Queueing Models and
Service Management
Vol. 2, No. 2, page 124-152, 2019

QMSM

A Discrete Workload Conservation Law with Applications to Discrete-Time Queueing Systems

Muhammad El-Taha

Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, ME 04104-9300, USA

(Received Febuary 2019, accepted April 2019)

Abstract: In this article we give a general conservation law for discrete-time systems. For queueing models in discrete-time the law relates the asymptotic average workload in the system to the conditional asymptotic average sojourn time and service times distribution function. The law is valid for multi-server systems, and anticipating and non-anticipating scheduling disciplines. The proof uses a sample-path discrete-time version for the well-known relation $H=\lambda G$. We apply the law to several queueing models and show that for anticipative and non-anticipative scheduling rules, the unconditional delay in a queue is related to the covariance of service times and queueing delays. The results in this article complement similar results for continuous-time models.

Keywords: Conservation law, multi-server queues, sample path analysis, scheduling, workload invariance.

1. Introduction

In this article we establish a discrete-time general law that relates the average work in a queueing system to the conditional waiting times distribution. This law will be used to give several applications including multi-server and single server systems. We note that this law is a discrete-time version of the continuous time law given by El-Taha [7]. Specifically, the law relates the long-run average unfinished work in the system to the conditional long-run average waiting (sojourn) time distribution and the empirical distribution function of service times. The law allows for a wide class of scheduling work-conserving disciplines including anticipative and non-anticipative rules. Under the additional condition that the workload in the system is *invariant* with respect to scheduling rules, which includes all single server systems and a class of multi-server systems, see El-Taha [6] and references there in, this law becomes a conservation law. We use sample path analysis, thus provide a proof under weak conditions that require primarily existence of limits and no probabilistic assumptions. Our method can be easily extended to the stationary framework by invoking the appropriate $SLLN$ and/or the ergodic theorem. There are a few research articles, we are

Corresponding author
Email : el-taha@maine.edu

aware of, that deals with conservation laws that include anticipative scheduling systems. O'Donovan [17] obtains an expression for a conservation law for the $GI/G/1$ single server model with anticipative service disciplines. Ayesta [1] gives a conservation law for work conserving single server queues that covers both anticipative and non-anticipative disciplines for general inter-arrival and general service times. El-Taha [7] generalizes these conservation laws for general input output systems including multi-server queueing models. This article gives discrete-time version of the law given by El-Taha [7].

Conservation laws that deal with non-anticipative scheduling disciplines relate the mean work in a queueing system to the waiting times distribution function. Kleinrock *et. al.* [16] introduce a preliminary version of the law for the $M/G/1$ model. Multi-server non-anticipative conservation laws are discussed in Heyman and Sobel [13] who extends Kleinrock's *et. al.* [16] law to multi-server models. See also Wolf [20, p. 455] who discusses the law given by Heyman and Sobel [13]. For more information on conservation laws the reader may consult Chapter 6 of El-Taha and Stidham [8], Chapter 10 of Wolf [20], Chapter 11 of Heyman and Sobel [13]. Additional articles that address conservation laws of multi-server systems include Bartsch and Bolch [3], Dacre and Glazerbrook [4], Federgruen and Groenevelt [9], Green and Stidham [11] and Shanthikumar and Yao [18].

The issue of the invariance of workload comes up in the context of establishing conservation laws. Workload invariance is necessary to establish the validity of the conservation law. In a $G/G/1$ single-server model, the workload in the system is *invariant* at all time instants for all sample paths for non-idling work-conserving scheduling rules (see, for example, Gelenbe and Mitrani [10, Theorem 6.1]. This invariance property is based on the simple observation that the workload in the system at any time instant is unaffected by the scheduling rule. This is true because the work in the system will be reduced at a unit rate per unit time as long as there is work in the system regardless of the scheduling rule. For single server systems (e.g., Kleinrock [16], Kleinrock [15], Gelenbe and Mitrani [10]) establish conservation laws that are valid across a large class of work conserving scheduling rules. Thus, for single server queues any a law that equates the mean workload in the system to other system quantities would be a conservation law.

The situation is different for multi-server systems. The *invariance* property of the workload in the system does not hold in a sample path sense at every time instant. Counter-examples where the invariance property in multi-server systems does not hold are given by El-Taha [6]. However, a sufficient condition for the *invariance* property, established by El-Taha [6], is that all service times are *i.i.d.*. This *i.i.d.* condition is more general than previously known, it is still restrictive in the sense that it does not permit scheduling rules where different classes have different service rates.

For multi-server systems it is possible for a law relating workload to waiting times to hold for each scheduling rule but that workload is not the same for all rules. This notion is first observed by Wolf [20, p. 455]. Therefore, for multi-server systems establishing a conservation law necessarily requires establishing invariance at the same time. El-Taha [7] generalizes a conservation law by Ayesta [1] to multi-server systems. This is accomplished by separating the two issues involved in establishing a conservation law. First the law is shown to hold for every discipline, and second the workload is shown to be discipline invariant under the $i.i.d.$ service times conditions given by El-Taha [6]. We follow the same approach in this article. The laws established in Theorem 3.1 and Theorem 3.2 are valid for every scheduling rule. In these laws the asymptotic average workload EV may be different for different disciplines in multi-server systems. However for single server systems and multi-server systems with $i.i.d.$ service times the asymptotic average workload EV is invariant and therefore Theorem 3.1 and Theorem 3.2 give conservation laws.

We point out that conservation laws rely on the property that the workload in the system is unaffected by the scheduling discipline (invariant) as long as the discipline is work conserving. The discipline can be preemptive or non-preemptive, non-anticipative or anticipative. This is true for single server systems where work, when present, is processed at a unit rate regardless of the scheduling discipline, see Gelenbe and Mitrani [10]. The situation is different for multi-server systems where workload invariance to scheduling disciplines cannot be asserted at every time instance. However, there are multi-server systems where mean workload in the system is invariant with respect to scheduling rules. The contribution of this article is to establish a law that is valid for a wide range of stochastic discrete-time systems, including multi-server and single server systems operating under a wide class of work conserving scheduling disciplines that include non-anticipative as well as anticipative disciplines; thus providing a discrete-time version of the law given by El-Taha [7]. This law would be a conservation law for single server systems, and for multi-server systems under additional conditions, e.g. El-Taha [6], that guarantee the workload is invariant with regard to scheduling rules. One interesting application states that the unconditional delay (time in queue) for an anticipating scheduling rule is less than the unconditional delay for a non-anticipating rule when the covariance of the service times and delay for anticipative rules is non-negative and vise versa.

This article is organized as follows. In Section 2 we give a brief discrete-time review of the well-known relation $H = \lambda G$ and discuss a related discrete-time conservation law that relates unfinished workload, and mean waiting and service times first two moments. In Section 3 we give a general discrete-time conservation law based on weak conditions. This law relates unfinished workload to the conditional waiting times and the service time distribution functions. An interesting feature of the law is that the derivative of conditional waiting in the continuous version is replaced by the forward/backward difference. We then specialize the law to queueing systems with anticipative scheduling rules. This result also extends to multi-server systems a conservation law given by Ayesta [1] for single server queues. In Section 4 we give several special cases and applications that include a specialization of the law to multi-server non-anticipative systems and give a pure sample path proof of this result. We also give an extension of the conservation law results to multi-class discrete-time systems; and give applications to non-preemptive anticipative disciplines and self service systems. Moreover, we show that the conservation law can be used to compare the mean waiting times of anticipative and non-anticipative scheduling rules. Additionally, we establish the optimality of the $c\mu$-rule for discrete-time queues. The Appendix contains a list of definitions and notation, and results that are needed in proofs of theorems in this article.

2. Discrete-Time $H = \lambda G$

Consider a deterministic sequence of integer time points, $\{T_k, k \geq 1\}$, with $0 \leq T_k \leq T_{k+1} < \infty$, $k \geq 1$, and define $A(n) := \max\{k \geq 1 : T_k \leq n\}$, $n \geq 0$, so that $A(n)$ is the number of points in $[0,n]$. We assume that $T_k \to \infty$ as $k \to \infty$, so that there are only a finite number of events in any finite time interval ($A(n) < \infty$ for all $n \geq 0$), and we note that $A(n) \to \infty$ as $n \to \infty$, since $T_k < \infty$ for all $k \geq 1$. Associated with each time point T_k, there is a function $f_k : I \to I$; where I is the set of non-negative integers. The bivariate sequence $\{(T_k, f_k(\cdot)), k \geq 1\}$ constitutes the basic data, in terms of which the behavior of the system is described. Now, let $f_k(n)$ denote the rate at which customer k incurs cost at time n, $k \geq 1$, $n \geq 0$, and define

$$H(n) := \sum_{k=1}^{\infty} f_k(n), \, n \geq 0, \tag{1}$$

$$G_k := \sum_{n=0}^{\infty} f_k(n), \, k \geq 1, \tag{2}$$

so that $H(n)$ is the total cost rate at time n and G_k is the total cost incurred by customer k.

To motivate our approach we consider Little's formula $L = \lambda W$, which relates mean number of customers in a system L to the product of the arrival rate λ and mean waiting time in the system W, and has an economic interpretation that sheds light on its generality and also suggests the current extension. Suppose customer k incurs a cost of one dollar per unit time while in the system (i.e., while $T_k \leq n \leq D_k$) and zero cost otherwise, where T_k (D_k) are the arrival (departure) time of customer k. Let $f_k(n) := 1\{T_k \leq n \leq D_k\}$. Then we can interpret the function $f_k(n)$ as the cost rate of customer k at time n. Under this interpretation, $L(t) = \sum_{k=1}^{\infty} f_k(n)$ is the total cost rate at time n and $W_k = \sum_{n=0}^{\infty} f_k(n)$ is the total cost incurred by customer k, so that $L = \lambda W$ says that the long-run average cost per unit time equals the arrival rate of customers times the long-run average cost per customer. The generalization to $H = \lambda G$ arises naturally if one allows a more general cost-rate function than the indicator of the event $\{T_k \le n \le D_k\}$. With $H(n)$ and G_k defined by (1) and (2), respectively, define the following limiting averages, when they exist:

$$\lambda := \lim_{n\to\infty} n^{-1}A(n),$$

$$H := \lim_{n\to\infty} n^{-1}\sum_{j=0}^{n} H(j),$$

$$G := \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} G_k .$$

Following Stidham [19] and Heyman and Stidham [14], suppose that the bivariate sequence $\{(T_k, f_k(\cdot)), k \ge 1\}$ satisfies the following condition:

Condition A. There exists a sequence $\{W_k, k \ge 1\}$ such that $W_k / T_k \to 0$ as $k \to \infty$; and $f_k(n) = 0$ for $n \notin [T_k, T_k + W_k]$.

Condition A says that all the cost associated with the k^{th} customer is incurred in a finite time interval beginning at the arrival of the customer, and that the lengths of these intervals cannot grow at the same rate as the points themselves, as $k \to \infty$. This is a stronger-than-necessary condition for $H = \lambda G$ (See El-Taha and Stidahm [8] for details), but it is satisfied in most applications to queueing systems, in which the time points T_k and $T_k + W_k$ correspond to customer arrivals and departures, respectively, and it is natural to assume that customers can only incur cost while they are physically present in the system.

The proof of the discrete-time $H = \lambda G$ follows the same steps as the continuous-time case given by El-Taha and Stidahm [8].

Theorem 2.1. *Suppose $n^{-1}A(n) \to \lambda$ as $n \to \infty$, where $0 \le \lambda \le \infty$, and Condition A holds. Then*

(i) if $m^{-1}\sum_{k=1}^{m} G_k \to G$ as $m \to \infty$, where $0 \le G \le \infty$, then $n^{-1}\sum_{j=0}^{n} H(j) \to H$ as $n \to \infty$, and $H = \lambda G$, provided λG is well defined;

(ii) if $n^{-1}\sum_{j=0}^{n} H(j) \to H$ as $n \to \infty$, where $0 \le H \le \infty$, then $m^{-1}\sum_{k=1}^{m} G_k \to G$ as $m \to \infty$, and $H = \lambda G$, provided $\lambda^{-1}H$ is well defined.

The $G/G/c$ queue: A conservation law between workload and waiting time

We now show how to use $H = \lambda G$ to derive a relation between the time-average workload and the customer-average waiting time in the queue in a multi-server system with a non-preemptive queue discipline. Consider the discrete-time *G/G/c* queue. The input data consists of the sequence $\{(T_k, S_k), k \ge 1\}$, where T_k is the arrival instant and S_k the work requirement of customer k. Let $A(n) := \max\{k : T_k \le n\}$ denote the number of arrivals in $[0, n]$. Customers need not be served in order of arrival, but a server is never idle when customers are waiting. In this application we shall assume that each server works at unit rate and that the queue discipline is non-preemptive. Let $W_{q,k}$ denote the waiting time in queue (delay) of the k^{th} customer. Assume each of the following limits exists and is finite:

$$ES := \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} S_k ,$$

$$ES^2 := \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} S_k^2 ,$$

$$EW_q := \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} W_{q,k} ,$$

$$ESW_q := \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} S_k W_{q,k} .$$

Here, ES is the long-run average service time, ES^2 is the long-run empirical second moment of service times, and EW_q is the long-run average waiting time in queue (excluding time in service). Note that these are sample-path averages, even though we use a notation suggestive of expectations. Let

$$f_k(n) = S_k 1\{T_k \le n < T_k + W_{q,k}\}$$
$$+(S_k - (n - T_k - W_{q,k}))\, 1\{T_k + W_{q,k} \le n \le T_k + W_{q,k} + S_k\} .$$

That is, $f_k(n)$ is the work remaining to be done for the k^{th} customer at time n. Thus

$$V(n) = \sum_{k=1}^{\infty} f_k(n)$$

is the total amount of unfinished work in the system at time n. Let

$$EV := \lim_{n\to\infty} n^{-1}\sum_{j=0}^{n} V(j) ,$$

when the limit exists. Now let $H(n) = V(n)$ and

$$G_k = \sum_{n=0}^{\infty} f_k(n) = S_k W_{q,k} + (S_k^2 + S_k)/2 ;$$

$$G = \lim_{m\to\infty} m^{-1}\sum_{k=1}^{n} [S_k W_{q,k} + (S_k^2 + S_k)/2] = ESW_q + E(S^2 + S)/2 .$$

Since λ, ES, and EW_q are well defined and finite, Conditio A holds with $W_k = W_{q,k} + S_k$, the waiting time of the k^{th} customer in the system. Applying $H = \lambda G$, we conclude that

$$EV = \lambda ESW_q + \lambda E(S^2 + S)/2 . \qquad (3)$$

The first term is the total amount of work associated with customers waiting in the queue, and the second term is the residual service time. In contrast, for continuous-time models the the residual service time is given by $\lambda ES^2/2$. Now suppose the sequences $\{S_k, k \ge 1\}$ and $\{W_{q,k}, k \ge 1\}$, are *asymptotically pathwise uncorrelated*, that is,

$$ESW_q = \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} S_k W_{q,k} = ES \cdot EW_q \,. \tag{4}$$

This will be true w. p. 1 for stochastic models with *service-time independent* scheduling rules, that is, models in which the rule for selecting the next job to process does not use information about the processing times of jobs. The *FIFO* queue discipline is an example of such a rule. In this case (3) reduces to:

$$EV = \lambda ESEW_q + \lambda E(S^2 + S)/2\,. \tag{5}$$

We will refer to this law later in this article.

3. The General Conservation Law

In this section we give the discrete-time general conservation law that states that in queueing context the asymptotic average workload in the system is related to the conditional asymptotic average sojourn time and the service times distribution function. The analysis uses a sample-path approach as in El-Taha and Stidham [8] and utilizes $H = \lambda G.$ Our problem setup is given at great generality. Our interpretations are given in the context of queueing systems.

For $k = 1,2,\ldots,$ consider the deterministic non-negative triplet sequence $\{T_k, S_k, \mathcal{W}_k(a,x)\}$; such that $0 \le T_k \le T_{k+1} < \infty,$ and $0 \le S_k < \infty$ are sequences of non-negative integer-valued numbers. We assume that $T_k \to \infty$ as $k \to \infty$ so that there are only a finite number of events in any finite time interval. For all non-negative integers a, x such that $a \le x \le S_k,$ let $\mathcal{W}_k(a,x)$ be a monotone, non-decreasing in $a,$ integer-valued function such that $\mathcal{W}_k(0,x) = 0$ at T_k and $\mathcal{W}_k(S_k, S_k) = W_k < \infty.$ When $a = x,$ we shall use the notation $\mathcal{W}_k(x,x) = \mathcal{W}_k(x)$. Moreover, for $k = 1,2,\ldots,$ let

$$V_k(n) = (S_k - \boldsymbol{A}_k(n - T_k, S_k))1\{T_k \le n \le T_k + W_k\},$$

where $\boldsymbol{A}_k(\tau, S_k) := \mathcal{W}_k^{-1}(\tau, S_k) = \min\{a : \mathcal{W}_k(a, S_k) \ge \tau\},$ $0 \le \tau \le W_k,$ is the generalized inverse of $\mathcal{W}_k(a, S_k)$. Moreover, for $x \in I,$ and $n \in I,$ (I is the set of non-negative integers), let

$$V(n) := \sum_{k=1}^{\infty} V_k(n)\,;$$

$$f_m(x) := m^{-1}\sum_{k=1}^{m} 1\{S_k = x\}\,;$$

$$F_m(x) := m^{-1}\sum_{k=1}^{m} 1\{S_k \le x\} = \sum_{y=0}^{x} f_m(y)\,;$$

$$\mathcal{W}_m(a,x) := \frac{\sum_{k=1}^{m} \mathcal{W}_k(a,x)\ 1\{S_k = x\}}{\sum_{k=1}^{m} 1\{S_k = x\}}; \tag{6}$$

and define the following limits when they exist

$$f(x) = \lim_{m\to\infty} f_m(x);$$

$$F(x) = \lim_{m\to\infty} F_m(x);$$

$$EW = \lim_{m\to\infty} m^{-1} \sum_{k=1}^{m} W_k;$$

$$\mathcal{W}(a,x) = \lim_{m\to\infty} \mathcal{W}_m(a,x);$$

$$EV = \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n} V(j).$$

We assume that $\mathcal{W}_m(a,x)$ and $F_m(x)$ converge respectively to $\mathcal{W}(a,x)$ and $F(x)$ uniformly in x as $m\to\infty$. Note that we use the suggestive notation EV and EW to indicate long-run averages. Later, we shall use the notation E[.] to indicate expected values.

We need additional definitions and notation. For some function g, let $\Delta_x^f g(x) = g(x+1) - g(x)$ and $\Delta_x^b g(x) := g(x) - g(x-1)$ denote the forward and backward first order differences respectively. Note that $\Delta_x^f g(x) = \Delta_x^b g(x+1)$. We will drop the subscript x when dealing with single valued functions. A queueing discipline is said to be *anticipative* if for all x, $\Delta_x^f \mathcal{W}(a,x) \neq 0$ (i.e., $\Delta_x^f \mathcal{W}(a,x) > 0$ or $\Delta_x^f \mathcal{W}(a,x) < 0$), and *non-anticipative* if $\Delta_x^f \mathcal{W}(a,x) = 0$ for all $a \le x$. We say a queueing system is work conserving if no work is created or destroyed while in the system. Moreover, we allow preemptive and non-preemptive queueing disciplines.

As in El-Taha [7], at this level of generality, think of the system as a black box where entities arrive at times $\{T_k\}$ at rate λ. Associated with each entity is a clock (variable) with $\{S_k\}$ units. The clock is set to 0 at time T_k (arrival) and advances to S_k exactly at departure time $T_k + W_k$, that is $\mathcal{W}_k(0,S_k) = 0$ at T_k and $\mathcal{W}_k(S_k,S_k) = W_k$, equivalently $\boldsymbol{A}_k(0,S_k) = 0$ and $\boldsymbol{A}_k(W_k,S_k) = S_k$. The timer can stop and start (interruptions) countably may times, and it can slow down and speed up, but cannot restart and no extra time can be added (work-conserving).

In the context of a queueing system, the above quantities have the following interpretation: We think of $\{T_k, S_k\}$ as the k^{th} customer arrival time, and service requirement respectively. Moreover, W_k is interpreted as the sojourn time (waiting time in system) of the k^{th} arrival, $\mathcal{W}_k(a,x)$ is interpreted as the k^{th} customer waiting time until it receives a units of service given that it requires $x \ge a$ units of service. In this case, we may write $\mathcal{W}_k(a,x) = \max\{u \geq 0 : \sum_{n=0}^{u} \delta_k(n+T_k) = a\}$, where $\delta_k(n) = 1$ if k^{th} arrival is in service at time n and 0 otherwise. It is interesting to note that $\mathcal{W}_k^{-1}(\tau, S_k)$ represents the attained service time, and thus $V_k(n)$ represents the residual service time of the k^{th} arrival at time n, $V(n)$ represents the total unfinished work in the system at time n. Moreover, $\mathcal{W}(a,x)$ represents the long run average conditional sojourn time of all customers with a units of attained service among all customers with $x \geq a$ units of service requirement; and EV represents the asymptotic average unfinished workload in the system. Additionally, $F(x)$ is the asymptotic frequency distribution of the service times. These interpretations are not necessary for the next two results. Now we state and prove the main theorem.

Theorem 3.1. *Suppose that all limits are well defined, the arrival rate is* $0 < \lambda < \infty$, *and that* $\frac{W_k}{T_k} \to 0$ *as* $k \to \infty$. *Then*

$$\lambda \sum_{x=0}^{\infty} \sum_{a=0}^{x} \mathcal{W}(a,x) f(x) = EV. \tag{7}$$

Proof. We use the discrete-time $H = \lambda G$, given in Section 2, to prove this result. Let $\tilde{f}_k(n) = (S_k - \boldsymbol{A}_k(n - T_k, S_k))1\{T_k \leq n \leq T_k + W_k\})$ be the remaining service requirement for k^{th} arrival, so that $H(n) = \sum_{k=1}^{\infty} \tilde{f}_k(n)$ represents the total unfinished work in the system at time n, thus

$$H := \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n} H(j) = EV.$$

Here, $g_k = \sum_{n=0}^{\infty} \tilde{f}_k(n)$ is the cumulative contribution of the k^{th} customer to the total unfinished work in the system, so that

$$\begin{aligned} g_k &= \sum_{n=0}^{\infty} (S_k - \boldsymbol{A}_k(n - T_k, S_k))1\{T_k \leq n \leq T_k + W_k\}) \\ &= \sum_{n=T_k}^{T_k+W_k} (S_k - \boldsymbol{A}_k(n - T_k, S_k)) \\ &= \sum_{\tau=0}^{W_k} (S_k - \boldsymbol{A}_k(\tau, S_k)) \\ &= \sum_{a=0}^{S_k} \mathcal{W}_k(a, S_k)\,. \end{aligned}$$

Now divide the numerator and denominator of (6) by m, and use the definition of $f_m(x)$ to obtain

$$\mathcal{W}_m(a,x) f_m(x) = m^{-1} \sum_{k=1}^{m} \mathcal{W}_k(a,x)\, 1\{S_k = x\}\,. \tag{8}$$

Sum both sides of (8) using $\sum_{x=0}^{\infty}\sum_{a=0}^{x}$. First summing r.h.s. of (8) then taking limit as $m\to\infty$ leads to

$$\begin{aligned}
&\lim_{m\to\infty}\sum_{x=0}^{\infty}\sum_{a=0}^{x} m^{-1}\sum_{k=1}^{m}\boldsymbol{w}_k(a,x)\,1\{S_k=x\}\\
&=\lim_{m\to\infty} m^{-1}\sum_{k=1}^{m}\sum_{x=0}^{\infty}\sum_{a=0}^{x}\boldsymbol{w}_k(a,x)\,1\{S_k=x\}\\
&=\lim_{m\to\infty} m^{-1}\sum_{k=1}^{m}\sum_{a=0}^{S_k}\boldsymbol{w}_k(a,S_k)\\
&=\lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} g_k\\
&=G.
\end{aligned}$$

Now sum the l.h.s. of (8) and take limits to get

$$\lim_{m\to\infty}\sum_{x=0}^{\infty}\sum_{a=0}^{x}\boldsymbol{w}_m(a,x)f_m(x)=\sum_{x=0}^{\infty}\sum_{a=0}^{x}\boldsymbol{w}(a,x)f(x).$$

Thus, we have shown that

$$G=\sum_{x=0}^{\infty}\sum_{a=0}^{x}\boldsymbol{w}(a,x)f(x).$$

Proof of the theorem follows by using $H=\lambda G$.

Remarks.

(i) The sufficient condition $\frac{W_k}{T_k}\to 0$ as $k\to\infty$ is needed to apply $H=\lambda G$. A weaker sufficient condition is $W=\lim_{n\to\infty}n^{-1}\sum_{k=1}^{n}W_k$ is well-defined and $<\infty$. In queueing systems, this sufficient condition is satisfied if the stronger condition $ES^2<\infty$.
(ii) Theorem 3.1 is valid for a wide range of queueing disciplines. The discipline can be non-anticipative or anticipative. When the discipline is non-anticipative, it is valid for all work conserving non-preemptive disciplines as well as preempt-resume scheduling rules.
(iii) Under additional conditions, Theorem 3.1 can also be an invariance relation in the sense that EV is invariant for all scheduling disciplines. This is true for any $G/G/c$ work-conserving non-anticipative queueing system, with possibly batch arrivals, if service times are $i.i.d.$ and the discipline is non-preemptive, for more details see El-Taha [7, 6]. Theorem 3.1 is invariant for single server $G/G/1$ systems, Gelenbe and Mitrani [10].

In the next result we give an alternative form of relation (7) under an additional mild condition. Note that by definition $ES^2=\lim_{m\to\infty}m^{-1}\sum_{k=1}^{m}S_k^2$. Because we use a deterministic framework the fact that $ES^2=\sum_0^{\infty}x^2f(x)<\infty,$ needs justification. It can be shown that this follows from the assumption

$$\lim_{\alpha\to\infty}\lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} S_k^2\, 1\{S_k > \alpha\} = 0\,;$$

which is a sample path analogue of uniform integrability; see El-Taha [5].

Theorem 3.2. *Under the conditions that* $ES^2 = \sum_0^\infty x^2 f(x) < \infty$, *and* $\lim_{x\to\infty} \mathcal{W}(a,x)/x < \infty$, *we have*

$$\lambda\sum_{x=0}^{\infty}\left[\mathcal{W}(x+1)+\sum_{a=0}^{x}\Delta_x^f \mathcal{W}(a,x)\right]F^c(x) = EV. \tag{9}$$

where $F^c(x) = 1 - F(x)$ *is the complement of the service times distribution function.*

Proof. We use Theorem 3.1 and discrete integration by parts; see Lemma 6.1 in the Appendix. Let $v(x) = \sum_{a=0}^{x}\mathcal{W}(a,x)$ and $\Delta^f u(x) = f(x)$, so that

$\Delta^f v_x = \sum_{a=0}^{x}\Delta_x^f\mathcal{W}(a,x) + \mathcal{W}(x+1)$, and $u(x) = -F^c(x-1)$. Now

$$G = \sum_{x=0}^{\infty}[\sum_{a=0}^{x}\mathcal{W}(a,x)]f(x)$$

$$= -F^c(x-1)\sum_{a=0}^{x}\mathcal{W}(a,x)\Bigg|_{x=0}^{\infty} + \sum_{x=0}^{\infty}F^c(x)\left[\sum_{a=0}^{x}\Delta_x^f\mathcal{W}(a,x)+\mathcal{W}(x+1)\right].$$

We need to show that

$$\lim_{x\to\infty} F^c(x-1)\sum_{a=0}^{x}\mathcal{W}(a,x) = 0\,.$$

Note that $\mathcal{W}(a,x) \le \mathcal{W}(x,x) = \mathcal{W}(x)$. Therefore

$$\lim_{x\to\infty} F^c(x-1)\sum_{a=0}^{x}\mathcal{W}(a,x) \le \lim_{x\to\infty}\mathcal{W}(x)\, x\, F^c(x-1)\,.$$

Now, the condition that service times have finite second moments, and Lemma 6.3 in the Appendix imply that $\lim_{x\to\infty} x^2 F^c(x) = 0$. Therefore,

$$G = \sum_{x=0}^{\infty}F^c(x)\left[\sum_{a=0}^{x}\Delta_x^f\mathcal{W}(a,x)+\mathcal{W}(x+1)\right],$$

which complete the proof of the Theorem.

Remark. We point out that the conservation laws in Theorem 3.1 and Theorem 3.2 are the discrete-time equivalents of Corollary 2.2 and Theorem 2.1 of El-Tha [7] respectively. However the conditions needed for these results are reversed. Specifically, the conditions for Theorem 3.1 are similar to those of Theorem 2.1, not the stricter conditions of Corollary 2.2 of [7]. Similarly, conditions for Theorem 3.2 are similar to those of Corollary 2.2, not the weak conditions of Theorem 2.1 of [7].

Theorem 3.1 provides a generalization of the single server conservation law obtained by O'Donovan [17] to the discrete-time multi-servers case. Start with (7) and make the change of variable $a = x - r$ to obtain

$$EV = \lambda \sum_{x=0}^{\infty} \sum_{r=0}^{x} \mathcal{W}(x-r, x) f(x) .$$

Use discrete integration by parts, as in Lemma 6.1 in the Appendix, on the inner sum with respect to r, and interchange the order of summation to obtain

$$EV = \lambda \sum_{r=0}^{\infty} \sum_{x=r}^{\infty} -(r+1) \Delta_r^f \mathcal{W}(x-r, x) f(x) ;$$

which is a discrete-time multi-server equivalent of equation (9) in O'Donovan [17].

We have shown that Theorem 3.1 and Theorem 3.2 are valid for any input-output system. The next corollary shows that the condition $\lim_{x \to \infty} \mathcal{W}(x)/x < \infty$ can be shown to hold in stable non-preemptive multi-server queueing systems.

Corollary 3.3. *Consider a discrete-time multi-server non-preemptive queueing model $G/G/c$ with $\rho = \lambda ES/c < 1$, $ES^2 = \sum_0^{\infty} x^2 f(x) < \infty$, and a work conserving scheduling discipline. Then (7) and (9) hold.*

Proof. The proof is similar to Corollary 2.3 of El-Taha [7]. We give an outline. We need to show that

$$\lim_{x \to \infty} \mathcal{W}(x)/x < \infty .$$

Now a full busy period starts when all servers become busy for the first time and ends when for the first time a server becomes idle; so $\mathcal{W}(x)$ is bounded above by x plus the full busy period. Let $E[B_f(y)]$ be the expected full busy period of a $G/G/c$ queueing model initiated by a workload of size y. The key idea here is that this full busy period is smaller than the corresponding busy period of a single server queue with service rate $c\mu$. By Lemma 6.2 in the Appendix, for the single server model with $\lambda < c\mu$, ($c\mu$ is the service rate of a single server) the expected busy period $E[B(y)]$ is bounded above. This complete the proof.

The non-preemptive condition in the Corollary 3.3 can be relaxed to allow *limited preemption* where preemption of service times is limited to one full busy period, i.e. customers in service after the end of a full busy period will complete their service without further preemption. Moreover, the non-preemptive condition in the Corollary can be removed completely in stable single-server systems.

4. Applications

In this section we give several applications of the conservation laws, specifically, we show how this law is applied in multi-class systems, non-preemptive anticipative systems, and self service and loss systems. We also compare the performance of anticipative vs non-anticipative systems.

4.1. Waiting time in queue

Here we give versions of Theorem 3.1 and Theorem 3.2 that relate unfinished work in the queue and waiting time in queue (excluding service times). Let $w(a,x) = w_q(a,x) + a$ where $w_q(a,x)$ is interpreted as the conditional mean waiting time in queue (excluding service time) until the customer receives a units of service of customers with x units of service requirement. For preemptive disciplines $w_q(a,x)$ can be thought of as the wasted time, which is time in the system beyond the attained service. Using Theorem 3.1 and $w(a,x) = w_q(a,x) + a$, we obtain

$$\begin{aligned} EV &= \lambda \sum_{x=0}^{\infty}\sum_{a=0}^{x}(a + w_q(a,x))f(x) \\ &= \lambda \sum_{x=0}^{\infty}(\frac{x(x+1)}{2} + \sum_{a=0}^{x} w_q(a,x))f(x) \\ &= \frac{\lambda E(S^2+S)}{2} + \lambda \sum_{x=0}^{\infty}\sum_{a=0}^{x} w_q(a,x)f(x); \end{aligned} \tag{10}$$

where $\frac{E(S^2+S)}{2}$ represents the expected residual work in service in a queueing system. Let EU be the long-run average work in queue defined similar to EV, then $EU = EV - \frac{\lambda E(S^2+S)}{2}$ will satisfy the following law.

$$EU = \lambda \sum_{x=0}^{\infty}\sum_{a=0}^{x} w_q(a,x)f(x). \tag{11}$$

Similarly, using Theorem 3.2, Corollary 6.5 of the Appendix and $w(a,x) = w_q(a,x) + a$ for all $a \le x$, we obtain

$$\begin{aligned} EV &= \lambda \sum_{x=0}^{\infty}\left[w_q(x+1) + x + 1 + \sum_{a=0}^{x} \Delta_x^f [a + w_q(a,x)] \right] F^c(x) \\ &= \frac{\lambda E(S^2+S)}{2} + \lambda \sum_{x=0}^{\infty}\left[w_q(x+1) + \sum_{a=0}^{x} \Delta_x^f w_q(a,x) \right] F^c(x). \end{aligned} \tag{12}$$

Thus

$$EU = \lambda \sum_{x=0}^{\infty}\left[w_q(x+1) + \sum_{a=0}^{x} \Delta_x^f w_q(a,x) \right] F^c(x). \tag{13}$$

Combining (11) and (13) we obtain

Corollary 4.1. *Under the conditions of Theorem 3.2*

$$EU = \lambda \sum_{x=0}^{\infty}\sum_{a=0}^{x} w_q(a,x)f(x) = \lambda \sum_{x=0}^{\infty}\left[w_q(x+1) + \sum_{a=0}^{x} \Delta_x^f w_q(a,x) \right] F^c(x).$$

In the next subsection we focus on systems with non-anticipative scheduling rules.

4.2. Systems with non-anticipative scheduling rules

Systems with non-anticipative scheduling represent an important class of models covered by Kleinrock [15] and Heyman and Sobel [13]. This is represented as a special case of Theorem 3.1 as we show in the next result.

Theorem 4.2. *Consider a discrete-time multi-server queueing model $G/G/c$ with work conserving non-anticipative scheduling discipline. When all relevant limits are well defined, we have*

$$\lambda\sum_{x=0}^{\infty}\mathcal{W}(x)F^{c}(x-1)=EV. \tag{14}$$

where $F^{c}(x)$ is the complement of the service times distribution function.

Proof. Since the discipline is non-anticipative $\mathcal{W}(a,x)=\mathcal{W}(a,a)=\mathcal{W}(a)$, then using (7) and interchange limits, we obtain

$$\begin{aligned}EV&=\lambda\sum_{x=0}^{\infty}\sum_{a=0}^{x}\mathcal{W}(a,x)f(x)\\&=\lambda\sum_{a=0}^{\infty}\mathcal{W}(a)\sum_{x=a}^{\infty}f(x)\\&=\lambda\sum_{a=0}^{\infty}\mathcal{W}(a)F^{c}(a-1);\end{aligned}$$

which completes the proof.

The proof of Theorem 4.2 can be obtained from Theorem 3.2 as a special case by noting that $\Delta_x^f\mathcal{W}(a,x)=0$ for non-anticipative disciplines.

This Theorem is a discrete-time version of a similar result given by Heyman and Sobel [13]. However, we have to be careful when using this result in multi-server systems. For example, this result holds for the $FCFS$, $LCFS$, and SPT disciplines, but EV is not, necessarily, the same for all disciplines. The invariance property holds under the additional stochastic assumptions that the scheduling discipline is non-preemptive and the service times are $i.i.d.$; see El-Taha [6].

4.3. Systems with non-preemptive anticipative disciplines

In this subsection we assume that scheduling rules are non-preemptive and give new results. For non-preemptive possibly anticipative disciplines we can write $\mathcal{W}(a,x)=\mathcal{W}_q(x)+a$, $a>0$ where $\mathcal{W}_q(x)=\mathcal{W}_q(a,x)$ is the conditional mean waiting time in queue (excluding service time) of customers with x units of service requirement. An example of a discipline that is non-preemptive and anticipative is the shortest processing time (SPT), also known as shortest job first (SJF). Using Corollary 3.3, an argument similar to (10), and assuming a non-preemptive discipline, we obtain

$$EV = \lambda \sum_{x=0}^{\infty} \sum_{a=1}^{x} (a + \mathcal{w}_q(x)) f(x)$$

$$= \frac{\lambda E(S^2 + S)}{2} + \lambda \sum_{x=0}^{\infty} x \mathcal{w}_q(x) f(x). \tag{15}$$

This conservation law has been obtained in continuous time in the single server case by Ayesta [1], and has been extended to the multi-server case by El-Taha [7]. See also Baccelli and Brémaud [2, p. 163].

Now suppose we have a stochastic multi-server system and let W_q be a random variable that represents the waiting time in queue. Using the expectation notation $E[.]$, (15) can be written as

$$E[V] = \frac{\lambda E[S^2 + S]}{2} + \lambda \sum_{x=0}^{\infty} E[SW_q \mid S = x] f(x);$$

$$E[V] = \frac{\lambda E[S^2 + S]}{2} + \lambda E[SW_q]. \tag{16}$$

Moreover, if the waiting time in queue is independent of service times, i.e. $\mathcal{w}_q(x) = E[W_q]$ for all x values. Then using (15) again we obtain

$$E[V] = \frac{\lambda E[S^2 + S]}{2} + \lambda E[S] E[W_q]. \tag{17}$$

Note also that (17) can be obtained from (16) by the assuming the scheduling discipline to be service time independent; see Section 2 of this article and El-Taha and Stidham [8, pp. 175-177]. The next result shows that a scheduling discipline is non-anticipative iff the covariance of W_q and S is 0.

Corollary 4.3. *A non-preemptive scheduling discipline is non-anticipative, i.e.* $\Delta_x^f \mathcal{w}_q(x) = 0$ *for all* $x \in I$*, if and only if* $Cov(W_q, S) = 0$.

Proof. Use (12) and note that non-preemption implies that $\mathcal{w}_q(a,x) = \mathcal{w}_q(x)$ to obtain

$$E[V] = \frac{\lambda E[S^2 + S]}{2} + \lambda \sum_{x=0}^{\infty} \left[\mathcal{w}_q(x+1) + \sum_{a=0}^{x} \Delta_x^f \mathcal{w}_q(x) \right] F^c(x)$$

$$= \frac{\lambda E[S^2 + S]}{2} + \lambda \sum_{x=0}^{\infty} \left[\mathcal{w}_q(x+1) + (x+1) \Delta_x^f \mathcal{w}_q(x) \right] F^c(x). \tag{18}$$

Comparing (16) and (18), we obtain

$$E[SW_q] = \sum_{x=0}^{\infty} \mathcal{w}_q(x+1) F^c(x) + \sum_{x=0}^{\infty} (x+1) \Delta_x^f \mathcal{w}_q(x) F^c(x). \tag{19}$$

Now suppose the scheduling rule is non-anticipative then $\Delta_x^f \mathcal{w}_q(x) = 0$, is equivalent to

$\mathcal{w}_q(x)$ is independent of x, that is $\mathcal{w}_q(x) = E[W_q]$. Therefore

$$E[SW_q] = \sum_{x=0}^{\infty} \mathcal{w}_q(x+1)F^c(x) = E[W_q]E[S].$$

Thus $Cov(W_q, S) = 0$. Conversely, suppose $Cov(W_q, S) = 0$. Then $\mathcal{w}_q(x) = E(W_q \mid S = x) = E[W_q]$. Therefore

$$\sum_{x=0}^{\infty} \mathcal{w}_q(x+1)F^c(x) = \sum_{x=0}^{\infty} E[W_q]F^c(x) = E[W_q]E[S].$$

Using (19), we conclude that $\Delta_x^f \mathcal{w}_q(x) = 0$.

The conclusion of Corollary 4.3 can be strengthened when service times are geometric/modified geometric as indicated in the Corollary below.

Corollary 4.4. *Let* S *be a r.v. with a modified geometric distribution defined as* $P(S = x) = q^x p$ *for* $x = 0,1,\cdots; p > 0$*, and* $q = 1 - p$*; and* 0*, otherwise. Then*

(i) $Cov(W_q, S) = qp^{-1}\sum_{x=0}^{\infty}(x+1)\Delta_x^f \mathcal{w}_q(x) f(x)$; *and*

(ii) $Cov(W_q, S) = 0 (> 0)(< 0)$ *if and only if for all* x, $\Delta_x^f \mathcal{w}_q(x) = 0 (> 0)(< 0)$ *respectively.*

Proof. It follows from the memoryless property of service times that $F^c(x) = E[S]f(x)$, where $0 < E[S] = q/p < \infty$. Therefore, using (19), we obtain.

$$E[SW_q] = E[S]E[W_q] + \sum_{x=0}^{\infty}(x+1)\Delta_x^f \mathcal{w}_q(x)F^c(x).$$

Thus (i) follows. Part (ii) follows from (i).

4.4. Anticipative Vs non-anticipative disciplines

In this subsection we show the effect on waiting time of selecting an anticipative service time discipline versus non-anticipative one. The results are discrete time versions of the continuous time results obtained by El-Taha [7]. We assume a discrete multi-server system for which the invariance property holds. The first result compares an anticipative service time discipline versus non-anticipative one, with possibly preemptive rules.

Corollary 4.5. *Consider a work-conserving stable queueing system that satisfies the invariance property, and let* φ_1 *denote a non-anticipative discipline, and* φ_2 *denote an anticipative discipline such that for all* $0 \le a < x$,

$$\Delta_x^f \mathcal{w}^{\varphi_2}(a, x) \ge 0.$$

Then

$$\sum_{x=0}^{\infty} \mathcal{w}^{\varphi_2}(x)F^c(x-1) \le \sum_{x=0}^{\infty} \mathcal{w}^{\varphi_1}(x)F^c(x-1). \tag{20}$$

Moreover, if service times have a geometric distribution defined as $P(S=x)=q^{x-1}p$, $x=1,2,\cdots;p>0,p+q=1$, *then*

$$E[W^{\phi_2}]\le E[W^{\phi_1}]. \tag{21}$$

The inequalities in (20) and (21) are reversed when $\Delta_x^f \boldsymbol{w}^{\varphi_2}(a,x)\le 0$.

Proof. Using Theorem 3.2 and Theorem 4.2, we have

$$\begin{aligned} E[V^{\varphi_2}] &= \lambda\sum_{x=0}^{\infty}\left[\boldsymbol{w}^{\varphi_2}(x+1)+\sum_{a=0}^{x}\Delta_x^f \boldsymbol{w}^{\varphi_2}(a,x)\right]F^c(x) \\ &= \lambda\sum_{x=0}^{\infty}\boldsymbol{w}^{\varphi_1}(x)F^c(x-1) \\ &= E[V^{\phi_1}]. \end{aligned}$$

Therefore, $\Delta_x^f \boldsymbol{w}^{\varphi_2}(a,x)\ge 0$ implies that $\sum_{x=0}^{\infty}\boldsymbol{w}^{\varphi_2}(x+1)F^c(x)\le\sum_{x=0}^{\infty}\boldsymbol{w}^{\varphi_1}(x)F^c(x-1)$. A change of variable proves the first part. When service times have a geometric distribution we have $F^c(x)=E[S]f(x+1)$. Substitute in (20) to obtain the second part of the result.

It is interesting that in Corollary 4.4 either the geometric or the modified geometric distribution would work while in Corollary 4.5 only the geometric distribution is needed for (21) to hold. Note also that Corollary 4.5 allows a scheduling rule to be preemptive. For non-preemptive disciplines it is easy to see that $\boldsymbol{w}(x)=\boldsymbol{w}_q(x)+x$. Using this property we have the following result.

Corollary 4.6. *Consider a work-conserving non-preemptive stable queueing system that satisfies the invariance property, and let* ϕ_1 *denote a non-anticipative discipline, and* ϕ_2 *denote an anticipative discipline such* $Cov(W_q^{\varphi_2},S)\ge 0$. *Then* $E[W_q^{\phi_2}]\le E[W_q^{\phi_1}]$. *The inequality is reversed if* $Cov(W_q^{\phi_2},S)\le 0$.

Proof. Using (16) and (17) we obtain

$$E[V^{\varphi_2}]=\frac{\lambda E[S^2+S]}{2}+\lambda E[SW_q^{\varphi_2}];$$

$$E[V^{\phi_1}]=\frac{\lambda E[S^2+S]}{2}+\lambda E[S]E[W_q^{\phi_1}].$$

By the invariance property $E[V^{\phi_2}]=E[V^{\phi_1}]$, so that $E[SW_q^{\phi_2}]=E[S]E[W_q^{\phi_1}]$. Thus

$$Cov(W_q^{\phi_2},S)+E[S]E[W_q^{\phi_2}]=E[S]E[W_q^{\phi_1}].$$

By the condition that $Cov(W_q^{\phi_2},S)\ge 0$, we conclude that $E[W_q^{\phi_2}]\le E[W_q^{\phi_1}]$.

Now, we consider a $B/G/1$ discrete-time single server model with Bernoulli arrivals, general service times, and a server that works at a unit rate. The workload in this model is invariant for all work-conserving scheduling rules. By Corollary 4.5, $\Delta_x^f \mathcal{W}^{\varphi_2}(a,x) \geq 0$ is a sufficient condition for an anticipating discipline to have a smaller waiting time than non-anticipating discipline. For the $B/G/1$ non-preemptive anticipative SPT discipline, one can adapt the continuous-time priority queue discipline material in Gross et. al. [12], pp.150-155), to show that for our discrete-time model

$$\mathcal{W}(a,x) = a + \frac{\lambda E[S^2+S]/2}{[1-\sigma(x)]^2}; \tag{22}$$

where $\sigma(x) = \lambda \sum_0^x yf(y) < 1$ for all $0 < x < \infty$. Because $\sigma(x) < \sigma(x+1)$ for all $x \geq 0$, One can see from (22) that for the SPT rule, $\Delta_x^f \mathcal{W}^{\varphi_2}(a,x) > 0$, and thus it has lower waiting time than any non-anticipating discipline.

4.5. No-wait multi-server systems

For multi-server loss and infinite server systems, $\mathcal{W}(a,x) = a$ and $\Delta_x^f \mathcal{W}(a,x) = 0$. Now for the infinite server system Theorem 3.2 leads to

$$E[V] = \lambda \sum_{x=0}^{\infty} (x+1) F^c(x) = \lambda E[S^2+S]/2 . \tag{23}$$

For loss systems with c servers the workload is obtained from (23), by noting that the effective arrival rate is give by $\lambda(1-p(c))$ where $p(c)$ (e.g., Corollary 7.8 of El-Taha and Stidham [8]) is the blocking probability. Thus

$$E[V] = \lambda(1-p(c))E[S^2+S]/2 .$$

4.6. Fixed priority multi-class systems

We give an application that involve deriving a conservation law for a fixed priority multi-class multi-server system. This result characterizes a conservation law for non-preemptive multi-server systems. For this system the limiting averages are assumed to be well-defined. For $j = 1,\cdots,J$, let $E[V_j]$ be class j workload in the system, $E[W_j]$ be class j long-run average waiting time in the system, $\gamma_j = E[S_j^2 + S_j]/2E[S_j]$, $E[S_j] = 1/\mu_j$, and $\rho_j = \lambda_j E[S_j]/c$. Note that $E[W_j] = E[W_{qj}] + E[S_j]$, where $E[W_{qj}]$ is class j long-run average waiting time in the queue.

Theorem 4.7. *Consider a work-conserving $G/G/c$ multi-server system that satisfies the invariance property with a set $\boldsymbol{J}$ containing J classes. Let Φ_J denote the set of all scheduling rules , $\varphi \in \Phi$ which are non-preemptive, within each class $j \in \boldsymbol{J}$. Then, under any $\varphi \in \Phi_J$ the workload $E[V]$ and the vector of mean waiting times in the system $(E[W_1],\ldots,E[W_J])$ satisfy,*

$$E[V]=\sum_{j\in \boldsymbol{J}}\lambda_j\sum_{x=0}^{\infty}\left[\mathcal{w}_j(x+1)+\sum_{a=0}^{x}\Delta_x^f\mathcal{w}_j(a,x)\right]F_j^c(x). \qquad (24)$$

Moreover, if in addition, the scheduling rules are non-preemptive and non-anticipative, then

$$\sum_{j\in \boldsymbol{J}}\rho_jE[W_j]=\frac{E[V]}{c}-\sum_{j\in \boldsymbol{J}}\rho_j\gamma_j+\sum_{j\in \boldsymbol{J}}\frac{\rho_j}{\mu_j}. \qquad (25)$$

Proof. Theorem 3.2 applies for each class j. Now (24) follows by noting that $E[V]=\sum_{j\in \boldsymbol{J}}E[V_j]$. Using equation (12), and noting that $\Delta_x^f\mathcal{w}_j(a,x)=0$ and $\mathcal{w}_q(x)=E[W_q]$ for non-anticipative and non-preemptive disciplines respectively, we obtain

$$\begin{aligned}E[V]&=\sum_{j\in \boldsymbol{J}}E[V_j]\\&=\sum_{j\in \boldsymbol{J}}\lambda_j\left[\sum_{x=0}^{\infty}E[W_{qj}]F_j^c(x)+E[S_j^2+S_j]/2\right]\\&=\sum_{j\in \boldsymbol{J}}\lambda_j\left[E[W_{qj}]E[S_j]+E[S_j^2+S_j]/2\right]\\&=c\sum_{j\in \boldsymbol{J}}[\rho_jE[W_{qj}]+c\rho_j\gamma_j]. \qquad (26)\end{aligned}$$

Now using $\rho_jE[W_{qj}]=\rho_jE[W_j]-\rho_j/\mu_j$, $j\in \boldsymbol{J}$; we obtain the desired result.

Theorem 4.9 is similar to the continuous version except that now $\gamma_j=E[S_j^2+S_j]/2E[S_j]$, instead of $\gamma_j=E[S_j^2]/2E[S_j]$ for the continuous case. Note also that equations (24) and (25) are valid for a wide variety of systems. However, at this level of generality $E[V]$ will depend on the scheduling rule. For $E[V]$ to be invariant with respect to scheduling rules we need the scheduling rules to be non-anticipative and the additional condition that in multi-server systems the service times are *i.i.d.* for all classes; see El-Taha [6] for details. Moreover, for (25) to be computationally useful we need to be able to compute $E[V]$. This can be done in the single server Bernoulli arrivals *FCFS* case. By *FCFS*, and *BASTA* (Bernoulli Arrivals See Time Averages),(e.g., Chapter 3 of El-Taha and Stidham [8]), it follows that $E[V]=E[W_q]$, so that using (5) leads to

$$E[W_q]=\rho E[W_q]+\lambda E[S^2+S]/2.$$

Simplify to obtain

$$E[V]=E[W_q]=\frac{\lambda E[S^2+S]}{2(1-\rho)}. \qquad (27)$$

Even though *FCFS* is used, this formula for $E[V]$ is valid for all invariant scheduling rules.

4.7. The $c\mu$-rule for discrete time queues

Consider a multi-class single-server discrete time system with Bernoulli arrivals and general *i.i.d*. service times, and let $\boldsymbol{J}=\{1,\ldots,J\}$ be the set of all customer classes. Let c_j be the class $j; j=1,2,\cdots J$ holding cost rate in queue, i.e., the cost per customer per unit time in queue. In this subsection we show how conservation and strong conservation laws can be used to give a proof to the $c\mu$-rule for discrete time queues. The $c\mu$-rule states that it is optimal to serve classes in the order of the largest $c_j\mu_j$ values, the remaining classes are assigned similarly. The objective is to find the optimal policy that minimizes $\sum_j^J c_j[L_{qj}]$, where $[L_{qj}]$ is the mean number of the j-class customers in the queue. By Little's law this is equivalent to

$$\sum_j^J c_j[L_{qj}]=\sum_j^J c_j\lambda[W_{qj}]=\sum_j^J (c_j\mu_j)\rho_j[W_{qj}]=\sum_j^J (c_j\mu_j)[U_j].$$

It follows from Subsections 4.1 and 4.2 that

$$E[U]=\rho E[W_q]; \tag{28}$$

$$E[V]=\rho E[W_q]+\lambda E[S^2+S]/2. \tag{29}$$

Also (28) and (29) are valid for each class $j\in\boldsymbol{J}$, that is

$$E[U_j]=\rho_j E[W_{qj}]; \tag{30}$$

$$E[V_j]=\rho_j E[W_{qj}]+\lambda_j E[S_j^2+S_j]/2. \tag{31}$$

We shall need the following preliminary result.

Lemma 4.8. *For any multi-class system*

$$\sum_j^J \rho_j\gamma_j=\sum_j^J \lambda_j E[S_j^2+S_j]/2=\lambda E[S^2+S]/2.$$

Proof. Sum both sides of (31) to get

$$\begin{aligned}\sum_{j=1}^J \lambda_j E[S_j^2+S_j]/2 &= \sum_{j=1}^J E[V_j]-\sum_{j=1}^J \rho_j E[W_{qj}]\\ &= E[V]-E[U]\\ &= E[V]-\rho E[W_q]\\ &= \lambda E[S^2+S]/2;\end{aligned}$$

where we used (31),(30), (28), and (29) in steps 1, 2, 3, and 4 respectively.

We are interested in *strict priority rules* that give strict preference for certain classes over others. In a multi-class queueing system, let $S\subset\boldsymbol{J}$ be a subset of classes and consider strict priority rules, i.e. rules that give strict priority to jobs in S over jobs in S^c. For each permutation, ψ of $\boldsymbol{J}$ there is a strict priority rule, which at each point serves a customer with the smallest index ψ_j among all customers that are currently present in the system. In other words, the class j with $\psi_j = 1$ has priority over all other classes, followed by the class k with $\psi_k = 2$, and so on.

Consider a scheduling rule that gives non-preemptive priority to jobs in S over jobs not in S, and service time independent within each class $j \in \boldsymbol{J}$. We refer to such *rules* as S - *rules* and jobs in S as S - *jobs*. Under S - *rules* we shall use the notation $E_S[V_S]$, and $E_S[U_S]$. Within the set of S - *rules* it can been shown (Gelenbe and Mitrani [10]) that $E_S[V_S]$ and $E_S[U_S]$ are invariant. This is a key step in establishing that a system obeys strong conservation laws. Also, let $\rho_s = \sum_{j\in S}\rho_j$.

Theorem 4.9. *Consider a* $B/G/1$ *queueing system with* J *classes. Let* Φ_J *denote the set of all scheduling rules,* $\varphi \in \Phi$ *which are non-preemptive. Service times of customers within each class are* i.i.d.. *Then, under any* $\varphi \in \Phi_J$ *the vector of mean waiting times in the queue* $(EW_{q1},\dots,E[W_{qj}])$ *satisfies the linear system,*

$$\sum_{j\in \boldsymbol{J}}\rho_j E[W_{qj}] = \frac{\rho}{1-\rho}\sum_{j=1}^{J}\rho_j\gamma_j\,; \tag{32}$$

$$\sum_{j\in S}\rho_j E[W_{qj}] \ge \frac{\rho_S}{1-\rho_S}\sum_{j=1}^{J}\rho_j\gamma_j\,; S\subseteq \boldsymbol{J}\,. \tag{33}$$

Moreover, the constraint (33) is satisfied with equality for any S *-rule, that is, for any rule* $\phi \in \Phi_J$ *that gives priority to jobs in classes* $j \in S$ *over jobs in classes* $j \in S^c$. *In other words, strong conservation laws hold for any* $\phi \in \Phi_J$.

Proof. Note that

$$E[W_j] = E[W_{qj}] - 1/\mu_j\,,$$

and substitute in (25) to get

$$\sum_{j=1}^{J}[U_j] = \sum_{j=1}^{J}\rho_j E[W_{qj}] = E[V] - \sum_{j=1}^{J}\rho_j\gamma_j\,.$$

Use (27), Lemma 4.8, and simplify to get (32). To prove (33), note that

$$\sum_{j\in S}E[V_j] = \sum_{j\in S}E[U_j] + \sum_{j\in S}\rho_j\gamma_j\,;$$

$$E_S[V_S] = E_S[U_S] + \sum_{j\in S}\rho_j\gamma_j\,;$$

$$E_S[U_S] = \rho_S E_S[W_{qS}]\,. \tag{34}$$

Now because $E_S[U_S]$ is invariant and

$$\sum_{j\in S} E[U_j] \geq E_S[U_S],$$

we obtain, using (30) and (34),

$$\sum_{j\in S} \rho_j E[W_{qj}] \geq \rho_S E_S[W_{qS}].$$

But

$$E_S[W_{qS}] = \frac{\lambda E[S^2+S]}{2(1-\rho_S)} = \frac{1}{1-\rho_S}\sum_{j\in J}\rho_j\gamma_j ,$$

so that

$$\sum_{j\in S}\rho_j E[W_{qj}] \geq \frac{\rho_S}{1-\rho_S}\sum_{j\in J}\rho_j\gamma_j .$$

This completes the proof.

Therefore, we have the following optimization problem:

$\min \sum_j^J (c_j\mu_j)E[U_j]$, subject to the constraints

$$\sum_{j\in J} E[U_j] = \frac{\rho}{1-\rho}\sum_{j=1}^{J}\rho_j\gamma_j ;$$

$$\sum_{j\in S} E[U_j] \geq \frac{\rho_S}{1-\rho_S}\sum_{j=1}^{J}\rho_j\gamma_j ; S \subseteq \boldsymbol{J} .$$

Now it follows from Theorem 2 of Green and Stidham [11] that for this optimization problem the optimal policy is the $c\mu$-rule. This is the strict-priority rule $\pi(\psi)$, where ψ is a permutation of the class indices such that $c_{\psi_1}\mu_{\psi_1} \geq c_{\psi_2}\mu_{\psi_2} \geq \cdots \geq c_{\psi_J}\mu_{\psi_J}$, is optimal over all policies $\pi \in \Pi$. That is, for the $B/G/1$ system with J classes, order the classes such $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_J\mu_J$. The the policy that serves classes in this order is optimal in the sense that it minimizes $\sum_j^J c_j[L_{qj}]$. Note that when service times are *i.i.d.* for all classes, the the optimal policy will serve the class with highest cost rate, then second highest holding cost rate, and so on.

5. Concluding Remarks

In this article, we give a discrete-time conservation law that is valid for discrete-time multi-server systems under scheduling rules that are non-anticipative as well as anticipative. It also extends a non-anticipative law for multi-server systems to allow for anticipative scheduling rules. Several applications that illustrate the usefulness of the law have been given. In particular, new results that characterize non-preemptive anticipative scheduling rules, as well as results that compare anticipative vs non-anticipative disciplines are provided. Discrete-time systems present a number of subtleties that do not come up in continuous-time counterparts. Our results complement those given for continuous-time systems.

6. Appendix

6.1. Definitions and notation

Here we give a list of definitions and the notation used in the article for easy reference.

- λ : arrival rate
- ES : long-run average service time
- ES^2 : long-run average second moment of service times
- EW : long-run average waiting time (including service time) in a queueing system
- EW_q : long-run average waiting time (excluding service time) in a queueing system
- EV : long-run average unfinished workload (including work in service) in a queueing system
- EU : long-run average unfinished workload in queue(excluding work in service) in a queueing system
- $\mathcal{W}(x)$: long-run average conditional waiting time (including time in service)for customers with x units of service.
- $\mathcal{W}_q(x)$: long-run average conditional waiting time in the queue for customers with x units of service.
- $\mathcal{W}_k(a,x)$: the k^{th} customer waiting time (time in the system) until it receives a units of service given that its service time is x units, $a \le x$.
- $\mathcal{W}(a,x)$: long-run average conditional sojourn time (time in the system) of customers with a units of attained service among all customers with x units of service requirements, $a \le x$.
- $\mathcal{W}_q(a,x)$: long-run average conditional waiting time in the queue of customers with a units of attained service among all customers with x units of service requirements, $a \le x$.
- $\mathcal{A}_k(\tau, S_k)$: the k^{th} customer attained service time given that the customer has been waiting for τ units and has S_k units of service.
- $f(x)$ = long run fraction of customers with x units of service , i.e. the probability mass function of service times.
- $F(x)$: cumulative distribution function of service times
- $F^c(x) = 1 - F(x)$
- ϕ^1 represents a non-anticipative discipline

- ϕ^2 represents an anticipative discipline
- EV_S: long-run average unfinished workload (including work in service) for classes in S in a queueing system
- EU_S: long-run average unfinished workload in queue(excluding work in service) for classes in S in a queueing system
- E_SV_S: long-run average unfinished workload (including work in service) for classes in S under strict priority rules
- E_SU_S: long-run average unfinished workload in queue (excluding work in service) for classes in S under strict priority rules

6.2. Discrete integration by parts

The result of this subsection is not new, but the style and notation make the result easily accessible to the readers of this article. Let $f(x)$ and $g(x)$ be discrete functions, $x = a, a+1, \cdots b,$; and 0 otherwise. Recall that $\Delta^f f(x) = f(x+1) - f(x)$ denotes the forward difference of $f(x)$. Our results focus on forward differences, but one can establish similar result for backward differences as well. The first result is a discrete integration by parts formula.

Lemma 6.1. *Let $f(x)$ and $g(x)$ be discrete non-negative functions. Then*

$$\sum_{x=a}^{b} f(x)\Delta^f g(x) = f(x)g(x)\Big|_{x=a}^{b+1} - \sum_{x=a}^{b} g(x+1)\Delta^f f(x). \tag{35}$$

Proof. Note that

$$\begin{aligned}\Delta^f(f(x)g(x)) &= f(x+1)g(x+1) - f(x)g(x)\\ &= f(x+1)g(x+1) - f(x)g(x+1) + f(x)g(x+1) - f(x)g(x)\\ &= g(x+1)\Delta^f f(x) + f(x)\Delta^f g(x).\end{aligned}$$

Therefore

$$f(x)\Delta^f g(x) = \Delta^f(f(x)g(x)) - g(x+1)\Delta^f f(x).$$

Now

$$\begin{aligned}\sum_{x=a}^{b}\Delta^f(f(x)g(x)) &= [f(a+1)g(a+1) - f(a)g(a)] + [f(a+2)g(a+2) - f(a+1)g(a+1)]\\ &\quad + \cdots + [f(b+1)g(b+1) - f(b)g(b)]\\ &= f(b+1)g(b+1) - f(a)g(a)\\ &= f(x)g(x)\Big|_{x=a}^{b+1}.\end{aligned}$$

Thus

$$\sum_{x=a}^{b} f(x)\Delta^f g(x) = f(x)g(x)\Big|_{x=a}^{b+1} - \sum_{x=a}^{b} g(x+1)\Delta^f f(x).$$

This completes the proof.

Using backward differences and an argument similar to Lemma 6.1, we obtain similar results that lead to an alternative backward difference proof of Theorem 3.2.

6.3. Busy period bound

This result is a bound on the busy period for a discrete-time single server queue. We allow for the possibility of multiple arrivals at arrival instants, i.e. batch arrivals. let T_k, X_k refer to batch arrival instants and batch size (number of arrivals in a batch). Also let $A_B(n), A_B(0,n)$ and $Y(0,n)$ be the number of batch arrivals , number of batch arrivals that see the system in state 0, and time in state 0 during time $[0,n]$ respectively; where state 0 indicates 0 customers in the system. Now, define the following limits when they exist:

$$EX_B = \lim_{m\to\infty} m^{-1}\sum_{k=1}^{m} X_k;$$

$$\lambda_B = \lim_{n\to\infty} A_B(n)/n;$$

$$\lambda_B(0) = \lim_{n\to\infty} A_B(0,n)/Y(0,n);$$

$$p(0) = \lim_{n\to\infty} Y(0,n)/n\,.$$

We interpret $EX_B, \lambda_B, \lambda_B(0$, and $p(0)$ as the long-run average batch size, batch arrival rate, batch arrival rate in state 0, and fraction of time in state 0 respectively. Note that the arrival rate of all customers $\lambda = \lambda_B EX_B$ and $p(0) = 1 - \lambda/\mu > 0$ ($\mu = 1/ES$). Similar to El-Taha and Stidham [8, p.26]

$$EI = \lim_{n\to\infty} Y(0,n)/A_B(0,n);$$

$$EB = \lim_{n\to\infty} (n - Y(0,n))/A_B(0,n);$$

where EI, and EB are the long-run average idle period, and busy period respectively. Moreover

$$EI = 1/\lambda_B(0);$$

$$EB = (1 - p(0))/\lambda_B(0)p(0) = \frac{\lambda ES}{\lambda_B(0)(1-\lambda ES)}.$$

Let $\{A_k = T_k - T_{k-1}, k \ge 1\}$ be the sequence of inter-arrival times, and define $A'_k = \min\{A_k, M\}$ where M is chosen such that $\lambda' ES < 1$; i.e., $\lambda < \lambda' < \mu$, . Here λ' is the arrival rate of the new modified system.

Lemma 6.2. *Let $\rho = \lambda ES < 1$, then for a discrete-time $GI/GI/1$ model with possibly batch arrivals where relevant limits are well-defined, we have*

$$EB \le \rho M / (\frac{\lambda}{\lambda'} - \rho).$$

Proof. We follow the the notation used in El-Taha and Stidham [8], pp. 25-26. Let $A_B(0,n)$, and $Y(0,n)$ ($A'_B(0,n)$, and $Y'(0,n)$) be the original (modified) system batch arrivals that find the system in state 0, and total time in state 0 during $[0,n)$ respectively. Note that $A'_B(0,n) \le A_B(0,n)$ and $Y'(0,n) \le Y(0,n)$. This follows by noting the effect of reducing inter-arrival times on, possibly, joining neighboring busy periods.

Now, the long-run average idle period for the modified system $I' = \lambda'_B(0)^{-1} \le M$. Moreover,

$$EB := \lim_{n\to\infty} \frac{n - Y(0,n)}{A_B(0,n)} \le \lim_{n\to\infty} \frac{n - Y'(0,n)}{A'_B(0,n)} := EB'.$$

Therefore

$$\begin{aligned} EB \le EB' &= \frac{\lambda' ES}{\lambda'_B(0)(1 - \lambda' ES)} \\ &\le \frac{\lambda' ESM}{(1 - \lambda' ES)} \\ &= \rho M / (\frac{\lambda}{\lambda'} - \rho) \end{aligned}$$

which completes the proof.

No stochastic assumptions are used in obtaining this bound. We only assumed that relevant limits are well defined. In the same spirit, we prove the second result that is needed in the proof of Theorem 3.2.

6.4. Additional results

The following results are needed in some proofs.

Lemma 6.3. *Suppose that* $E[S^2] = \sum_{x=0}^{\infty} x^2 f(x) < \infty$*, then*

$$\lim_{x\to\infty} x^2 F^c(x) = 0.$$

Proof. Note that

$$E[S^2] = \sum_{y=0}^{x-1} y^2 f(y) + \sum_{y=x}^{\infty} y^2 f(y).$$

Now, taking limits of both sides as $x \to \infty$ leads to

$$E[S^2] = \sum_{y=0}^{\infty} y^2 f(y) + \lim_{x\to\infty} \sum_{y=x}^{\infty} y^2 f(y).$$

Therefore

$$\lim_{x\to\infty}\sum_{y=x}^{\infty} y^2 f(y) = 0. \tag{36}$$

Now

$$\begin{aligned}\lim_{x\to\infty} x^2 F^c(x) &= \lim_{x\to\infty}\sum_{y=x}^{\infty} x^2 f(y)\\ &\le \lim_{x\to\infty}\sum_{y=x}^{\infty} y^2 f(y)\\ &= 0.\end{aligned}$$

Lemma 6.4. *Let $f(x)$ be any be discrete non-negative probability mass function and $h(x)$ be any discrete non-negative function. Then*

$$\sum_{x=0}^{\infty} h(x) f(x) = \sum_{x=0}^{\infty} \Delta^f h(x) F^c(x) + h(0).$$

Proof. Note that $h(x) = \sum_{y=0}^{x-1} \Delta^f h(y) + h(0)$. Then

$$\begin{aligned}\sum_{x=0}^{\infty} h(x) f(x) &= \sum_{x=0}^{\infty} (h(0) + \sum_{y=0}^{x-1} \Delta^f h(y)) f(x)\\ &= \sum_{x=0}^{\infty} h(0) f(x) + \sum_{y=0}^{\infty}\sum_{x=y+1}^{\infty} \Delta^f h(y)\\ &= \sum_{y=0}^{\infty} \Delta^f h(y) F^c(y) + h(0).\end{aligned}$$

We need the following Corollary in the proof of Corollary 4.1.

Corollary 6.5. *Let $h(x) = x^2$, then*

$$E(X^2) \equiv \sum_{x=0}^{\infty} x^2 f(x) = \sum_{x=0}^{\infty} (2x+1) F^c(x).$$

Acknowledgment

The author would like to thank anonymous reviewers for suggestions that helped improve this paper.

References

[1] Ayesta, U. (2007). A unifying conservation law for single-server queues. *Journal of Applied Probability*, 44, 1078–1087.

[2] Baccelli, F., & Brémaud, P. (1994). *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences, Applications of Mathematics, 26*. Springer-Verlag, New York.

[3] Bartsch, B., & Bolch, G. (1978). Conservation law for G/G/m queueing systems. *Acta Informatica*, 810, 105–109.

[4] Dacre, M., Glazebrook, K., & Ni$\tilde{n}$ o-Mora, J. (1999). The achievable region approach to the optimal control of stochastic systems. *Journal of Royal Statistical Society B*, 61, 747–791.

[5] El-Taha, M. (2016). Asymptotic time averages and frequency distributions. *International Journal of Stochastic Analysis*, Volume 2016, Article ID 2741214, 10 pages, *http://dx.doi.org/10.1155/2016/2741214*.

[6] El-Taha, M. (2016). Invariance of workload in queueing systems. *Queueing Systems*, 83, 181–192.

[7] El-Taha, M. (2017). A general workload conservation law with applications to queueing systems. *Queueing Systems*, 85, 361–381.

[8] El-Taha, M., & Stidham, Jr. S. (1999). *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishing, Boston, 1999.

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

[10] Gelenbe, E., & Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems*. Academic Press, London.

[11] Green, T., & Stidham, Jr. S. (2000). Sample-path conservation laws, with applications to scheduling queues and fluid systems. *Queueing Systems*, 36, 175–199.

[12] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. (2008). *Fundamentals of Queueing Theory*. John Wiley, New Jersey, 4th edition, 2008.

[13] Heyman, D. P., & Sobel, M. (1982). *Stochastic Models in Operations Research, Volume I.* McGraw-Hill, New York.

[14] Heyman, D. P., & Stidham, Jr. S. (1980). The relation between customer and time averages in queues. *Operations Research*, 28, 983–994.

[15] Kleinrock, L. (1976). *Queueing Systems, Volume II.* Wiley Intersciences, New York.

[16] Kleinrock, L., Muntz, R. R., & Hsu, J. (1971). Tight bounds on average response time for time-shared computer systems. In *Proceedings of the IFIP Congress*, Volume 1, 124–133.

[17] O'Donovan, T. M. (1974). Distribution of attained service and residual service in general queueing systems. *Operations Research*, 22, 570–574.

[18] Shanthikumar, J. G., & Yao, D. D. (1992). Multiclass queueing systems: Polynomial structure and optimal scheduling control. *Operations Research*, 40, S293–S299. Supplement 2.

[19] Stidham, Jr. S. (1990). On the relation between time averages and customer averages in queues, in. *Variational Methods and Stochastic Analysis eds. H. J. Kimn and D. M. Chung, Proceedings of Workshop in Pure Mathematics*, 9, 243–278.

[20] Wolff, R. (1989). *Stochastic Modeling and the Theory of Queues.* Prentice Hall, New Jersey.