Noname manuscript No.
(will be inserted by the editor)

On duality relations in finite queueing models

Nasser Barjesteh · Ravi R. Mazumdar · Catherine P. Rosenberg

Received: date / Accepted: date

Abstract Motivated by the starvation and overflow of queues, we study duality relations in finite queues when the arrival and service processes are interchanged. In particular we study the relations between arrival and departure Palm distributions and their relations to stationary distributions. We consider both the case of point process inputs as well as fluid inputs. We obtain inequalities between the probability of queue being empty and the probability of queue being full for both the time stationary and Palm distributions by interchanging arrival and service processes. In the fluid queue case we show that there is an equality between arrival and departure distributions. The techniques are based on monotonicity arguments and coupling. The usefulness of the bounds is illustrated via numerical results.

Keywords Queues · Duality · Point processes · Palm distributions · Stationary distributions · Fluid

1 Introduction

Explicit results for stationary distributions in finite capacity queueing models with general arrival and service processes are rare, and yet, as applications evolve there is the need to move beyond classical Markovian queueing models. However, in many applications, we are interested in specific measures such as the overflow probability in finite queues for which there exist many results under fairly general hypotheses on the input processes via the use of large deviations or heavy traffic limits as the loading on the queues increases. Recently, motivated by applications in energy systems and multimedia streaming,

N. Barjesteh · R.R. Mazumdar · C.P. Rosenberg
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1
E-mail: nbarjest@uwaterloo.ca,
E-mail: mazum@ece.uwaterloo.ca,
E-mail: cath@uwaterloo.ca

the probability of starvation of finite buffer queues has become the focus of attention. The probability of starvation corresponds to the probability that a queue becomes empty, i.e., the end of a busy period. For instance, [1] relates the starvation of storage units in energy systems to the starvation of finite buffer queues. Similarly, in [19], authors are interested in maximizing the quality of experience (QoE) of media streaming service by optimizing the number of *prefetched* packets in video streaming in order to avoid having periods of buffer starvation with no packets to playback. A key measure is the distribution of the number of buffer starvations within a sequence of N consecutive packet arrivals and they propose a trade-off between the start-up delay and the starvation.

Starvation in queues is defined in many ways and different applications use different notions. Furthermore, due to the nature of applications, different assumptions on the nature of the queueing models are more appropriate. In classical queueing models where the arrivals are point processes, the probability of a departure leaving no customers in the queue and the stationary probability of queue being empty provide useful information about the starvation. Similarly, in a fluid flow model, underflow rate is of interest in most applications. Unfortunately, explicit results for these quantities are very difficult to obtain. Moreover, although one can relate the probability of a queue being empty to the probability of loss in the point process model using

$$\rho(1 - P_L) = 1 - \pi(0)$$

where ρ is the traffic intensity, P_L is the probability of loss, and $\pi(0)$ is the stationary probability of the queue being empty, this equality is of no practical use. This is because,in the heavy traffic case, $\rho(1 - P_L) \approx 1$ and the bounds obtained for $\pi(0)$ are not very useful. However, motivated by the $M/M/1/K$ queueing model, it is of interest to relate the starvation probability measures to the overflow probabilities in a queue with the arrival and service processes switched since well known and powerful methods that are valid for general stationary inputs can be exploited especially if we require the starvation probability to be small. The relationship between these two measures for queues with arrival and service distributions interchanged are what we refer to as duality in queueing models. Duality results concerning the stationary queue length distribution seen at arbitrary times and departure instances, stationary distribution of workload process, and underflow and overflow rates can be very useful in studying the starvation of finite buffer queues by relating the quantities of interest to other known and more tractable quantities. Duality has been studied in other contexts as for example in risk models [17, 2] where the dual or risk process corresponds to processes with negative jumps and is a dual process of the workload process with point process inputs and one of the measures of interest is the hitting probability to the origin corresponding to bankruptcy.

Let us begin by describing the problem and defining quantities of interest in our analysis. In this paper, we consider classical queueing models with both

discrete point process arrivals and fluid queues with continuous arrivals of work. In this paper the primal queue is assumed to be a $G_1/G_2/1/K$ and will be called $S1$ and its dual is a system in which the inter-arrival times have the same distribution as the service times in the primal queue and vice versa. In other words, the dual queue is a $G_2/G_1/1/K$ queue and denoted by $S2$. Let $\{TA[n]\}$ denote the arrival times and $\{T_D[n]\}$ the departure times. Let us denote the queue length at time t by $Q(t)$, and the queue length at arrival and departure instances by $Q_A[n] = Q(T_A[n])$ and $Q_D[n] = Q_{D^+}[n] = Q(T_D[n])$ respectively. Let $\pi(.) = \mathbb{P}(Q = .)$ denote the stationary distribution. Let us denote the event of the arrival of a customer by $[\Delta A_0 = 1]$, the event of the departure of a customer by $[\Delta D_0 = 1]$, and the event of customer entering the queue by $[\Delta Q_0 = 1]$. Note that because an arrival might get rejected, $[\Delta Q_0 = 1] \subset [\Delta A_0 = 1]$. Then, $\pi_A(n) = \mathbb{P}_A\{Q_{0-} = n\} = \mathbb{P}(Q_{0-} = n|\Delta A_0 = 1)$, and $\pi_D(n) = \pi_{D^-}(n+1) = \mathbb{P}(Q_{0-} = n+1|\Delta D_0 = 1)$ denote the Palm probabilities associated with the arrivals and departures. In this paper, we will use $\pi_D(.)$ to denote $\pi_{D+}(.)$,the Palm distribution just after a departure. For single server queues with FIFO discipline and finite buffer size of K, the following relationship holds between the various distributions [3,12].

$$\pi_D(n) = \frac{\pi_A(n)}{1 - \pi_A(K)} \tag{1}$$

$$\lambda_n \pi(n) = \lambda_A \pi_A(n) \tag{2}$$

$$\mu_n \pi(n) = \lambda_D \pi_D(n-1) \tag{3}$$

$$\lambda_n \pi(n) = \mu_{n+1} \pi(n+1) \tag{4}$$

where λ_n and μ_n denote the conditional arrival intensity and conditional departure intensity in state n and λ_A and λ_D denote the mean arrival rate and mean departure rate. λ_n and μ_n are not known in general and difficult to calculate for general input and service distributions. Equation (1) follows from an argument similar to the one proposed in [12, p.40] for infinite buffer queues. Note equation (1) holds for all work-conservative disciplines.

For the fluid flow model, we will use a notation similar to [18] and denote the cumulative input fluid and available processing in interval $[0,t]$ respectively by $\{C(t) : t > 0\}$ and $\{S(t) : t > 0\}$. Let $W(t)$ denote workload at time t, K denote the buffer size of the queue, $U(t) \triangleq \int_0^t \mathbb{1}\{W(s) = K\} d(C(s) - S(s))$ denote the overflow process (the amount of flow that is lost up to time t), and $L(t) \triangleq \int_0^t \mathbb{1}\{W(s) = 0\} d(S(s) - C(s))$ denote the underflow process. Moreover, let us denote the overflow and underflow rates, respectively by $\Lambda = \lim_{t\to\infty} t^{-1}U(t)$ and $\nu = \lim_{t\to\infty} t^{-1}L(t)$. Furthermore, as in the point process model, we will call the primal queue $S1$ and the dual, which has a cumulative available processing distributed as the cumulative input fluid in the primal queue and vice versa, $S2$.

The concept of duality in queues was first proposed by N.U. Prabhu in [4] and later discussed in [5]. In [5] a loop cyclic queueing system is described in

which K customers circulate between two finite buffer queueing systems and customers that leave one queue enter the other. Since then this notion has been studied by many authors. These authors have primarily focused on investigating the relationship between the stationary queue length distribution of $G_1/G_2/1/K$ seen at an arbitrary time and its counterpart in $G_2/G_1/1/K$, and very few results are proposed for the relationship between the stationary queue length distribution seen by arrivals or departures in the above-mentioned queues. M. Hlynka in [6] showed that $\pi_1(i) = \pi_2(K-i)$, $\forall i$ holds only for a $M/M/1/K$ queueing system and the equality fails to hold for more general arrival processes and service times. He showed that there exists a quasi-dual queue for which the equality holds. The so-called quasi-dual queue has a modified first service time, meaning that the first service time of each busy period has a different distribution, and its arrivals stop when the queue is full. The results of [6] were used in [7] to obtain duality results for queue length distribution in queueing systems with arrival and service control and in [8] to compute the loss probability of an overloaded $GI/M/1/K$ queue.

In this paper, we show some new duality results for queue length distribution of queues with more general arrival processes and service times, seen at arbitrary times and at departure instances. We will also investigate duality results concerning finite buffer fluid queues. In the sequel the probability distribution seen at arbitrary times will simply be referred to as the (time) stationary distribution while the Palm probabilities associated with arrivals and departures will be called arrival and departure distributions. Our approach is via monotonicity arguments and coupling whereby we construct processes on a common probability space to compare them.

The paper is organized as follows. In Section 2 we begin by defining the interrupted arrival and virtual service disciplines and recall the results of [7] for duality in queues with such disciplines. We also provide some new results for queues with a virtual service discipline. Next, we compare the queue length distribution of a $GI/GI/1/K$ queue under FIFO and interrupted arrival discipline and use the results to obtain some new relationships (bounds) between the dual quantities in finite buffer FIFO queues. In Section 2.1, we consider the special case of $M/G/1/K$ queues and in Section 3, we investigate duality results concerning finite buffer fluid flow queues in the general case and the special case of fluid queues with ON-OFF fluid input and constant available processing. Section 4 presents some numerical results for finite buffer queues that confirm our results as well how good the bounds are. Section 5 provides a summary of the results.

2 Duality between overflow and starvation in classical queueing models

As stated in [6], an exact duality relation in the form $\pi_1(i) = \pi_2(K-i)$, $\forall i$ does not hold for queues with general arrivals or service times. The lack of

equality is because when a queue in empty, its server stays idle and waits for an arrival, but when the queue is full, the arrival process does not stop and only the packets that find the queue full at the time of their arrival are rejected. In other words, the time till next departure from the moment a packet arrives at an empty queue is distributed as a service time, but the time till next arrival from the moment a packet leaves the full queue has a residual inter-arrival time distribution. Thus, if one desires to find an equivalent dual queue i one must consider queues with controlled arrivals and services as suggested in [7]. We now define two concepts introduced in [7] that are used in this paper.

Definition 1

1. Virtual service discipline: The server does not become idle when the queue length is zero but starts a virtual service. If an arrival occurs during a virtual service, then service time of the first arrival is not a regular service time but the remaining time of the ongoing virtual service. If no arrivals occur, the server starts another virtual service. In other words, the first customer served in a busy period receives a special service while the rest of the customers receive regulars services.
2. Interrupted/Stopped arrival discipline: The arrival stream is turned off when the buffer is full and turned on when the buffer space becomes available. Therefore, no customer losses occur and the time till next arrival from the moment a customer leaves the full queue has an inter-arrival time distribution, as opposed to a queue with FIFO discipline, in which customers are rejected when queue is full and the time till next arrival from the moment a customer leaves the full queue has a residual inter-arrival time distribution.

In [7, Thm 1] a duality relation between the queue length distributions of $G_1/G_2/1/K$ and $G_2/G_1/1/K$ with interrupted arrivals is given and in [7, Lem 1] it is shown that a similar relation holds between the queue length of S1 seen at arrivals and the queue length of the dual at departure instances. In [7, Thm 2], it is shown that a similar duality relation holds for the stationary queue length seen at arbitrary times for queues with a virtual service discipline. The paper does not discuss the queue length seen at arrival or departure instances.

Here in Theorem 1 , we will recall the results in [7] and provide a duality relation between the queue length of S1 seen at departure instances and the queue length of its dual at departure times.

Theorem 1

1. *Let* $s1$ *denote a* $G_1/G_2/1/K$ *queue with an interrupted arrival discipline, and* $s2$ *a* $G_2/G_1/1/K$ *queue with the same discipline. Then, the corresponding stationary and Palm distributions at departure times distributions*

satisfy:

$$\pi_{s1}(i) = \pi_{s2}(K-i)\ ,\ \forall 0 \leq i \leq K \tag{5}$$

and

$$\pi_{D_{s1}}(i) = \pi_{A_{s2}}(K-1-i) = \pi_{D_{s2}}(K-1-i)\ ,\ \forall 0 \leq i \leq K-1 \tag{6}$$

2. *Similarly, let* $v1$ *denote a* $G_1/G_2/1/K$ *with a virtual service discipline and* $v2$ *a* $G_2/G_1/1/K$ *with the same discipline. Then,*

$$\pi_{v1}(i) = \pi_{v2}(K-i)\ \ ,\ \forall 0 \leq i \leq K \tag{7}$$

$$\pi_{D_{v1}}(i) = \pi_{D_{v2}}(K-1-i)\ \ ,\ \forall 0 \leq i \leq K-1 \tag{8}$$

Proof Equation (5) and the left hand side equality of (6) are essentially the results presented in *Theorem 1* and *Remark 1* of [7] and do not need proof. Since queues with interrupted arrival discipline do not reject arrivals, similar to infinite buffer queues, $\pi_{A_s}(m) = \pi_{D_s}(m)$ for all m. Hence, we obtain the right hand side equality of equation (6).
Although equation (7) has been shown in [7], we provide a similar proof based on coupling alongside equation (8), which is new and used subsequently. Imagine *system* $v1$ is in tandem with some other queueing system with virtual service discipline that we will call *system* $2'$. Moreover, the sum of the number of customers in the two tandem queues is K. Therefore, customers that arrive at *system* $v1$ and are accepted correspond to the departures of *system* $2'$. Moreover, we enforce that the customers that are rejected in *system* $v1$ correspond to virtual departures (the instances in which virtual service times finish) of *system* $2'$ and vice versa. Hence, inter-arrival times of *system* $v1$ are distributed as service times of *system* $2'$ and vice versa. Then, since $Q_{v1}(t) + Q_{2'}(t) = K$, $\pi_{v1}(i) = \pi_{2'}(K-i)$ for all $0 \leq i \leq K$. Moreover, $\pi_{D_{v1}}(i) = \pi_{A_{2'}}(K-1-i|\Delta Q_0 = 1)\ 0 \leq i \leq K-1$. Hence,

$$\begin{aligned}\pi_{D_{v1}}(i) = \pi_{A_{2'}}(K-1-i|\Delta Q = 1) &= \frac{\pi_{A_{2'}}(K-1-i)}{\sum_{j=1}^{K-1}\pi_{A_{2'}}(j)} \\ &= \frac{\pi_{A_{2'}}(K-1-i)}{1-\pi_{A_{2'}}(K)} = \frac{\pi_{A_{2'}}(K-1-i)}{1-P_L}\end{aligned} \tag{9}$$

and vice versa. Substituting equation (1), that holds for all work conserving disciplines, into equation (9) results in $\pi_{D_{v1}}(i) = \pi_{D_{2'}}(K-1-i)$ for all $0 \leq i \leq K-1$. Next, notice that the inter-arrival and service times of *system* $v2$ are distributed as their counterparts in *system* $2'$. Thus, *systems* $2'$ and 2 are stochastically equivalent. Hence, equations (7) and (8) hold for any distribution of G_1 and G_2. □

Remark 1 The above proof does not require the inter-arrival and service time distributions to be independent and hence holds for stationary queues with general stationary arrivals and services that could be state-dependent.

As stated before, an exact duality theorem (in the form of an equality) does not exist for FIFO queues with general arrival processes and service times. This is because of the lack of symmetry between the arrivals to a full queue (that are lost) and lack of departures from the empty queue. Theorem 1 suggests that a duality relation is possible if we alter the behaviour of the queues when they are full [empty] to resemble the behaviour of their duals when they are empty [full] by using an interrupted arrival [virtual service] discipline.
We now provide a comparison between the queue length distribution of a $G_1/G_2/1/K$ queueing system with interrupted arrival or virtual service discipline with a $G_1/G_2/1/K$ queueing system with FIFO discipline via stochastic majorization.
In the sequel, when use the notation $X \leq_{st} [\geq_{st}] Y$ to denote stochastic dominance, i.e. X is stochastically smaller [greater] than random variable Y if $\mathbb{P}\{X \geq \alpha\} \leq [\geq] \mathbb{P}\{X \geq \alpha\}\ \forall \alpha > 0$.

Theorem 2 *Let queue-I denote a $G_1/G_2/1/K$ queue with interrupted arrivals and queue-II denote a $G_1/G_2/1/K$ queue also with interrupted arrivals. Let $R_{A|Q_D^{K-1}}$ denote the time till next arrival from the moment a customer leaves the full queue and let A be a random variable that is distributed as regular inter-arrival times of the arrival process of system I.*

1. *If $R_{A|Q_D^{K-1}} \leq_{st} A$, then*

$$\mathbb{P}\{Q_{D_I}[n] \leq i\} \leq \mathbb{P}\{Q_{D_{II}}[n] \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \tag{10}$$
$$\mathbb{P}\{Q_I(t) \leq i\} \leq \mathbb{P}\{Q_{II}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \tag{11}$$

2. *Conversely, if $R_{A|Q_D^{K-1}} \geq_{st} A$, then*

$$\mathbb{P}\{Q_{D_I}(t) \leq i\} \geq \mathbb{P}\{Q_{D_{II}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \tag{12}$$
$$\mathbb{P}\{Q_I(t) \leq i\} \geq \mathbb{P}\{Q_{II}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \tag{13}$$

Proof Let us denote the counting process associated with the arrivals by $G_A(t)$, the counting process associated with the departure process by $G_S(t)$, the probability distribution of a given stochastic process $B(t)$ by $L(B)$, the service time of the n-th accepted customer by $S[n]$, the length of the inter-arrival time at the end of which the n-th accepted customer arrives at the queue by $A[n]$, queue length at the time of the n-th accepted customer by $Q_{A'}[n]$, and the residual inter-arrival time at time t by $R_A[t]$.
The argument is similar to the argument presented in [10, Thm 1]. We construct two new queueing systems on the same probability space such that $\widetilde{Q}_{D_I}[n] \geq \widetilde{Q}_{D_{II}}[n]$ for all n, $L(\widetilde{Q}_{D_I}) = L(Q_{D_I})$, and $L(\widetilde{Q}_{D_{II}}) = L(Q_{D_{II}})$. To do this, use any arrival and service processes for *system I* such that $L(G_{A_I}) = L(\widetilde{G}_{A_I})$ and $L(G_{S_I}) = L(\widetilde{G}_{S_I})$ and by construction, take $\widetilde{A}_{II}[n] = \widetilde{A}_I[n]$ if $\widetilde{Q}_{A'_I}[n] < K-1$, $\widetilde{A}_{II}[n] \geq \widetilde{A}_I[n]$ if $\widetilde{Q}_{A'_I}[n] = K-1$, and $\widetilde{S}_{II}[n] = \widetilde{S}_I[n]$ for all n.
Let us show that if a proper starting point is assumed for the queues, this construction guarantees that for all n, either $\widetilde{Q}_{D_I}[n] > \widetilde{Q}_{D_{II}}[n]$ or $\widetilde{Q}_{D_I}[n] =$

$\widetilde{Q}_{D_{II}}[n]$ and $\widetilde{R}_{A_I}(\widetilde{T}_{D_I}[n]) \leq \widetilde{R}_{A_{II}}(\widetilde{T}_{D_{II}}[n])$. Let us use mathematical induction and show that if the above property holds for n, it holds for $n+1$, as well. Consider $t_i[n] = \widetilde{T}_{D_i}[n+1] - \widetilde{S}_i[n]$, which is equal to $\widetilde{T}_{D_i}[n]$ if and only if $\widetilde{Q}_{D_i}[n] > 0$. If $\widetilde{Q}_{D_i}[n] = 0$, $\widetilde{Q}_i(t_i[n]) = 1$ and $R_{A_i}(t_i[n]) = \widetilde{A}_i[n+2]$. It is straight forward to show that the above property holds at $t = t_i[n]$. Thus, for any $0 \leq t \leq \widetilde{S}_i[n+1]$ we have either $\widetilde{Q}_I(\widetilde{T}_{D_I}[n]+t) > \widetilde{Q}_{II}(\widetilde{T}_{D_{II}}[n]+t)$ or $\widetilde{Q}_I(\widetilde{T}_{D_I}[n]+t) = \widetilde{Q}_{II}(\widetilde{T}_{D_{II}}[n]+t)$ and $\widetilde{R}_{A_I}(\widetilde{T}_{D_I}[n]+t) \leq \widetilde{R}_{A_{II}}(\widetilde{T}_{D_{II}}[n]+t)$ because $\widetilde{A}_I[n] \leq \widetilde{A}_{II}[n]$ and $\widetilde{S}_I[n] = \widetilde{S}_{II}[n]$. Hence, the property holds for $n+1$, as well. Therefore, $\widetilde{Q}_{D_I}[n] \geq \widetilde{Q}_{D_{II}}[n]$ for all n. Thus, $\mathbb{1}\{\widetilde{Q}_{D_I}[n] \leq j\} \leq \mathbb{1}\{\widetilde{Q}_{D_{II}}[n] \leq j\}$. From the finiteness of the queues and the strong law of large numbers, we conclude that $\mathbb{P}\{\widetilde{Q}_{D_I}[n] \leq i\} \leq \mathbb{P}\{\widetilde{Q}_{D_{II}}[n] \leq i\}$; meaning that equation (10) holds.
Furthermore for $m \geq 0$, we have

$$\int_{\widetilde{T}_{D_i}[n]}^{\widetilde{T}_{D_i}[n+1]} \mathbb{1}\{\widetilde{Q}_i(s) > m\}ds = \int_{\widetilde{T}_{D_i}[n]}^{t_i[n]} \mathbb{1}\{\widetilde{Q}_i(s) > m\}ds + \int_{t_i[n]}^{\widetilde{T}_{D_i}[n+1]} \mathbb{1}\{\widetilde{Q}_i(s) > m\}(s)ds$$

Where $\int_{\widetilde{T}_{D_{II}}[n]}^{t_{II}[n]} \mathbb{1}\{\widetilde{Q}_{II}(s) > m\}ds = 0$, $\int_{\widetilde{T}_{D_I}[n]}^{t_I[n]} \mathbb{1}\{\widetilde{Q}_I(s) > m\}ds \geq 0$, and $\int_{t_{II}[n]}^{\widetilde{T}_{D_{II}}[n+1]} \mathbb{1}\{\widetilde{Q}_{II}(s) > m\}ds \leq \int_{t_I[n]}^{\widetilde{T}_{D_I}[n+1]} \mathbb{1}\{\widetilde{Q}_I(s) > m\}ds$. Therefore,

$$\int_{\widetilde{T}_{D_{II}}[n]}^{\widetilde{T}_{D_{II}}[n+1]} \mathbb{1}\{\widetilde{Q}_{II}(s) > m\}ds \leq \int_{\widetilde{T}_{D_I}[n]}^{\widetilde{T}_{D_I}[n+1]} \mathbb{1}\{\widetilde{Q}_I(s) > m\}ds \quad \forall n \geq 1$$

Hence,

$$\mathbb{E}_D[\int_0^{T_{D_{II}}} \mathbb{1}\{Q_{II}(s) > m\}ds] \leq \mathbb{E}_D[\int_0^{T_{D_{II}}} \mathbb{1}\{Q_I(s) > m]ds\}$$

Where T_{D_i} is the stationary inter-point time of Q_{D_i}. Moreover, since $\widetilde{T}_{D_{II}}[n+1] - \widetilde{T}_{D_{II}}[n] \geq \widetilde{T}_{D_I}[n+1] - \widetilde{T}_{D_I}[n]$ for all n, $\mathbb{E}_D[T_{D_{II}}] \geq \mathbb{E}_D[T_{D_I}]$. Thus, $\lambda_{D_{II}} \leq \lambda_{D_I}$, where λ_{D_i} is the mean rate of Q_{D_i}. Using the Palm inversion formula introduced in [12], we obtain that

$$\mathbb{P}\{Q_{II} > m\} = \mathbb{E}[\mathbb{1}\{Q_{II} > m\}] = \lambda_{D_{II}} \cdot \mathbb{E}_D[\int_0^{T_{D_{II}}} \mathbb{1}\{Q_{II}(s) > m\}ds]$$
$$\leq \lambda_{D_I} \cdot \mathbb{E}_D[\int_0^{T_{D_I}} \mathbb{1}\{Q_I(s) > m\}ds] = \mathbb{E}[\mathbb{1}\{Q_I > m\}] = \mathbb{P}\{Q_I > m\}$$

Thus,

$$\mathbb{P}\{Q_{II} \leq m\} = 1 - \mathbb{P}\{Q_{II} > m\} \geq 1 - \mathbb{P}\{Q_I > m\} = \mathbb{P}\{Q_I \leq m\}$$

Hence, equation (11). Proof of part 2, follows in a similar manner. □

Now, we investigate the relationship between the residual inter-arrival time $R_{A|Q_D^{K-1}}$ and the inter-arrival time under the Palm distribution. Let us denote the residual inter-arrival time seen by a departure that leaves i customers in the queue by $R_{A|Q_D^i}$ and the residual service time seen by an arrival that finds i customers in the queue by $R_{S|Q_A^i}$. To the best of our knowledge, the distribution of the residual times defined above, for queues with general arrival processes and service times is unknown. In fact, according to [13], one can extract the queue length distribution of a $GI/GI/1/K$ queueing system if one can calculate $E[R_{A|Q_D^i}]$ and $E[R_{S|Q_A^i}]$. The residual life of a renewal process seen at arbitrary times has been well studied and the ratio of the expectation of the residual life to the expectation of a regular inter-point time is a function of the coefficient of variation (C) of the inter-point times, see [5, 12, **?**] for example. For $C \leq 1$, the mean residual life is less than or equal to the mean inter-point time and for $C \geq 1$, the opposite. One can even compare the random variables in the stochastic ordering sense. There are no general stochastic orderings available for the residual life and they depend on the form of the distribution, i.e. whether it is more or less variable than the exponential case. In order to do so, we need the following stochastic orders [15].

Definition 2 [15]

1. The random variable X is IFR if and only if, $[X-t|X \geq t] \geq_{st} [X-t_1|X \geq t_1]$ whenever $t \leq t_1$ and X is DFR if and only if, $[X-t|X \geq t] \leq_{st} [X-t_1|X \geq t_1]$ whenever $t \leq t_1$. [1]
2. The non-negative random variable X is NBU if and only if $X \geq_{st} [X-t|X \geq t]$ for all $t \geq 0$ and X is NWU if and only if, $X \leq_{st} [X-t|X \geq t]$ for all $t \geq 0$.[2]

According to the definition, IFR [DFR] random variables are a subclass of NBU [NWU] random variables.

We have the following result.

Theorem 3 *If the inter-arrival times in a $GI/GI/1/K$ queueing system are NBU [NWU], the residual inter-arrival time $R_{A|Q_D^{K-1}}$ is stochastically smaller [greater] than a regular inter-arrival time.*

Proof Let us denote the distribution of the backward recurrence times of the arrival process at the instances of interest by $F_{B_A}(x) = \mathbb{P}\{B_A \leq x\}$, and the distribution of the inter-arrival times by $F_A(x) = \mathbb{P}\{A \leq x\}$. Then if $R_{A|Q_D^{K-1}} \leq_{st} A$,

$$\begin{aligned}\mathbb{P}\{R_A \geq x\} &= \int_0^\infty \mathbb{P}\{R_A \geq x, B_A = \alpha\} d\alpha \\ &= \int_0^\infty \mathbb{P}\{R_A \geq x | B_A = \alpha\} \mathbb{P}\{B_A = \alpha\} d\alpha\end{aligned}$$

[1] By IFR [DFR], we mean Increasing[Decreasing] Failure Rate

[2] By NBU [NWU], we mean New Better[Worse] than Used

$$= \int_0^\infty \mathbb{P}\{A - B_A \geq x | A \geq \alpha\} \mathbb{P}\{B_A = \alpha\} d\alpha \leq \int_0^\infty \mathbb{P}\{A \geq x\} \mathbb{P}\{B_A = \alpha\} d\alpha$$
$$= \mathbb{P}\{A \geq x\} \cdot \int_0^\infty \mathbb{P}\{B_A = \alpha\} d\alpha = \mathbb{P}\{A \geq x\}$$

The proof of $R_{A|Q_D^{K-1}} \geq_{st} A$ for NWU inter-arrival times follows similarly. □

Combining the above results, gives us the opportunity to compare the queue length distribution of a $GI/GI/1/K$ queueing system with the queue length distribution of its dual under certain conditions on the inter-arrival and service time distributions.

Theorem 4 *Let us call $G_1/G_2/1/K$ as system 1 and the queue $G_2/G_1/1/K$ as system 2. The following properties hold,*

1. *If the inter-arrival and service times are NBU, then*

$$\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) \leq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j)$$
$$= \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad \forall\, 0 \leq i \leq K-1 \tag{14}$$

and

$$\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) \leq \sum_{j=K-i}^{K} \pi_2(j)$$
$$= \mathbb{P}\{Q_2(t) \geq K-i\}, \quad \forall\, 0 \leq i \leq K \tag{15}$$

2. *If the inter-arrival and service times are NWU, then*

$$\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) \geq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j)$$
$$= \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad \forall\, 0 \leq i \leq K-1 \tag{16}$$

and

$$\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) \geq \sum_{j=K-i}^{K} \pi_2(j)$$
$$= \mathbb{P}\{Q_2(t) \geq K-i\}, \quad \forall\, 0 \leq i \leq K \tag{17}$$

Proof The proof is straightforward. We will prove part 1 and part 2 follows similarly. Assume there exists a $G_1/G_2/1/K$ queue with interrupted arrival discipline that we will call *system* $1'$ and a $G_2/G_1/1/K$ queue with interrupted arrival discipline that we will call *system* $2'$. Using Theorems 2 and 3 for the primal queue, we have

$$\mathbb{P}\{Q_{D_1}(t) \leq i\} \leq \mathbb{P}\{Q_{D_{1'}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1$$

$$\mathbb{P}\{Q_1(t) \leq i\} \ \leq \ \mathbb{P}\{Q_{1'}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K$$

Similarly, for the dual queue, we have

$$\mathbb{P}\{Q_{D_2}(t) \leq i\} \ \leq \ \mathbb{P}\{Q_{D_{2'}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1$$

$$\mathbb{P}\{Q_2(t) \leq i\} \ \leq \ \mathbb{P}\{Q_{2'}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K,$$

Moreover, Theorem 1 asserts that

$$\begin{aligned}\mathbb{P}\{Q_{D_{1'}}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_{1'}}(j) \ &= \sum_{j=K-1-i}^{K-1} \pi_{D_{2'}}(j) \\ = \mathbb{P}\{Q_{D_{2'}}(t) \geq K-1-i\}, \quad &\forall\, 0 \leq i \leq K-1\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\{Q_{1'}(t) \leq i\} = \sum_{j=1}^{i} \pi_{1'}(j) \ &= \sum_{j=K-i}^{K} \pi_{2'}(j) \\ = \mathbb{P}\{Q_{2'}(t) \geq K-i\}, \quad &\forall\, 0 \leq i \leq K\end{aligned}$$

Substituting the equations resulted from Theorems 2 and 3 into the result of Theorem 1, concludes the proof. □

Corollary 1 *Theorem 4 relates* $\mathbb{P}\{Q_1(t) \leq i\}$ *to* $\mathbb{P}\{Q_2(t) \geq K-i\}$. *Using* $\sum_{j=0}^{K} \mathbb{P}\{Q(t) = j\} = 1$, *we can extract the same result for* $\mathbb{P}\{Q_2(t) \leq i\}$ *and* $\mathbb{P}\{Q(t)_1 \geq K-i\}$. *The same holds for the queue length distribution at departure instances.*

Corollary 2 *A special case of Theorem 4,* $i = 0$, *results in* $P_{s_1} = \pi_{D_1}(0) \leq \pi_{D_2}(K-1)$ *and* $\pi_1(0) \leq \pi_2(K-1)$ *for NBU inter-arrival and service times. The opposite holds for NWU inter-arrival and service times. Such bounds are very helpful in studying the starvation of finite buffer queues.*

Since a exact duality theorem in the form of an equality does not hold for FIFO queueing systems with general arrival and service times, we have combined the duality relations introduced in Theorem 1 and stochastic orders to get some useful bound for the queue length distribution of a FIFO queueing system.

Remark 2 Though the relations proposed in Theorem 4 only hold for two specific classes of inter-arrival and service times, they encompass many common models such as queues with deterministic, uniform, Erlang-K , and exponential inter-arrival and service times. Moreover, the exponential random variable is the only random variable that is both NBU and NWU. Hence, in $M/G/1/K$ queues, we need only consider the service time distribution.

2.1 $M/G/1/K$

In this section, we consider a $M/G/1/K$ queueing system as an example. Instead of using the duality result of queues with interrupted arrival discipline, we will prove a theorem similar to Theorem 2 for the special case of queues with Poisson arrivals.

Theorem 5 *Let system I be a $M/G/1/K$ queue with FIFO discipline, and system II a $M/G/1/K$ with virtual service discipline. Let $R_{S|Q_A^0}$ denote the service time of the first customer of each busy period in system II and let a regular service time be denoted by S.*

1. *If $R_{S|Q_A^0} \leq_{st} S$, then*

$$\mathbb{P}\{Q_{D_I}(t) \leq i\} \leq \mathbb{P}\{Q_{D_{II}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \tag{18}$$
$$\mathbb{P}\{Q_I(t) \leq i\} \leq \mathbb{P}\{Q_{II}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \tag{19}$$

2. *Conversely, if $R_{S|Q_A^0} \geq_{st} S$, then*

$$\mathbb{P}\{Q_{D_I}(t) \leq i\} \geq \mathbb{P}\{Q_{D_{II}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \tag{20}$$
$$\mathbb{P}\{Q_I(t) \leq i\} \geq \mathbb{P}\{Q_{II}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \tag{21}$$

Due to PASTA property introduced in [16], a similar result holds for the queue length distribution seen at arrival instances.

Proof We will prove part 1 and the proof of part 2 follows similarly. The argument is similar to Theorem 2 and we will use the same notation. We will construct two new queueing systems on the same probability space such that the distributions of the arrival and service processes of the primal queueing systems are preserved. Use any arrival and service processes for *system I* such that $L(G_{A_I}) = L(\widetilde{G}_{A_I})$ and $L(G_{S_I}) = L(\widetilde{G}_{S_I})$. We will propose a construction for *system II* that ensures that $\widetilde{Q}_{D_I}[n] \geq \widetilde{Q}_{D_{II}}[n]$ for all n. Let us use mathematical induction. We assume that $\widetilde{Q}_{D_I}[k] \geq \widetilde{Q}_{D_{II}}[k]$ for $k \leq m$ and we want to prove that $\widetilde{Q}_{D_I}[m+1] \geq \widetilde{Q}_{D_{II}}[m+1]$. If $\widetilde{Q}_{D_{II}}[m] > 0$, we choose the $(m+1)$-th service time of *system II* equal to the one in *system I*. Since arrivals are Poisson and exponential random variables are memoryless, we can construct the same arrivals for *system II* during the $(m+1)$-th service time. Hence, $\widetilde{Q}_{D_I}[m+1] \geq \widetilde{Q}_{D_{II}}[m+1]$. On the other hand, if $\widetilde{Q}_{D_{II}}[m] = 0$, because the exceptional first service time is assumed to be stochastically smaller than a regular service time, we will choose a service time with a length less than or equal to the length of the service time of *system I* for *system II*. Then, because exponential random variables are memoryless, the number of arrivals during the $(m+1)$-th service time of *system II* will be less than or equal to the number of arrivals in *system I*. Thus, $\widetilde{Q}_{D_I}[m+1] \geq \widetilde{Q}_{D_{II}}[m+1]$, meaning that equation 18 holds.
For proving equation 19, we will use a result proposed in *Theorem 1* of [9]. Let us denote the epochs that the k-th arrival comes to enter the system by A_k, the

epochs that the k-th admitted customer is admitted by B_k, and the probability of loss by P_L. Since the arrival processes of the above-defined systems are the same, for every sample path of *system I*, there exists a sample path of *system II* that satisfies $A_{k_I} = A_{k_{II}}$. Moreover, *Theorem 1* of [9] suggests that for the above-mentioned systems, $B_{k_I} \geq B_{k_{II}}$. Hence, we have $P_{L_I} \geq P_{L_{II}}$. Therefore,

$$\begin{aligned}\mathbb{P}\{\widetilde{Q}_{A_I}(t) \leq i\} &= (1 - P_{L_I}) \cdot \mathbb{P}\{\widetilde{Q}_{D_I}(t) \leq i\} \\ \leq (1 - P_{L_{II}}) \cdot \mathbb{P}\{\widetilde{Q}_{D_{II}}(t) \leq i\} &= \mathbb{P}\{\widetilde{Q}_{A_{II}}(t) \leq i\}\end{aligned}$$

Using the PASTA property and the fact that the queue length distributions of the constructed and primal queues are the same, we obtain equation (19). □

Thus, duality results of both the interrupted arrival discipline and the virtual service discipline can be used to extract bounds for finite buffer FIFO queues. Next, we propose a bound similar to the one introduced in Theorem 4.

Corollary 3 *Let queue* 1 *be a* $M/G/1/K$ *queue and its dual* $GI/M/1/K$ *be called system* 2.

1. *If service times of the* $M/G/1/K$ *queueing system (G) are NBU,*

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) &\leq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\ = \mathbb{P}\{Q_{D_2}(t) \geq K - 1 - i\}, &\quad \forall\, 0 \leq i \leq K-1 \end{aligned} \tag{22}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) &\leq \sum_{j=K-i}^{K} \pi_2(j) \\ = \mathbb{P}\{Q_2(t) \geq K - i\}, &\quad \forall\, 0 \leq i \leq K \end{aligned} \tag{23}$$

2. *If service times of the* $M/G/1/K$ *queueing system (G) are NWU,*

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) &\geq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\ = \mathbb{P}\{Q_{D_2}(t) \geq K - 1 - i\}, &\quad \forall\, 0 \leq i \leq K-1 \end{aligned} \tag{24}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) &\geq \sum_{j=K-i}^{K} \pi_2(j) \\ = \mathbb{P}\{Q_2(t) \geq K - i\}, &\quad \forall\, 0 \leq i \leq K \end{aligned} \tag{25}$$

Proof We omit the proof since it is similar to Theorem 4. □

So far using the duality relations for queueing systems with interrupted arrival and virtual service discipline and stochastic orders, we have found some relations between the queue length distribution of a finite buffer FIFO queueing system with NBU or NWU inter-arrival and service times and the counterpart in its dual. Next, we will consider fluid flow queues and study duality results in such finite buffer queues.

3 Fluid Flow Model

In this section, we will consider duality results concerning finite buffer fluid flow queues and relate the workload distribution of a queue, which we will call *system* 1, with cumulative fluid input $C_1(t)$ and cumulative available processing $S_1(t)$ to the workload distribution of a queue, which we will call *system* 2, with cumulative fluid input $C_2(t) \sim S_1(t)$ and cumulative available processing $S_2(t) \sim C_1(t)$. Then, we will show an interesting relation between the overflow and underflow rates and processes of the queues defined above. In the end, we will propose a duality result concerning fluid flow queues with ON-OFF fluid input and constant available processing. To analyse fluid queues we need the notion of a fluid Palm measure. see [3,12] for details.

Proposition 1 *Let us denote the fluid Palm measure of the remaining workload of a finite buffer fluid flow queue, associated with fluid input and output of the queue, respectively by* $\mathbb{P}_A\{W \leq x\}$ *and* $\mathbb{P}_D\{W \leq x\}$*. Also, let us denote the fluid Palm measure associated with the fluid input that entered the queue by* $\mathbb{P}_{A'}\{W \leq x\}$*. Then,*

$$\mathbb{P}_A\{W \leq x\} \leq \mathbb{P}_{A'}\{W \leq x\} = \mathbb{P}_D\{W \leq x\} \tag{26}$$

Proof By changing equation 2.41 of [12] to fit a finite buffer queue, we obtain the following results, in which C'_t is the cumulative fluid input that has entered the queue up to time t.

$$f(W_t) = f(W_0) + \int_0^t f'(W_s)\mathbb{1}\{W_s > 0\}d(C'_s - S_s)$$

Taking the expectation of both sides, we obtain

$$\mathbb{E}\{f(W_t)\} = \mathbb{E}\{f(W_0)\} + \mathbb{E}\{\int_0^t f'(W_s)\mathbb{1}\{W_s > 0\}d(C'_s - S_s)\}$$

Hence,

$$\mathbb{E}\{\int_0^t f'(W_s)\mathbb{1}\{W_s > 0\}d(C'_s - S_s)\} = 0$$

Then,

$$\mathbb{E}\{\int_0^t f'(W_s)\mathbb{1}\{W_s > 0\}dC'_s\} = \mathbb{E}\{\int_0^t f'(W_s)\mathbb{1}\{W_s > 0\}dS_s\}$$

By taking $f'(W) \triangleq \mathbb{1}\{W > x\}$, we have

$$t \cdot \mathbb{E}\{\int_0^1 \mathbb{1}\{W_s > x\}dC'_s\} = t \cdot \mathbb{E}\{\int_0^1 \mathbb{1}\{W_s > x\}dS_s\}$$

Hence,

$$\lambda_{A'} \cdot \mathbb{P}_{A'}\{W_s > x\} = \lambda_D \cdot \mathbb{P}_D\{W_s > x\}$$

And since the queue is stable, the rate of the input that is accepted to the queue equals the output rate. In other words, $\lambda_{A'} = \lambda_D$.

$$\mathbb{P}_{A'}\{W_s > x\} = \mathbb{P}_D\{W_s > x\}$$

The left hand side inequality follows from the definition of fluid Palm measure in equation 1.22 of [12]. Fluid input A is the sum of the fluid input that entered the queue A' and the portion of fluid input that is lost. The portion that is lost only sees the full queue. Hence, according to the definition of fluid Palm measure, the fluid Palm measure of $\{W \leq x\}$ for any $x \leq K$ associated with the fluid input is less than or equal to the fluid Palm measure associated with the fluid input that enters the queue. □

The result of Proposition 1 resembles the relation between the queue length distribution at arrival and departure instances of a finite buffer queue when using the point process model.

$$\mathbb{P}_A\{Q \leq i\} \;\leq\; \mathbb{P}_{A'}\{Q \leq i\} \;\triangleq\; \frac{\mathbb{P}_A\{Q \leq i\}}{1 - \mathbb{P}_A\{Q = K\}} \;=\; \mathbb{P}_D\{Q \leq i\}\;,\; \forall\, i < K$$

where, the left-most term is the queue length distribution at arrival instances, the term in the middle is the queue length distribution at the instances of arrivals that enter the queue, and the right-most term is the queue length distribution at departure instances. Then we can show the following duality results for general fluid flow queues.

Theorem 6 *Let us call the fluid queue with cumulative input $C_1(t)$ and available processing $S_1(t)$ system 1 and its dual with cumulative input $C_2(t) \sim S_1(t)$ and available processing $S_2(t) \sim C_1(t)$, system 2. Then,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{27}$$

$$\mathbb{P}_D\{W_1(t) \leq \beta\} = \mathbb{P}_D\{W_2(t) \geq K - \beta\} \tag{28}$$

$$\mathbb{P}_A\{W_1(t) \leq \beta\} \leq \mathbb{P}_A\{W_2(t) \geq K - \beta\} \tag{29}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{30}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1}L_1(t) = \lim_{t\to\infty} t^{-1}U_2(t) = \Lambda_2 \tag{31}$$

Proof Let us use the "role inversion" model introduced in [5] and later used in [6]. This model inverts the roles of a customer and an empty space in finite buffer queues. In other words, empty buffer spaces of *system* 1 can be thought of as the occupied spaces (customers) in a second queue, which we will call *system* $1'$, and vice versa. The input fluid to *system* 1 is the available processing for *system* $1'$ and vice versa. Thus, *systems* $1'$ and 2 are stochastically equivalent, meaning that their remaining workloads have the same distribution. Therefore, since remaining workload of *systems* 1 and 2 are related as $W_1(t) = K - W_{1'}(t)$, for every sample path of *system* 1, there exists a sample path of *system* 2 that satisfies $W_1(t) = K - W_2(t)$, hence equation (27).
Note that when *system* 1 is full, the fluid input sees a full system with a rate equal to the instantaneous fluid input rate[3] but the fluid output of *system* $1'$ sees an empty queue with a rate equal to the instantaneous rate of its fluid input, which is less than or equal to the instantaneous available processing because *system* $1'$ is empty. Thus,

$$\mathbb{P}_{A'}\{W_1(t) \leq \beta\} = \mathbb{P}_D\{W_{1'}(t) \geq K - \beta\}$$

Using Theorem 1 and the fact that *systems* 1 and 2 are stochastically equivalent, we obtain (28). Moreover, using equation (28) and the inequality of Theorem (1), we obtain (29).
Based on the definition of underflow and overflow processes and the role inversion model, $U_1(t) = L_{1'}(t)$. Then, since *systems* $1'$ and 2 are stochastically equivalent, for every sample path of the overflow process of *system* 1, there exists a sample path of the underflow process of *system* 2 that satisfies $U_1(t) = L_2(t)$ and vice versa. Hence, we obtain (30) and (31) follows by definition of the rates. □

In some cases, it is of interest to compare the fluid Palm measure of the remaining workload associated with the input of the primal queue with the fluid Palm measure associated with the output of its dual. Corollary 4 studies this relation.

Corollary 4 *For systems* 1 *and* 2 *defined in Theorem 6, we have*

$$\mathbb{P}_A\{W_1(t) \geq K - \beta\} \geq \mathbb{P}_D\{W_2(t) \leq \beta\} \tag{32}$$

Proof Substituting the inequality proposed in Proposition 1 into equation (28) results in

$$\mathbb{P}_A\{W_1(t) \leq \beta\} \leq \mathbb{P}_D\{W_2(t) \geq K - \beta\}$$

Since this inequality holds for all values of β, it also holds for $\beta' = \beta - \frac{\delta}{n}$, where δ is some positive real number that satisfies $\beta - \frac{\delta}{n} \geq 0$ for all $n \geq 1$.

[3] the rate is the derivative of C_t

Since the sequences of events $\{C_n\} \triangleq \{W_1(t) \leq \beta - \frac{\delta}{n}\}$ and $\{D_n\} \triangleq \{W_2(t) \geq K - (\beta - \frac{\delta}{n})\}$ are increasing,

$$\mathbb{P}_A\{\bigcup_{n=1}^{\infty} C_n\} = \lim_{n\to\infty} \mathbb{P}_A\{C_n\}$$
$$\mathbb{P}_D\{\bigcup_{n=1}^{\infty} D_n\} = \lim_{n\to\infty} \mathbb{P}_D\{D_n\}$$

By substituting this into the inequality above, we obtain

$$\begin{aligned}\mathbb{P}_A\{W_1(t) < \beta\} = \mathbb{P}_A\{\textstyle\bigcup_{n=1}^{\infty} C_n\} = \lim_{n\to\infty} \mathbb{P}_A\{C_n\} \\ = \lim_{n\to\infty} \mathbb{P}_A\{W_1(t) \leq \beta - \tfrac{\delta}{n}\}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_D\{W_2(t) > K - (\beta - \tfrac{\delta}{n})\} = \mathbb{P}_D\{\textstyle\bigcup_{n=1}^{\infty} D_n\} = \lim_{n\to\infty} \mathbb{P}_D\{D_n\} \\ = \lim_{n\to\infty} \mathbb{P}_D\{W_2(t) \geq K - (\beta - \tfrac{\delta}{n})\}\end{aligned}$$

Let us define sequences of real numbers $a_n \triangleq \mathbb{P}_A\{W_1(t) \leq \beta - \frac{\delta}{n}\}$ and $b_n \triangleq \mathbb{P}_D\{W_2(t) \geq K - (\beta - \frac{\delta}{n})\}$. Since $a_n \leq b_n$ for all $n \geq 1$, their limit must follow the same ordering, meaning that $\lim_{n\to\infty} \mathbb{P}_A\{W_1(t) \leq \beta - \frac{\delta}{n}\} \leq \lim_{n\to\infty} \mathbb{P}_D\{W_2(t) \geq K - (\beta - \frac{\delta}{n})\}$. Hence,

$$\mathbb{P}_A\{W_1(t) < \beta\} \leq \mathbb{P}_D\{W_2(t) > K - \beta\}$$

Then,

$$\begin{aligned}\mathbb{P}_A\{W_1(t) \geq \beta\} = 1 - \mathbb{P}_A\{W_1(t) < \beta\} &\geq 1 - \mathbb{P}_D\{W_2(t) > K - \beta\} \\ &= \mathbb{P}_D\{W_2(t) \leq K - \beta\}\end{aligned}$$

By interchanging β and $K - \beta$, we obtain equation (32). □

Remark 3 In Theorem 4, duality relations for point process inputs are given by inequalities, while theorem 6 relates the remaining workload distribution of a fluid queue to the remaining workload distribution of its dual via an equality.

Finite buffer fluid queues with an ON-OFF fluid input and constant available processing have received much attention recently due to applications in modern communication systems and energy systems. We will study these queues as a special case of the more general fluid input case. Before doing so, let us define a few new quantities. Define the loss event L, to be the event that the fluid queue is full and the instantaneous input rate is greater than the instantaneous available processing. In other words, $\{L\} = \{W(t) = K\} \bigcap \{C(t) > S(t)\}$. Similarly, define the starvation event S to be the event that the fluid queue is empty and the instantaneous input rate is less than the instantaneous

available processing. In other words, $\{L\} = \{W(t) = 0\} \bigcap \{C(t) < S(t)\}$. The loss and starvation rates are define as follows.

$$R_L \triangleq \lim_{t\to\infty} \frac{1}{C(t)} \int_0^t \mathbb{1}\{\omega(s) \in L\} d(C(s) - S(s)) = \lim_{t\to\infty} \frac{U(t)}{C(t)} \tag{33}$$

$$R_S \triangleq \lim_{t\to\infty} \frac{1}{S(t)} \int_0^t \mathbb{1}\{\omega(s) \in S\} d(S(s) - C(s)) = \lim_{t\to\infty} \frac{L(t)}{S(t)} \tag{34}$$

In other words, the loss rate is the long run portion of the lost input and the starvation rate is the long run portion of the unused (lost) processing.

Theorem 7 *Consider two ON-OFF fluid queues with constant available processing and call them systems* 1 *and* 2*. Denote the ON periods, OFF periods, input rate during the ON period, and the constant available processing of system* i *by* A_i*,* B_i*,* H_i*, and* C_i*, respectively. Moreover, assume* $A_2 \sim B_1$*,* $B_2 \sim A_1$*,* $H_2 = H_1$ *and* $C_2 = H_1 - C_1$*. Then,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{35}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{36}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1} L_1(t) = \lim_{t\to\infty} t^{-1} U_2(t) = \Lambda_2 \tag{37}$$

$$R_{L_1} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \cdot R_{S_2} \tag{38}$$

Notice that a relation similar to equations 28 and 29 does not hold for the aforementioned queues.

Proof From Theorem 6, dual of *system* 1 is a fluid queue with a constant input of C_1 and an ON-OFF available processing with ON and OFF periods distributed respectively as A_1 and B_1 with a constant available processing during the ON periods of H_1. We will call the dual queue, *system* 3. Now, let us define a third queue, which we will call *system* 2, with an ON-OFF fluid input, constant input rate during the ON period, and a constant available processing such that $W_2(t) = W_3(t)$. For this to hold, the third queue must be ON when the available processing in the dual queue is OFF and vice versa. Furthermore, constant available processing and input rate during the ON periods in the third queue must respectively be equal to $H_1 - C_1$ and C_1 to ensure that the two queues have the same remaining workload processes. Hence, based on $W_2(t) = W_3(t)$ and equation (27), the relation (35) follows. Moreover, since *systems* 2 and 3 both overflow at an instantaneous rate of C_1 and underflow at an instantaneous rate of $H_1 - C_1$, and $W_2(t) = W_3(t)$, we have $L_2(t) = L_3(t)$ and $U_2(t) = U_3(t)$. By substituting this into equation (30), we obtain equations (36) and (37).
Based on the definition of *system* 2, we have

$$\frac{\mathbb{E}[S_2(t)]}{\mathbb{E}[S_3(t)]} = \frac{(H_1 - C_1) \cdot t}{H_1 \cdot \frac{\mathbb{E}[A_1]}{\mathbb{E}[A_1]+\mathbb{E}[B_1]} \cdot t} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \tag{39}$$

Then, since $L_2(t) = L_3(t)$, by the definition of the starvation and loss rates, we have

$$R_{S_3} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \cdot R_{S_2} \tag{40}$$

Moreover, using equation (30) and the fact that $C_1(t) = S_3(t)$, because they are duals, we obtain $R_{L_1} = R_{S_3}$ and hence (38).

□

Using the above equation, we can translate the problem of finding the overflow/underflow rate of a fluid queue with ON-OFF input and constant available processing to the problem of finding the underflow/overflow in another fluid queue. We can thus generalize the theorem to fluid queues with bounded fluid inputs. We define a bounded fluid input process to be a fluid process in which the derivative of the cumulative input is bounded. In other words, $\frac{dC(t)}{dt} \leq M$.

Corollary 5 *Consider two fluid queues with constant available processing and fluid inputs bounded by the value of M and call them systems 1 and 2. Moreover, assume that $C_2(t) \sim M \cdot t - C_1(t)$, $S_1(t) = C \cdot t$, and $S_2(t) = (M - C) \cdot t$. Then,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{41}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{42}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1} L_1(t) = \lim_{t\to\infty} t^{-1} U_2(t) = \Lambda_2 \tag{43}$$

Proof Similar to Theorem 7, the dual of *system* 1 is a fluid queue, which we will call *system* 3, with cumulative fluid input $C_3(t) = S_1(t) = C \cdot t$ and cumulative available processing $S_3(t) = C_1(t)$. Now, let us define a third queue, which we will call *system* $2'$, with constant available processing such that $W_{2'}(t) = W_3(t)$. For this to hold, we must have $C_{2'}(t) = M \cdot t - S_3(t) = M \cdot t - C_1(t)$ and $S_{2'}(t) = (M - C) \cdot t$. Based on $W_{2'}(t) = W_3(t)$, and equation (27), we obtain $\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_{2'}(t) \geq K - \beta\}$. Moreover, *systems* $2'$ and 2 are stochastically equivalent. Hence, we obtain equation (41). Moreover, since $C_{2'}(t) = M \cdot t - S_3(t)$, $S_{2'}(t) = (M - C) \cdot t$, and $C_3(t) = C \cdot t$, *systems* $2'$ and 3 overflow and underflow at equal instantaneous rates. Thus, $L_{2'}(t) = L_3(t)$ and $U_{2'}(t) = U_3(t)$. By substituting this and the fact that *systems* $2'$ and 2 are stochastically equivalent into equation (30), we obtain equations (42) and (43).

□

This result can be particularly helpful in queues that have a fluid input with a finite number of states (the state of the input can be dependent on an external chain) and constant input rate during each state.

4 Numerical Results

In this section, we present numerical results for a few finite buffer models, and investigate the tightness of the bounds that have been obtained.

4.1 Point Process Model

We consider four different NBU random variables for the inter-arrival and service times of the finite buffer queues. IN particular we consider exponential, deterministic,uniform and bimodal random variables. The first three are NBU while bimodal random variable is not NBU for all values of its CV[4]. Hence, its parameters are chosen such that they ensure that the random variable is NBU. Tables 1 and 2 present the numerical results for the queue length distribution seen at departure instances of a finite buffer queue with a buffer size of 10 and Table 3 exhibits the numerical results for the stationary queue length distribution seen at an arbitrary time for a queue of the same buffer size.
Since the inter-arrival and service times of all queues are NBU, queue length distributions of the primal and dual FIFO queues must satisfy equations (24) and (25). Since the values in the second columns of the tables are less than or equal to the values in the corresponding third columns, the aforementioned equations hold true. Moreover, the numerical results show that the bounds are tight.
One of the primary uses of the results presented in section 2 is bounding quantities such as $\pi(0) = \mathbb{P}\{Q = 0\}$ and $\pi_D(0) = \mathbb{P}_D\{Q = 0\}$, that characterize the starvation of finite buffer queues, by more well-known quantities of its dual. According to Theorem 4, we can find upper or lower bounds for $\pi(0)$ and $\pi_D(0)$ of some finite buffer queueing systems using $\pi(K)$ and $\pi_D(K-1)$ of their duals. Table 4 provides some numerical results for $\pi(0)$ of a $D/M/1/K$ and $\pi(K)$ of its dual, $M/D/1/K$, where $K = 10$. Let us call the $D/M/1/K$ and $M/D/1/K$ queueing systems *systems* 1 and 2, respectively. From to the PASTA property, $\pi_2(K) = \mathbb{P}_A\{Q_2 = K\} = P_{L_2}$. Therefore, since numerous results for computing the probability of loss of finite buffer queues especially in the case of queues with Poisson arrivals are already well known in the literature, we can find tight bounds for $\pi_1(0)$. Similarly, we can obtain a lower bound for $\pi(0)$ of a queueing system with NWU inter-arrival times and exponential service times using the probability of loss of its dual.

4.2 Fluid Flow Model

Now, let us present the numerical results for a finite buffer fluid flow queue. Figure 1 depicts the normalized fluid Palm measure of the remaining workload associated with the input in a finite buffer fluid queue with ON-OFF input

[4] coefficient of variation

Table 1 Queue length distribution seen by departures in the primal and the dual queue - U/B (U and B stand for uniform and bipolar distributions) - with an accuracy of 10^{-6}

$U/B/1/10$, $\rho = 0.3$		$B/U/1/10$, $\rho = 3.33$	
$\mathbb{P}\{Q_D^1(t) \leq 0\}$	0.832274	0.88759	$\mathbb{P}\{Q_D^2(t) \geq 9\}$
$\mathbb{P}\{Q_D^1(t) \leq 1\}$	0.981835	0.989923	$\mathbb{P}\{Q_D^2(t) \geq 8\}$
$\mathbb{P}\{Q_D^1(t) \leq 2\}$	0.998573	0.999213	$\mathbb{P}\{Q_D^2(t) \geq 7\}$
$\mathbb{P}\{Q_D^1(t) \leq 3\}$	0.999875	0.999944	$\mathbb{P}\{Q_D^2(t) \geq 6\}$
$\mathbb{P}\{Q_D^1(t) \leq 4\}$	0.999991	0.999997	$\mathbb{P}\{Q_D^2(t) \geq 5\}$
$\mathbb{P}\{Q_D^1(t) \leq 5\}$	0.999995	0.999998	$\mathbb{P}\{Q_D^2(t) \geq 4\}$
$\mathbb{P}\{Q_D^1(t) \leq 6\}$	1	1	$\mathbb{P}\{Q_D^2(t) \geq 3\}$
$\mathbb{P}\{Q_D^1(t) \leq 7\}$	1	1	$\mathbb{P}\{Q_D^2(t) \geq 2\}$
$\mathbb{P}\{Q_D^1(t) \leq 8\}$	1	1	$\mathbb{P}\{Q_D^2(t) \geq 1\}$
$\mathbb{P}\{Q_D^1(t) \leq 9\}$	1	1	$\mathbb{P}\{Q_D^2(t) \geq 0\}$

Table 2 Queue length distribution seen by departures in the primal and the dual queue (M/D) with an accuracy of 10^{-6}

$D/M/1/10$, $\rho = 0.769$		$M/D/1/10$, $\rho = 1.3$	
$\mathbb{P}\{Q_D^1(t) \leq 0\}$	0.001862	0.002845	$\mathbb{P}\{Q_D^2(t) \geq 9\}$
$\mathbb{P}\{Q_D^1(t) \leq 1\}$	0.006775	0.007920	$\mathbb{P}\{Q_D^2(t) \geq 8\}$
$\mathbb{P}\{Q_D^1(t) \leq 2\}$	0.015696	0.017263	$\mathbb{P}\{Q_D^2(t) \geq 7\}$
$\mathbb{P}\{Q_D^1(t) \leq 3\}$	0.030897	0.032487	$\mathbb{P}\{Q_D^2(t) \geq 6\}$
$\mathbb{P}\{Q_D^1(t) \leq 4\}$	0.057794	0.058865	$\mathbb{P}\{Q_D^2(t) \geq 5\}$
$\mathbb{P}\{Q_D^1(t) \leq 5\}$	0.104764	0.105497	$\mathbb{P}\{Q_D^2(t) \geq 4\}$
$\mathbb{P}\{Q_D^1(t) \leq 6\}$	0.186994	0.187076	$\mathbb{P}\{Q_D^2(t) \geq 3\}$
$\mathbb{P}\{Q_D^1(t) \leq 7\}$	0.328818	0.329275	$\mathbb{P}\{Q_D^2(t) \geq 2\}$
$\mathbb{P}\{Q_D^1(t) \leq 8\}$	0.574113	0.574638	$\mathbb{P}\{Q_D^2(t) \geq 1\}$
$\mathbb{P}\{Q_D^1(t) \leq 9\}$	1	1	$\mathbb{P}\{Q_D^2(t) \geq 0\}$

Table 3 Queue length distribution seen at an arbitrary time in the primal and the dual queue (U/D) with an accuracy of 10^{-6}

$U/D/1/10$, $\rho = 0.8$		$D/U/1/10$, $\rho = 1.25$	
$\mathbb{P}\{Q^1(t) \leq 0\}$	0.199828	0.321988	$\mathbb{P}\{Q^2(t) \geq 10\}$
$\mathbb{P}\{Q^1(t) \leq 1\}$	0.720737	0.844115	$\mathbb{P}\{Q^2(t) \geq 9\}$
$\mathbb{P}\{Q^1(t) \leq 2\}$	0.961449	0.97949	$\mathbb{P}\{Q^2(t) \geq 8\}$
$\mathbb{P}\{Q^1(t) \leq 3\}$	0.994831	0.997313	$\mathbb{P}\{Q^2(t) \geq 7\}$
$\mathbb{P}\{Q^1(t) \leq 4\}$	0.999273	0.999649	$\mathbb{P}\{Q^2(t) \geq 6\}$
$\mathbb{P}\{Q^1(t) \leq 5\}$	0.999902	0.999954	$\mathbb{P}\{Q^2(t) \geq 5\}$
$\mathbb{P}\{Q^1(t) \leq 6\}$	0.999992	0.999994	$\mathbb{P}\{Q^2(t) \geq 4\}$
$\mathbb{P}\{Q^1(t) \leq 7\}$	0.999999	1	$\mathbb{P}\{Q^2(t) \geq 3\}$
$\mathbb{P}\{Q^1(t) \leq 8\}$	1	1	$\mathbb{P}\{Q^2(t) \geq 2\}$
$\mathbb{P}\{Q^1(t) \leq 9\}$	1	1	$\mathbb{P}\{Q^2(t) \geq 1\}$
$\mathbb{P}\{Q^1(t) \leq 10\}$	1	1	$\mathbb{P}\{Q^2(t) \geq 0\}$

Table 4 Bounding $\pi^1(0)$ with $\pi^2(K)$, where S1 is $D/M/1/K$ and S2 is $M/D/1/K$ ($K = 10$)

ρ	0.1	0.5	0.9	1.3	1.5
$\pi^1(0) = \mathbb{P}\{Q^1(t) = 0\}$	0.89977	0.49989	0.11269	0.00162	0.00018
$\pi^2(K) = \mathbb{P}\{Q^2(t) = K\}$	0.90021	0.50008	0.11353	0.00210	0.00027

and constant available processing alongside the counterpart of its dual.[5] The blue curve is the cumulative fluid Palm measure of the primal queue and the red curve is the tail of the fluid Palm measure of its dual. As a direct result of equation (29), the blue curve falls below the red one. Hence, one can be used to obtain an upper or lower bound on the other. Moreover, since the fluid Palm measure associated with the output equals the fluid Palm measure associated with the input that enters the queue, the only difference between the fluid Palm measure associated with the output and the the fluid Palm measure associated with the input is the portion of the input that is lost. This portion is usually very smaller in comparison to the overall input. Hence, the fluid Palm measure of the remaining workload associated with the input in the primal queue should be very close to the tail of the fluid Palm measure of the remaining workload associated with the input in the dual queue. This is depicted in Figure 1.

5 Conclusion

In this paper we have studied duality relationships for finite $G1/G2/1/K$ queues.The queue length duality relationship of queueing systems with controlled arrival and service processes was extended to FIFO queueing systems with NBU and NWU inter-arrival and service times. These results were used to obtain bounds for the queue length distribution seen at arbitrary times and at departure instances. Such results can be utilized in many applications including multimedia streaming and energy systems. For instance, Theorem 4 suggests that in queueing systems with NBU [NWU] inter-arrival and service times, $\pi_1(0) \leq [\geq]\pi_2(K)$ and $\pi_{D_1}(0) \leq [\geq]\pi_{D2}(K-1)$. In general, we cannot relate the probability of starvation, $P_S = \pi_D(0)$, of the primal system to the probability of loss, $P_L = \pi_A(K)$, of its dual. But, in the special case of queueing systems with Poisson arrival process and NWU service times, $P_{S_1} \geq \pi_A(0) = \pi_1(0) \geq \pi_2(K) = P_{L_2}$. Hence, one can use the results on probability of loss to compute an upper bound on the probability of starvation. Similarly, if one is interested in finding $\pi(0)$ of a finite buffer queueing system with NBU or NWU inter-arrival and service times, one can use the results on $\pi(K)$ to find an upper or lower bound on the quantity of interest.
The remaining workload distribution and the fluid Palm measure of the remaining workload associated with the input and output of the primal queue

[5] ON and OFF period average lengths in the primal queues are respectively 1.101 and 1.15 seconds. The input rate in the ON period is 50 bps and available processing is 20 bps.

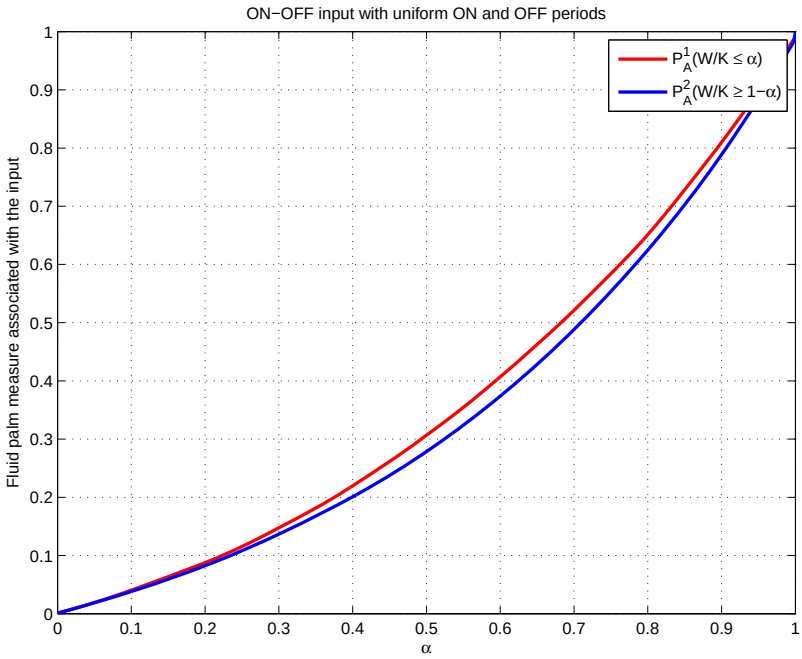

Fig. 1 Workload distribution of the primal and dual fluid queues with an ON-OFF input

were related to their counterparts in the dual queue in Theorem 6. It was shown that a duality holds between the overflow and underflow rates for fluid flow queues, which was specialized for fluid queues with ON-OFF inputs.

References

1. Ardakanian, O, Keshav, S., Rosenberg, C.P.: On the use of teletraffic theory in power distribution systems. In: Proceedings of E-Energy, May 2012, pp. 1-10 (2012)
2. Asmussen, S., Albrecher, H.: Ruin Probabilities. 2nd edn. Advanced Series on Statistical Science & Applied Probability 14, Hackensack, NJ (2010)
3. Baccelli, F., Brémaud, P.: Elements of Queueing Theory, Palm Martingale Calculus and Stochastic Recurrences. 2nd edn. Applications of Mathematics, Springer-Verlag, Berlin (2003)
4. Prabhu, N.U.: Queues and Inventories, a Study of Their Base Stochastic Processes. John Wiley & Sons, New York (1965)
5. Kleinrock, L.: Queueing Systems, Volume 1. John Wiley & Sons (1975)
6. Hlynka, M., Wang, T.: Comments on duality of queues with finite buffer size. Operation Research Letters 14, pp. 29-33 (1993)
7. Karaesmen, F., Gupta, S.M.: Duality relations for queues with arrival and service control. Computer & Operatoins Research, Vol. 24, No. 6, pp. 529-538 (1997)
8. Gouweleeuw, F.N.: The loss probability in an overloaded queue using the dual queue, Operations Research Letters 21, pp. 101-106 (1997)
9. Berger, A.W., Whitt, W.: Comparisons of multi-server queues with finite waiting rooms. Communications in Statistics. Stochastic Models, 8:4, pp. 719-732 (1992)
10. Sonderman, D.: Comparing multi-server queues with finite waiting rooms II: different numbers of servers. Advances in Applied Probability, Vol. 11, No. 2, pp. 448-455 (1979)

11. Hock, N.C.: Queueing Modeling Fundamentals. John Wiley & Sons Ltd (1996)
12. Mazumdar, R.R.: Performance Modelling, Loss Networks, and Statistical Multiplexing. Morgan & Claypool Publishers (2009)
13. Kim, N.K., Chae, K.C.: Transform-free analysis of the $GI/G/1/K$ queue through the decomposed Littles formula. Computers & Operations Research 30, pp. 353365 (2003)
14. Skellam, J.G., Shenton, L.R.: Distribution associated with random walk and recurrent events. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 19, No. 1, pp. 64-118 (1957)
15. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer Series in Statistics (2007)
16. Wolff, R.W.: Poisson arrivals see time averages. Operations Research, Vol. 30, No. 2, pp. 223-231 (1982)
17. Rolski, T., Schmidli, H., Schmidt, V., Teugels, J.: Stochastic Processes for Insurance and Finance. Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester (1999)
18. Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues. Springer Series in Operations Research (2001)
19. Xu, Y., Altman, E., El-Azouzi, R., Elayoubi, S., Haddad, M., Jimenez, T.: Probabilistic analysis of buffer starvation in markovian queues. In: Proceedings of IEEE Infocom, pp. 1826-1834 (2012).