are true:*

1. $\bar{S}_{ik}$ *is independent of* $\mathcal{F}_{ik}$, *for every* $k \geq 1$.

2. *For a fixed* $i,\ 1 \leq i \leq N$, *the sequences* $\{\bar{S}_{ik},\ k \geq 1\}$ *and* $\{S_{ik},\ k \geq 1\}$, *are identically distributed.*

Remark: Note that the second statement of Lemma 1 is not true for the service times of the successive departing customers. In other words, if $\bar{S}_k$ is the service time of the kth departing customer (irrespective of its class,) it is not true in general that under a nonpreemptive work-conserving policy, $\{\bar{S}_k,\ k \geq 1\}$ is identically distributed to $\{S_k,\ k \geq 1\}$.

3 The conservation law

The key to the main result of this section is the Strong Law of Large Numbers for Martingales which is presented in Lemma 2. Its proof can be found in [11]. As will be seen in the proof of Theorem 1, the corollary to this lemma, Corollary 1, will allow us to state the conservation law *without the need* to assume the existence of limits of sample averages of waiting times.

Lemma 2 *Let* $\{X_k,\ k \geq 1\}$ *be a sequence of random variables and* $\{\mathcal{F}_k,\ k \geq 1\}$ *an increasing sequence of* σ*-fields, with* X_k *measurable with respect to* $\mathcal{F}_k$, *for each* k. *Let* Z *be a random variable and let c be a constant such that* $E(|Z| \cdot \max\{0, \log|Z|\}) < \infty$ *and* $P(|X_k| > x) \leq c \cdot P(|Z| > x)$ *for each* $x \geq 0$ *and* $k \geq 1$. *Then*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{m=1}^{k} [X_{m+1} - E(X_{m+1}|\mathcal{F}_m)] = 0 \ a.e.$$

The next corollary follows easily from Lemma 2.

Corollary 1 *Let* $\{\mathcal{W}_k,\ k \geq 1\}$, $\{\mathcal{S}_k,\ k \geq 1\}$ *be sequences of random variables and* $\{\mathcal{F}_k, k \geq 1\}$ *be an increasing family of* σ*-fields. Suppose that* $\mathcal{W}_k$ *is* $\mathcal{F}_k$*-measurable for all* $k \geq 1$ *and that* $\mathcal{S}_{k-1}$*is* $\mathcal{F}_k$*-measurable for all* $k \geq 2$. *Suppose also that* $\mathcal{S}_k$ *is independent of* $\mathcal{F}_k$, *for all* $k \geq 1$. *Let there be a random variable* Z, *a constant* $\delta > 1$, *and a constant* c *such that* $E|Z|^\delta < \infty$ *and*

$P(|\mathcal{W}_k\mathcal{S}_k| > x) \leq cP(|Z| > x)$ *for each* $x \geq 0$ *and* $k \geq 1$. *Then*

$$\lim_{k\to\infty}\left(\frac{\sum_{m=1}^{k}\mathcal{W}_m\mathcal{S}_m}{k} - \frac{\sum_{m=1}^{k}\mathcal{W}_mE\mathcal{S}_m}{k}\right) = 0 \quad a.e. \tag{4}$$

Proof: Let $X_1 \stackrel{\Delta}{=} 0$, $X_k \stackrel{\Delta}{=} \mathcal{W}_{k-1}\mathcal{S}_{k-1}, \quad k \geq 2$. Clearly, X_k is $\mathcal{F}_k$-measurable for all $k \geq 1$. Also, since $\mathcal{S}_k$ is independent of $\mathcal{F}_k$ and $\mathcal{W}_k$ is $\mathcal{F}_k$-measurable, we have that for $k \geq 1$, $E(X_{k+1}|\mathcal{F}_k) = E(\mathcal{W}_k\mathcal{S}_k|\mathcal{F}_k) = \mathcal{W}_kE\mathcal{S}_k$. Since $\delta > 1$, the condition $E|Z|^\delta < \infty$ implies that $E(|Z|\max\{0, \log|Z|\}) < \infty$. Equation (4) follows now by applying Lemma 2 to the sequence $\{X_k,\ k \geq 1\}$. □

Recall that W_{im} (or $W_{im}(\pi)$, when dependence on the policy needs to be emphasized,) denotes the waiting time of the mth departing customer from class i when policy π is used and $\bar{S}_{ik}$ denotes the service time of the kth departing customer from class i. We shall use Corollary 1 in Theorem 1 below, by replacing $\mathcal{W}_k$ with W_{ik} and $\mathcal{S}_k$ with $\bar{S}_{ik}$. To verify the conditions of Corollary 1 in this case, we need the lemma that follows and some conditions on the moments of the service times. Let $s_i^{(m)} \stackrel{\Delta}{=} ES_{i1}^m$ denote the mth moment of the random variable S_{i1}. Let $s_i \stackrel{\Delta}{=} ES_{i1}^1$ and define ρ_i, the utilization of class i, as $\rho_i = \lambda_i s_i$.

Lemma 3 *Suppose that* $\sum_{i=1}^{N}\rho_i < 1$ *and that for some constant* $\gamma > 2$, $s_i^{(\gamma)} < \infty$, $i = 1, \cdots, N$. *Then under any nonpreemptive, work-conserving policy, there exists a constant* c_i *and a nonnegative random variable* Z_i *such that* $EZ_i^{\gamma/2} < \infty$, *and for each* $x \geq 0$ *and* $k \geq 1$, *we have that*

$$P(W_{ik}(\pi)\bar{S}_{ik} > x) \leq c_iP(Z_i > x).$$

The proof of Lemma 3 is given in [9]. The main idea is to use the fact that under any work-conserving policy, the waiting time of a customer is bounded by the length of the busy period in which the customer arrived, a quantity which is independent of the policy. The details of the proof are based on renewal arguments and are omitted for the sake of brevity.

We are now ready to state and prove the conservation law for the $GI|G|1$ queueing system. For an admissible scheduling policy π, let $V^{(\pi)}(t)$ denote the work in system (i.e., the sum of

remaining service times of all customers in the system at time t,) when policy π is used. Let us also denote by $FIFO$ the policy that serves customers in First In First Out order. It is well known that for this policy

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T V^{(FIFO)}(t)dt = V^* = constant \quad a.e.$$

Theorem 1 *Suppose that* $\sum_{i=1}^N \rho_i < 1$ *and that for some constant* $\gamma > 2$, $s_i^{(\gamma)} < \infty, \quad i = 1,\ldots,N$. *Then, for all work-conserving, nonpreemptive policies, we have that*

$$\lim_{k\to\infty}\sum_{i=1}^N \rho_i\left(\frac{\sum_{m=1}^k W_{im}(\pi)}{k}\right) = V^* - \frac{1}{2}\sum_{i=1}^N \lambda_i s_i^{(2)} \quad a.e.$$

Proof: Let π be an admissible policy. Since the admissible policies are nonidling, the work in system at time t is independent of the policy, [13], and therefore, for all $t \geq 0$,

$$V^{(FIFO)}(t) = \sum_{i=1}^N V_i^{(\pi)}(t), \tag{5}$$

where $V_i^{(\pi)}(t)$ is the work in system at time t due to customers from class i only.

Let K_{im} denote the number of class i customers served during the mth busy period and $K_m = \sum_{i=1}^N K_{im}$. Let also $M_{ik} = \sum_{m=1}^k K_{im}$ and $M_k = \sum_{m=1}^k K_m$. Since $\sum_{i=1}^N \rho_i < 1$, we have that $EK_1 < \infty$, [5]. Using the Strong Law of Large Numbers, we have that $\lim_{k\to\infty} M_{ik}/k = P(C=i)EK_1$ and $\lim_{k\to\infty} M_k/k = EK_1$. Therefore,

$$\lim_{k\to\infty}\frac{M_{ik}}{M_k} = \lim_{k\to\infty}\frac{M_{ik}/k}{M_k/k} = P(C_1 = i) = \frac{\lambda_i}{\lambda} \quad a.e. \tag{6}$$

From [3, Theorem 6] we have that

$$\lim_{k\to\infty}\frac{1}{k}\sum_{m=1}^k W_k^a S_k = \frac{V^*}{\lambda} - \frac{1}{2}\sum_{i=1}^N \frac{\lambda_i}{\lambda} s_i^{(2)} \quad a.e., \tag{7}$$

where W_k^a denotes the waiting time of the kth arriving customer. Taking the limit in (7) over the subsequence M_k and rearranging terms, we have that

$$\begin{aligned}\lim_{k\to\infty}\frac{1}{M_k}\sum_{i=1}^N\sum_{m=1}^{M_{ik}} W_{im}(\pi)\bar{S}_{im} &= \lim_{k\to\infty}\sum_{i=1}^N \frac{M_{ik}}{M_k}\frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)\bar{S}_{im}}{M_{ik}} \\ &= \frac{V^*}{\lambda} - \frac{1}{2}\sum_{i=1}^N \frac{\lambda_i}{\lambda} s_i^{(2)} \quad a.e. \end{aligned} \tag{8}$$

In (8), the sum $\sum_{m=1}^{M_{ik}} W_{im}(\pi)\bar{S}_{im}$ appears, while for the conservation law we need only the sum $\sum_{m=1}^{M_{ik}} W_{im}(\pi)$. It is here where Corollary 1 is used. Towards this end, let us define,

$$r_{ik} \stackrel{\Delta}{=} \frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)\bar{S}_{im}}{M_{ik}} - s_i \frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)}{M_{ik}}. \tag{9}$$

Using Lemma 1 and Lemma 3 we see that (for a fixed i,) the sequences $\mathcal{W}_k = W_{ik}(\pi)$, $\mathcal{S}_k = \bar{S}_{ik}$ and $\mathcal{F}_k = \mathcal{F}_{ik}$ satisfy all the conditions of Corollary 1. Since $\lim_{k\to\infty} M_{ik} = \infty \ \ a.e.$, we conclude that

$$\lim_{k\to\infty} r_{ik} = 0 \quad a.e. \tag{10}$$

From (8), (9) and (10) we have that

$$\lim_{k\to\infty} \sum_{i=1}^{N} \left(\frac{M_{ik}}{M_k} \ s_i \ \frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)}{M_{ik}} + \frac{M_{ik}}{M_k} r_{ik} \right) = \frac{V^*}{\lambda} - \frac{1}{2} \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} s_i^{(2)} \quad a.e. \tag{11}$$

To proceed, we need to replace in (11) the sequence M_{ik}/M_k with its limit, λ_i/λ, and this can be done if we show that for a fixed i, the sequence $\{\sum_{m=1}^{M_{ik}} W_{im}(\pi)/M_{ik}, \quad k \geq 1\}$ is bounded almost everywhere. Let B_k denote the length of the kth busy period; $\{B_k, \ k \geq 1\}$ is a sequence of i.i.d. random variables, independent of the scheduling policy. Observe that since $ES_1^2 < \infty$, we have that $EB_1^2 < \infty$, $EK_1^2 < \infty$, [10], and therefore, from Schwartz's inequality we have that

$$E[K_1 B_1] \leq (EK_1^2 \ \cdot \ EB_1^2)^{1/2} < \infty.$$

Let $\bar{B}_{im}$ denote the busy period during which the mth customer from class i departs. Since $W_{im}(\pi) \leq \bar{B}_{im}$, we have that

$$\begin{aligned} \limsup_{k\to\infty} \{ \frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)}{M_{ik}} \} &\leq \limsup_{k\to\infty} \{ \frac{\sum_{m=1}^{M_{ik}} \bar{B}_{im}}{M_{ik}} \} = \limsup_{k\to\infty} \frac{\sum_{m=1}^{k} B_m K_{im}}{\sum_{m=1}^{k} K_{im}} \\ &\leq \lim_{k\to\infty} \frac{\sum_{m=1}^{k} B_m K_m}{\sum_{m=1}^{k} K_{im}} = \frac{E[B_1 K_1]}{P(C=i) E K_1} < \infty \quad a.e. \end{aligned} \tag{12}$$

Using now (6),(11),(10) and (12) we have that

$$\lim_{k\to\infty} \sum_{i=1}^{N} \lambda_i s_i \frac{\sum_{m=1}^{M_{ik}} W_{im}(\pi)}{M_{ik}} = V^* - \frac{1}{2} \sum_{i=1}^{N} \lambda_i s_i^{(2)} \quad a.e. \tag{13}$$

It remains to show that (13) holds for all k (and not just for the subsequences M_{ik}). This is done by using the standard arguments by which partial reward limits are established from the known limits of renewal reward processes, see e.g. [14, Section 2.3]. □

When the arrival process of each class is Poisson, it is stated in [12, p. 432] that under any nonpreemptive, work-conserving policy the limits of the sample averages of the waiting times exist. If this statement were correct, Theorem 1 would be useful only when interarrival times are not Poisson. It was hinted by Federgruen and Groenvelt [7], however, that the statement may not be true. We present next a counterexample that shows that the assumption of Poisson arrival processes does not guarantee ergodicity of the scheduling policies. Therefore, stronger conditions on these policies must be imposed to establish ergodicity.

Counterexample. Consider an M/G/1 queue with two classes of customers. The two classes are characterized by their arrival rates (denoted as λ_1 and λ_2) and service rates (denoted as μ_1 and μ_2.) Assume that the system is stable (i.e., $\lambda_1/\mu_1 + \lambda_2/\mu_2 < 1$), the second moments of the service requirements are finite, and that $\lambda_1/\mu_1 > 0$, $\lambda_2/\mu_2 > 0$. Define a time-dependent policy π, as follows:

> At the beginning of the nth busy period ($n \geq 1$,) policy π gives highest priority to class 1 customers, if $2^{2k} \leq n \leq 2^{2k+1} - 1$, for some $k \geq 0$. Otherwise, (i.e., if $2^{2k+1} \leq n \leq 2^{2(k+1)} - 1$,) π gives highest priority to class 2 customers.

We shall show that under policy π, limits of sample averages of waiting times do not exist. Let W_m denote the waiting time of the mth served customer from class 1, $m \geq 1$. Define the following sets:

$H(n) \stackrel{\Delta}{=} \{m$: customer m belongs to class 1 and is served during busy period l, $1 \leq l \leq n$, in which class 1 has highest priority $\}$

$L(n) \stackrel{\Delta}{=} \{m$: customer m belongs to class 1 and is served during busy period l, $1 \leq l \leq n$, in which class 1 has lowest priority$\}$

Of course, $H(n) \cup L(n) = \{1, 2, \cdots, M_n\}$, where M_n denotes the number of class 1 customers served in the first n busy periods. Let M_n^h (resp. M_n^l) denote the cardinality of the set $H(n)$ (resp. $L(n)$.) Finally, let $h(n),\ l(n)$ denote the number of busy periods among the first n busy periods during which class 1 has highest and lowest priority respectively. Clearly, $h(n)+l(n) = n$. Then, $\bar{W}_n$, the sample average waiting time of class 1 customers during the first n busy periods, is given by

$$\bar{W}_n \stackrel{\Delta}{=} \frac{1}{M_n} \sum_{m=1}^{M_n} W_m = \frac{n}{M_n} \frac{h(n)}{n} \frac{M_n^h}{h(n)} \frac{\sum_{m \in H(n)} W_m}{M_n^h} + \frac{n}{M_n} \frac{l(n)}{n} \frac{M_n^l}{l(n)} \frac{\sum_{m \in L(n)} W_m}{M_n^l}. \tag{14}$$

We shall show now that $\bar{W}_n$ does not converge. Consider first the subsequence $\{n_k = (2^{2k} - 1),\ k \geq 1\}$. Observe that $h(n_k) = \frac{1}{3}(2^{2k} - 1)$ and $l(n_k) = \frac{2}{3}(2^{2k} - 1)$. The random variables $W_m,\ m \in H(n),$ are identically distributed to the corresponding random variables in a queueing system that gives always nonpreemptive priority to customers in class 1. Similarly, the random variables $W_m,\ m \in L(n)$, are identically distributed to the corresponding random variables in a queueing system that gives always nonpreemptive priority to customers in class 2.

Let $B_k,\ I_k,\ k = 1, 2, \cdots$ denote the length of the kth busy period and kth idle period respectively. The random variables $B_k + I_k,\ k = 1, 2, \cdots$ are independent, identically distributed. Moreover, their distribution does not depend on the scheduling policy, since the policies considered are nonidling. Let EB (EI) denote the mean of B_k (I_k.) Then, using the Strong Law of Large Numbers we obtain as in the proof of Theorem 1, that $M_n/n \to \lambda_1(EB + EI)$. Similarly, since $\lim_{n\to\infty} h(n) = \infty$ and $\lim_{n\to\infty} l(n) = \infty$, we have that $M_n^h/h(n) \to \lambda_1(EB + EI)$ and $M_n^l/l(n) \to \lambda_1(EB + EI)$. Therefore, taking limits along the subsequence n_k in (14) we conclude that

$$\lim_{k\to\infty} \bar{W}_{n_k} = \frac{1}{3} W^h + \frac{2}{3} W^l \quad a.e.,$$

where W^h (resp. W^l) denote the average waiting time of class 1 customers, under the policy which always gives strict nonpreemptive priority to class 1 (resp. class 2). Similarly, if we consider the subsequence $\{\bar{n}_k = 2^{2k+1} - 1,\ k \geq 0\}$, we find that $h(\bar{n}_k) = (2 \cdot 2^{2k+1} - 1)/3$ and $l(\bar{n}_k) = (2^{2k+1} - 2)/3$. Therefore, from (14),

$$\lim_{k\to\infty} \bar{W}_{\bar{n}_k} = \frac{2}{3} W^h + \frac{1}{3} W^l \quad a.e.$$

Since $W^h < W^l$, we see that the sample averages converge to two different points along the two subsequences $\{n_k\}, \{\bar{n}_k\}$. The system does not reach steady-state under this particular time-dependent policy. □

4 Inequality constraints for M/G/1 Systems

Theorem 1 describes a hyperplane in the N-dimensional space, on which the limit of a linear combination of the sample averages of waiting times must lie. Clearly, not all points in this hyperplane can be obtained by employing some work conserving polilcy. The performance space (i.e., the subset of points that are achievable by some admissible scheduling policy,) is well known for ergodic systems. For an $M|G|1$ system, we are able to further characterize the performance space. As we shall see, the constants $F(R)$ that appear in the right hand side of the inequalities in Theorem 2 are the same as the ones that appear when only ergodic policies are allowed.

For the rest of this section, we assume that the interarrival times are exponential. As in [8, 7], the basic idea is to determine a lower bound on the time average of the work in system due to customers that belong to a subset R of the classes.

Suppose that the system operates under a nonpreemptive, work-conserving policy π. For any set $R,\ R \subset \{1, 2, \cdots, N\}$, let $V_R^{(\pi)}(t)$ be the work in system at time t, due to customers that belong to the classes in R, that is,

$$V_R^{(\pi)}(t) \stackrel{\Delta}{=} \sum_{i \in R} V_i^{(\pi)}(t). \tag{15}$$

We are interested in developing a lower bound for the quantity

$$\liminf_{T \to \infty} \frac{\int_0^T V_R^{(\pi)}(t) dt}{T}. \tag{16}$$

In [7], it was shown that for all $t \geq 0$ and for each sample path, $V_R^{(\pi)}(t)$ is minimized by a policy which gives nonpreemptive priority to classes in R. Moreover, since the order of service of customers from classes in R cannot affect $V_R^{(\pi)}(t)$, it is sufficient to assume that the order of

service of customers from classes in R is FIFO. The question that needs to be resolved is whether the order of service of customers from R^c, the complement of R, can affect the time average in (16). Under the assumption that steady state exists, it is shown in [8, 7] that the order of service of customers in R^c does not affect the time average in (16). Since we do not make any assumptions about the ergodicity of the policies, however, we need a different approach here.

Assume that a work-conserving policy π gives nonpreemptive priority to classes in R. Let τ_{im} be the time instant the mth arriving customer from a class i in R^c is scheduled for service and let $\bar{\tau}_{im}$ be the first time in the interval $[\tau_{im} + S_{im}, \infty)$ at which there are no customers from classes in R in the system. Observe that either $\bar{\tau}_{im} = \tau_{jl}$ for some $l = 1, 2, \cdots$ and $j \in R^c$, or $\bar{\tau}_{im}$ is the end of a busy cycle. Also, let τ_k be the first time within the kth busy period that a customer from a class in R^c is scheduled for service. Let Γ_k denote the time instant when the kth busy period begins. At time $T = \Gamma_{k+1}$, the integral in (16) can be written as follows:

$$\int_0^{\Gamma_{k+1}} V_R^{(\pi)}(t)dt = \sum_{m=1}^{k} \int_{\Gamma_m}^{\Gamma_{m+1}} V_R^{(\pi)}(t)dt = \sum_{m=1}^{k} \bar{U}_m^{(\pi)} \ , \tag{17}$$

where $\bar{U}_m^{(\pi)} \triangleq \int_{\Gamma_m}^{\Gamma_{m+1}} V_R^{(\pi)}(t)dt$. We may rewrite $\bar{U}_k^{(\pi)},\ k = 1, 2, \cdots$ in the form

$$\bar{U}_k^{(\pi)} = U_k^{(\pi)} + \sum_{i \in R^c} \sum_{m=M_{i(k-1)}+1}^{M_{ik}} U_{im}^{(\pi)}, \tag{18}$$

where

$$U_{im}^{(\pi)} \triangleq \int_{\tau_{im}}^{\bar{\tau}_{im}} V_R^{(\pi)}(t)dt \ \text{ and } \ U_k^{(\pi)} \triangleq \int_{\Gamma_k}^{\tau_k} V_R^{(\pi)}(t)dt \ .$$

Using this decomposition of $\bar{U}_k^{(\pi)}$ we can show that the distribution of the sequence $\bar{U}_k^{(\pi)},\ k \geq 1$, cannot be affected by a nonpreemptive, work-conserving policy that gives priority to classes in R. This is the content of the next lemma.

Lemma 4 *For any nonpreemptive, work-conserving policy π that gives priority to classes in the set R over classes in the set R^c, the sequence $\{\bar{U}_k^{(\pi)}, k \geq 1\}$ consists of i.i.d. random variables, identically distributed to the sequence $\{\bar{U}_k^{(\pi_f)},\ k \geq 1\}$ generated by the policy π_f that gives*

nonpreemptive priority to classes in R, serves customers from class in R in FIFO order and customers from classes in R^c in FIFO order.

Proof: Recall that π is work-conserving, the arrival process is Poisson (and independent of the service requirements,) and the service requirements are independent random variables. Observe that the value of $V_R^{(\pi)}(t)$ for $t \in [\tau_{im}, \bar{\tau}_{im})$ depends only on S_{im} and the service times and the arrival process of customers from classes in R in that time interval. It follows that the random variables $\{U_{im}^{(\pi)},\ i \in R^c,\ m \geq 1\}$ are independent and also independent of $\{U_k^{(\pi)},\ k \geq 1\}$. Moreover, the distribution of $U_{im}^{(\pi)}$ is independent of the policy π. Since π is nonidling and gives nonpreemptive priority to customers in R, $\{U_k^{(\pi)}, k \geq 1\}$ is a sequence of i.i.d. random variables and the distribution of $\{U_k^{(\pi)}, k \geq 1\}$ is independent of the policy. Observe also that the summation in (18) includes the $same$ customers, $independent$ of the policy. This observation and the fact that $\{U_k^{(\pi)}, k \geq 1\}$ and $\{U_{im}^{(\pi)},\ i \in R^c,\ m \geq 1\}$ are independent sequences of independent random variables, imply that $\{\bar{U}_k^{(\pi)}, k \geq 1\}$ is a sequence of independent random variables and that the distribution of $\{\bar{U}_k^{(\pi)}, k \geq 1\}$ is the same for all work-conserving policies that give nonpreemptive priority to customers in R. To conclude the proof, observe that under the policy that serves customers in R^c in $FIFO$ order, the previous sequence consists of i.i.d. random variables. □

The main result regarding the performance space for the $M|G|1$ queueing system is given by the following theorem.

Theorem 2 *Suppose that $\sum_{i=1}^N \rho_i < 1$, and for some constant $\gamma > 2$, we have $s_i^{(\gamma)} \leq \infty$, for $i = 1, \ldots, N$. Then for any work-conserving, nonpreemptive policy π, and for all $R \subset \{1, \cdots, N\}$ we have that*

$$\liminf_{k\to\infty} \sum_{i\in R} \rho_i \left(\frac{\sum_{m=1}^k W_{im}(\pi)}{k} \right) \geq F(R) \quad a.e., \tag{19}$$

where $F(R)$ is a constant, independent of the scheduling policy. The limit exists and equality is achieved, if classes in R have nonpreemptive priority over classes in R^c.

Proof: Using Lemma 4 and assuming that customers from classes in R^c are served according to a FIFO policy, we conclude as in [8, Chapter 6], that under any nonpreemptive, work-conserving policy π that gives priority to classes in R,

$$\lim_{k\to\infty} \frac{\int_0^{\Gamma_k} V_R^{(\pi)}(t)dt}{\Gamma_k} = \frac{w_0}{1-\sum_{i\in R}\rho_i} + \sum_{i\in R}\rho_i s_i \stackrel{\Delta}{=} \bar{F}(R) \quad a.e., \tag{20}$$

where w_0 is the average residual work of the customer in service. Standard arguments can be used now, to show that when classes in R have nonpreemptive priority,

$$\lim_{T\to\infty} \frac{\int_0^{T} V_R^{(\pi)}(t)dt}{T} = \bar{F}(R) \quad a.e. \tag{21}$$

Since $V_R^{(\pi)}(t)$ is minimized by a policy that gives nonpreemptive priority to classes in R, we conclude that under any nonpreemptive, work-conserving policy,

$$\liminf_{T\to\infty} \frac{\int_0^{T} V_R^{(\pi)}(t)dt}{T} \geq \bar{F}(R) \quad a.e. \tag{22}$$

The limit exists and equality is achieved if classes in R have nonpreemptive priority over classes in R^c. Following the methodology of Theorem 1 and using (22) we have that

$$\liminf_{k\to\infty} \sum_{i\in R}\rho_i \left(\frac{\sum_{m=1}^{k} W_{im}(\pi)}{k}\right) \geq \bar{F}(R) - \sum_{i\in R}\frac{1}{2}\lambda_i s_i^{(2)} \stackrel{\Delta}{=} F(R) \quad a.e.$$

□

5 Conclusions.

We considered a $GI|G|1$ queueing system with N priority classes. The system operates under a scheduling policy which may be nonergodic or for which ergodicity may not be verifiable in advance. Adaptive and time-dependent scheduling rules represent an important class of such policies. We have derived an invariance relation that "conserves" a weighted sum of appropriate Cesaro sums of waiting times. We have also demonstrated inequalities that characterize further the achievable performance region, for the special case of Poisson arrivals.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments, which considerably improved the presentation of this paper.

References

[1] P.P. Bhattacharya, L. Georgiadis, P. Tsoucas, and I. Viniotis. Optimality and finite time behavior of an adaptive multi-objective scheduling algorithm. Technical report RC 15967, IBM T.J. Watson Research Center, 1990. Mathematics of Operations Research, to appear.

[2] O. J. Boxma and W. P. Groenendijk. Waiting times in discrete-time cyclic-service systems. *IEEE Transactions on Communications*, 36:164–170, 1988.

[3] S. L. Brumelle. On the relation between customer and time averages in queues. *J. Appl. Prob.*, 8:508–520, 1971.

[4] E. G. Coffman and I. Mitrani. A characterization of waiting time performance realizable by single server queues. *Operations Research*, 28:810–821, 1980.

[5] J. W. Cohen. *The Single Server Queue*, volume 8 of *Applied Mathematics and Mechanics*. North-Holland, Amsterdam, 1969.

[6] A. Federgruen and H. Groenevelt. Characterization and optimization of achievable performance in general queueing systems. *Operations Research*, 36:733–741, 1988.

[7] A. Federgruen and H. Groenevelt. M/G/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science*, 34(9):1121–1138, 1988.

[8] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems.* Academic Press, New York, 1980.

[9] L. Georgiadis and I. Viniotis. On the conservation law and the performance space of single server systems. Technical report RC 15673, IBM T.J. Watson Research Center, 1990. Operations Research, to appear.

[10] S. Ghahramani and R. W. Wolff. A new proof of finite moment conditions for $gi/g/1$ busy periods. *Queueing Systems*, 4:171–178, 1989.

[11] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Applications.* Academic Press, New York, 1980.

[12] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research*, volume I. McGraw-Hill, New York, 1982.

[13] L. Kleinrock. *Queueing Systems*, volume 2. John Wiley, 1976.

[14] R. W. Wolff. *Stochastic Modeling and the Theory of Queues.* Industrial and Systems Engineering. Prendice Hall, 1989.

[15] R.W. Wolff. Work-conserving priorities. *Journal of Applied Probability*, 7:327–337, 1970.