Duality relations in finite queueing models

by

Nasser Barjesteh

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Motivated by applications in multimedia streaming and in energy systems, we study duality relations in finite queues. Dual of a queue is defined to be a queue in which the arrival and service processes are interchanged. In other words, dual of the $G_1/G_2/1/K$ queue is the $G_2/G_1/1/K$ queue, a queue in which the inter-arrival times have the same distribution as the service times of the primal queue and vice versa. Similarly, dual of a fluid flow queue with cumulative input $C(t)$ and available processing $S(t)$ is a fluid queue with cumulative input $S(t)$ and available processing $C(t)$. We are primarily interested in finding relations between the overflow and underflow of the primal and dual queues. Then, using existing results in the literature regarding the probability of loss and the stationary probability of queue being full, we can obtain estimates on the probability of starvation and the probability of the queue being empty. The probability of starvation corresponds to the probability that a queue becomes empty, i.e., the end of a busy period.

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 the queue being empty and the probability of the 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 that leads to an equality between the probability of starvation in the primal queue and the probability of overflow in the dual queue. The techniques are based on monotonicity arguments and coupling. The usefulness of the bounds is illustrated via numerical results.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Catherine Rosenberg and co-supervisor Prof. Ravi Mazumdar for their patience, motivation, immense knowledge in networking and queueing theory, and continuous support during my M.A.Sc. study and research. I could not have imagined having better supervisors and mentors for my graduate studies.

My sincere thanks also goes my fellow labmates: Jagadish Ghimire, Syamantak Datta Gupta, Dr. Sharad Birmiwal, Dr. Raja Sappidi, Arpan Mukhopadhyay, Dariush Fooladivanda, and Sajjad Mosharraf Dehkordi for the stimulating discussions, their helpful advice, and all the fun we have had in the last two years.

Last but not least, I would like to thank my parents for their understanding and endless love through the duration of my studies and giving me the strength to fulfill my dreams. My younger brother, Moein, and my dear friends deserve my whole-hearted thanks, as well.

This thesis is only the beginning of my journey.

Dedication

This thesis is dedicated to my parents and brother who have supported me all the way since the beginning of my studies.

Also, this thesis is dedicated to all those who believe in the richness of learning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How much time did you waste waiting in line, this week? It seems we can not escape frequent delays and they are getting worse. We encounter waiting lines in many industrial problems, as well. In communication systems, voice and data packets queue up in switches and routers for transmission. In manufacturing systems, products wait in line to get service from facilities. In health care, patients wait in line to see a doctor, get a bed, or receive medical services. Queueing theory studies the phenomena of standing, waiting, and serving and provides useful tools for performance evaluation of waiting lines. Moreover, it has been demonstrated that many problems in actuarial sciences and mathematical finance (particularly in risk analysis) can be translated to similar problems in queueing theory. Thus, over the years, many researchers have devoted their time to studying the behavior of queues. The alteration in queueing policy in banks is a simple example of how useful queueing analysis can be in everyday life. A few decades ago, each teller had a separate line. Queueing analysts investigated the effect of combining the queues and realized that the mean waiting time of customers can be significantly decreased by combining the lines. Thus, banks started using centralized numbering systems.

Queueing theory as a discipline is almost a century old. It was developed by Erlang, even though Johannsen seems to be the first person to publish

a paper in this field. Johannsen's paper [19] did not have a mathematical approach to the problem. Thus, in 1908, Johannsen asked Erlang to treat the problem from a mathematical point of view and Copenhagen Telephone Company appointed Erlang as a research associate on telephone problems. In the beginning, Erlang had no laboratory staff to help him. So, he had to carry out all the measurements on his own. He was often seen in the streets of Copenhagen, accompanied by a workman carrying a ladder, which was used to climb down into manholes. In 1909, Erlang laid the foundation for the place of Poisson distribution in queueing theory by showing that telephone calls arrive based on a Poisson process in [10]. His papers written during the next 20 years contain some of the most significant concepts and techniques; the notion of statistical equilibrium and the method of writing state balance equations are two such examples. In [12], he tackled the first optimization problem in queueing theory and in 1917, he wrote his most important paper on the subject [11]. As a result of the growing interest in his work, many of his papers were translated into English, French and German. He wrote his papers quite briefly, sometimes omitting the proofs, which made the papers difficult to understand especially for non-specialists in this field. It is known that a researcher from Bell Labs (in the United states) learned Danish, so he could read Erlang's papers in the original language. Interest in his work continued after his death and by 1944 Erlang was used in Scandinavian countries for denoting the unit of traffic. International recognition followed at the end of World War II [1]. Researchers continued working on this topic and numerous results followed [6, 27, 22, 18, 37].

Electrical engineers study queueing models and investigate the impact of different disciplines and distributions on the quantities of interest in order to increase the efficiency of communication systems and decrease expenses. In this thesis, motivated by applications in multimedia streaming and energy systems, we consider duality in finite buffer queueing systems and study the relations between arrival Palm distribution, departure Palm distribution, and stationary distribution of a finite buffer queue with the corresponding distributions in the dual queue.

1.1 Motivation

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, 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, [4] relates the starvation of storage units in energy systems to the starvation of finite buffer queues. Similarly, in [39], 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 in a single server queue (point process model), one can relate the probability of the queue being empty to the prob-

ability of loss 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 instants, 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 [35, 5] 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.

1.2 Mathematical description

Let us begin by describing the problem and defining the quantities of interest in our analysis. In this thesis, we consider classical queueing models with both discrete point process arrivals and fluid queues with continuous arrivals of work. The primal queue is assumed to be a $G_1/G_2/1/K$ and will be called *system* 1 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 is denoted by *system* 2. Let $\{T_A[n]\}$ denote the arrival times and $\{T_D[n]\}$ the departure times, where n indicates the n-th arrival or departure. Let us denote the queue length at time t by $Q(t)$, and the queue length at arrival and departure instants by $Q_A[n] \triangleq Q(T_A[n]^-)$ and $Q_D[n] \triangleq Q_{D^+}[n] \triangleq Q(T_D[n]^+)$ respectively. Let $\pi(\cdot) \triangleq \mathbb{P}(Q = \cdot)$ 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, $\pi_A(n) \triangleq \mathbb{P}_A\{Q_{0-} = n\} \triangleq \mathbb{P}(Q_{0-} = n|\Delta A_0 = 1)$, and $\pi_D(n) \triangleq \pi_{D^-}(n+1) \triangleq \mathbb{P}(Q_{0-} = n+1|\Delta D_0 = 1)$ denote the Palm probabilities associated with the arrivals and departures. In this thesis, we will use $\pi_D(\cdot)$ to denote $\pi_{D+}(\cdot)$, 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 [6, 23].

$$\pi_D(n) = \frac{\pi_A(n)}{1 - \pi_A(K)} \tag{1.2.1}$$

$$\lambda_n \pi(n) = \lambda_A \pi_A(n) \tag{1.2.2}$$

$$\mu_n \pi(n) = \lambda_D \pi_D(n-1) \tag{1.2.3}$$

$$\lambda_n \pi(n) = \mu_{n+1} \pi(n+1) \tag{1.2.4}$$

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 are difficult to calculate for general input and service distributions. Equation (1.2.1) follows from an argument similar to the one proposed in [23, p.40] for infinite buffer queues. Note equation (1.2.1) holds for all work-conservative disciplines.

For the fluid flow model, we will use a notation similar to [37] 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, $S2$.

1.3 Literature survey

The concept of duality in queues was first proposed by Prabhu in [27] and later discussed by Kleinrock in [22]. In [22] 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.

In [16], Heathcote exhibits a relationship between the limit distribution of the times between points of increase of the maximum queue length process in the $GI/M/1$ queueing system and the distribution of the busy period of the dual $M/G/1$ queueing system. [29] generalizes the result proposed in [16] for phase-type queueing systems, $GI/PH/1$ and $PH/G/1$, so it can be used for well-known distributions such as Erlang and hyper-exponential. The main result of [29] implies in particular that in an unstable $GI/M/1$ queueing system the maximum queue length process grows approximately like the counting process of an appropriately defined Markov renewal process.

In [36], Takacs proposes an alternate approach to queueing problems and uses the classical ballot theorem to obtain direct probabilistic and combinatorial derivations of numerous classical results of queueing theory, many of which had been previously obtained by the use of Laplace transform in his previous works. In [24], Niu continues in this spirit. He further generalizes the ballot theorem used in [36] and provides generalizations of several well-known formulas relating to busy periods and waiting times in $M/G/1$ and $GI/M/1$ queues. In particular, a duality relation between the joint distribution of several variables associated with the busy period in $M/G/1$ and the corresponding joint distribution in $GI/M/1$ is presented. In [25], Niu generalizes this duality relation to $GI/G/1$ queues with exceptional/modified first services. [25] clarifies the original result presented in [24] and proves that the generalized ballot theorem is not necessary for obtaining duality relations.

[13] obtain relations between the joint distribution of several variables in a $GI/G/1$ queue and the joint distribution of variables associated with the busy period in the dual queue. It is assumed that the primal queue has the preemptive-resume last-in-first-out (LIFO, aka last-come-first-served) discipline while the dual queue may have any work conserving queueing discipline. These relations generalize the results obtained by Niu in [24, 25] for $M/G/1$ and $GI/G/1$ queueing systems.

Ramaswami in [28] defines a double transform $R(z, s)$, which is a generalization of Neuts' rate matrix R, for Markov renewal processes of $GI/M/1$ type and writes $R(z, s)$ in terms of the double transform $G(z, s)$ of Neuts for a properly defined dual process of $M/G/1$ type. For the $GI/PH/1$ queue, the $G(z, s)$ matrix of the duality theorem is identified as the matrix of the dual $PH/GI/1$ queue with a time reversed representation of the PH-distribution. In [30], Ramaswami proposes a useful connection between Markov-modulated fluid flows (MMFF) and quasi-birth-and-death (QBD) models. MMFF models are real-valued stochastic processes with piece-wise linear sample paths whose instantaneous rate of change depend on the state of an associated continuous-time Markov Chain (CTMC) called the phase process. QBD

models are discrete state space processes on the integers modulated by a Markov chain and make jumps of at most one unit up or down in each step. This result has been used by Ramaswami and other in a series of papers. [2] proposes one further similarity between MMFFs and QBD processes by proposing a set of duality results using time reversal.

[7] is a good example of how one can benefit from duality theorems in deriving some interesting results pertaining to workload, queue length, and busy period of queues. [7] derives the stationary distribution of the workload of the server, or the virtual waiting time of $G/M/1$ queue with patience time K. It first obtains the expected number of downcrossings of a level in the workload process during a busy period and then the expected length of a busy period. The dual property between the $M/G/1$ queueing systems and the $G/M/1$ queueing systems is utilized to derive the expectations. By applying the level crossing argument to the expectations, the stationary density of the workload is derived.

[9] considers the queueing system $GI/G/l$ with impatient customers, meaning that customers will depart after a limited amount of time K if their service has not yet been completed, and shows that the waiting time of the nth customer is the dual of K minus the waiting time of the nth customer in the dual system.

[9] considers queues with a maximum of C waiting spaces and uses duality principles to relate the time of the first overflow of the $M/G/1$ and $G/M/1$ queues. The time is measured in terms of the number of arriving or departing customers before the overflow.

[32] establishes a relation between the queue sizes of a queuing system with Poisson arrivals, batch services with general service times, and its dual $G/M/1$ queue with batch arrivals and finite waiting space. Moreover, it proves that the knowledge of this relation is useful in the simultaneous study of the busy periods of such infinite buffer queues.

[15] instead of trying to relate the queue length or busy period of the primal queue to the counterparts of its dual, introduces a dual dam. It claims that events defined for the original queue can be related to events for the dual dam and vice versa. It particularly emphasizes on results pertaining to a single busy period.

[26] considers the notion of duality in infinite buffer dams and proposes interesting results based on the construction of an auxiliary generalized mountain process. It considers an infinite buffer $G/M/1$ dam with a general release rule, and constructs a dual $M/G/1$ dam with state-dependent jumps and no dry periods, whose workload process has the same stationary distribution up to some transformation. The stationary distribution of the dual dam can be computed in closed form.

[21] studies duality in the context of multiserver queues with no waiting space ($M/G/s/s$ aka Erlang's loss system). This queueing system is particularly interesting because it plays an important role in modeling and designing telephone exchanges. [21] provides a duality relation between the joint distribution of several variables associated with busy servers in an $M/G/s/s$ queueing system and the corresponding joint distribution associated with idle input sources in a $GI/M/1/s/s$ queueing system. In particular, it proves that the joint distribution of the number of busy servers and the attained service times in the $M/G/s/s$ queueing system are equivalent to the joint distribution of the number of idle sources and the attained idle times in the dual $GI/M/1/s/s$ queueing system. The $GI/M/1/s/s$ queueing system is a queue in the following sense: We assume that the number of input sources is s and that each input source independently generates customers according to a renewal process when idle and sleeps otherwise. The distribution of inter-point times is given by H, the same as the service time distribution in the $M/G/s/s$ queue. Moreover, assume that only one server is available and that service times are independent exponentially distributed random variables with mean $\frac{1}{\lambda}$. A well-known application of the $GI/M/1/s/s$ queueing

system is the machine interference model where the group of input sources is a set of s machines, an arrival corresponds to a machine breakdown, and the server corresponds to a repair-man. Another application of this queueing system is in modeling multiple access systems such as time-sharing systems or multiple access communication channels.

Hlynka in [17] 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 [17] were used in [20] to obtain duality results for queue length distribution in queueing systems with arrival and service control and in [14] to compute the loss probability of an overloaded $GI/M/1/K$ queue.

1.4 Contribution of this work

In this work, 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 instants. These results were used to obtain bounds for the queue length distribution seen at arbitrary times and at departure instants. We did not not derive relationships between starvation and overflow for general arrival and service processes. But, our results show that in special cases, we can bound the probability of starvation, $\pi_D(0)$, and the stationary probability of queue being empty, $\pi(0)$, with the probability of loss, $\pi_A(K)$, and the stationary probability of queue being full, $\pi(K)$. Then, the existing results in the literature concerning the probability of loss and the stationary probability of queue being full can be used to obtain bounds on the parameters of interest. Such results can be utilized in many applications including multimedia streaming and energy systems.

Similar results are provided for fluid flow queues with the exception that the duality relations for point process inputs are given by inequalities, while we will show that the remaining workload distribution of a fluid flow queue can be related to the remaining workload distribution of its dual via an equality. This is simply because fluid flow queues are limits of point process queueing systems when the service time approaches zero and increase in arrival rate compensates for it.

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.

1.5 Outline of the report

The thesis is organized as follows. In chapter 2 we begin by defining the interrupted arrival and virtual service disciplines and recall the results of [20] 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 2.2, we consider the special case of $M/G/1/K$ queues. In chapter 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. Chapter 4 presents some numerical results for finite buffer queues that confirm our results as well as how good the bounds are. Chapter 5 provides a summary of the results.

Chapter 2

Duality between overflow and starvation in classical queueing models

2.1 $G/G/1/K$

As stated in [17], 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 is 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 until next departure from the moment a packet arrives at an empty queue is distributed as a service time, but the time until 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 one must consider queues with controlled arrival and services as suggested in [20]. In queues with controlled arrival, the time until next arrival from the moment a packet leaves the full queue has an inter-arrival time distribution as opposed to the residual inter-arrival time distribution encountered in queues with FIFO discipline. Thus, as stated in Theorem 2.1.1, one can claim that a duality relation exists between the queue

length distribution of the primal $G_1/G_2/1/K$ queue with controlled arrival discipline and the counterpart in the dual queue. This duality relation will prove very useful in investigating duality relations in FIFO queues. We now define two concepts introduced in [20] that are used in this paper.

Definition 2.1.1.

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 until 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 until next arrival from the moment a customer leaves the full queue has a residual inter-arrival time distribution.*

In [20, 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 [20, Lemma 1] it is shown that a similar relation holds between the queue length of the first queue seen at arrivals and the queue length of the dual at departure instants. In [20, 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 instants.

Here in Theorem 2.1.1 , we will restate the results in [20] and provide a new duality relation between the queue length of $G_1/G_2/1/K$ with vir-

tual service seen at departure instants and the queue length of its dual at departure times.

Theorem 2.1.1.

1. *Let $i1$ denote a $G_1/G_2/1/K$ queue with an interrupted arrival discipline, and $i2$ a $G_2/G_1/1/K$ queue with an interrupted arrival discipline. Then, the corresponding stationary and Palm distributions at departure times satisfy:*

$$\pi_{i1}(i) = \pi_{i2}(K-i)\ ,\ \forall 0 \leq i \leq K \tag{2.1.1}$$

and

$$\pi_{D_{i1}}(i) = \pi_{A_{i2}}(K-1-i) = \pi_{D_{i2}}(K-1-i),\ \forall 0 \leq i \leq K-1 \tag{2.1.2}$$

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 a virtual service discipline. Then,*

$$\pi_{v1}(i) = \pi_{v2}(K-i) \quad , \quad \forall 0 \leq i \leq K \tag{2.1.3}$$

$$\pi_{D_{v1}}(i) = \pi_{D_{v2}}(K-1-i) \quad , \quad \forall 0 \leq i \leq K-1 \tag{2.1.4}$$

Proof. Equation (2.1.1) and the left hand side equality of (2.1.2) are essentially the results presented in *Theorem 1* and *Remark 1* of [20] and do not need proof. Since queues with interrupted arrival discipline do not reject arrivals, similar to infinite buffer queues, $\pi_{A_i}(m) = \pi_{D_i}(m)$ for all m. Hence, we obtain the right hand side equality of equation (2.1.2).
Although equation (2.1.3) has been shown in [20], we provide a similar proof based on coupling alongside equation (2.1.4), 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 instants 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,

$$\begin{aligned}\pi_{D_{v1}}(i) = \pi_{A_{2'}}(K-1-i|\Delta Q = 1) &= \frac{\pi_{A_{2'}}(K-1-i)}{\sum_{j=1}^{K-1}\pi_{A_{2'}}(j)} \\ &= \frac{\pi_{A_{2'}}(K-1-i)}{1-\pi_{A_{2'}}(K)} = \frac{\pi_{A_{2'}}(K-1-i)}{1-P_L}\end{aligned} \tag{2.1.5}$$

and vice versa. Substituting equation (1.2.1), that holds for all work conserving disciplines, into equation (2.1.5) 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 $v2$ are stochastically equivalent. Hence, equations (2.1.3) and (2.1.4) hold for any distribution of G_1 and G_2. □

Remark 2.1.1. *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 2.1.1 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, we use the notation $X \leq_{st} [\geq_{st}] Y$ to denote stochastic dominance, i.e., random variable X is stochastically smaller [greater] than random variable Y if $\mathbb{P}\{X \geq \alpha\} \leq [\geq] \mathbb{P}\{Y \geq \alpha\}$, $\forall \alpha > 0$.

Theorem 2.1.2. *Let system f denote a $G_1/G_2/1/K$ queue with FIFO discipline and system i denote a $G_1/G_2/1/K$ queue with interrupted arrivals. Let $R_{A|Q_D^{K-1}}$ denote the time until 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 f.*

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

$$\begin{aligned} \mathbb{P}\{Q_{D_f}[n] \leq j\} &\leq \mathbb{P}\{Q_{D_i}[n] \leq j\} \quad , \quad \forall\, 0 \leq j \leq K-1 \quad (2.1.6) \\ \mathbb{P}\{Q_f(t) \leq j\} &\leq \mathbb{P}\{Q_i(t) \leq j\} \quad , \quad \forall\, 0 \leq j \leq K \quad (2.1.7) \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}\{Q_{D_f}(t) \leq j\} &\geq \mathbb{P}\{Q_{D_i}(t) \leq j\} \quad , \quad \forall\, 0 \leq j \leq K-1 \quad (2.1.8) \\ \mathbb{P}\{Q_f(t) \leq j\} &\geq \mathbb{P}\{Q_i(t) \leq j\} \quad , \quad \forall\, 0 \leq j \leq K \quad (2.1.9) \end{aligned}$$

Proof. Let us denote the counting process associated with the arrivals by $N_A(t)$, the distribution of its inter-point times with $G_A(x)$, the counting process associated with the departure process by $N_D(t)$, the distribution of the service times by $G_S(x)$, 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 forward recurrence time of the arrival process $N_A(t)$ (residual inter-arrival time) at time t by $R_A[t]$. Moreover, let us denote the stationary distribution of the continuous/discrete-time stochastic process X by π_X. Let us prove part 1. The argument is similar to the argument presented in [[34], Theorem 1]. We construct two new queueing

systems[1] on the same probability space such that $\widetilde{Q}_{D_f}[n] \geq \widetilde{Q}_{D_i}[n]$ for all n, $\pi_{\widetilde{Q}_{D_f}} = \pi_{Q_{D_f}}$, and $\pi_{\widetilde{Q}_{D_i}} = \pi_{Q_{D_i}}$. To do this, use any arrival and service processes for *system* f such that $\pi_{N_{A_f}} = \pi_{\widetilde{N}_{A_f}}$ and $\pi_{N_{D_f}} = \pi_{\widetilde{N}_{D_f}}$ and by construction, take $\widetilde{A}_i[n] = \widetilde{A}_f[n]$ if $\widetilde{Q}_{A'_f}[n] < K-1$, $\widetilde{A}_i[n] \geq \widetilde{A}_f[n]$ if $\widetilde{Q}_{A'_f}[n] = K-1$, and $\widetilde{S}_i[n] = \widetilde{S}_f[n]$ 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 $\widetilde{Q}_{D_f}[n] > \widetilde{Q}_{D_i}[n]$ or $\widetilde{Q}_{D_f}[n] = \widetilde{Q}_{D_i}[n]$ and $\widetilde{R}_{A_f}(\widetilde{T}_{D_f}[n]) \leq \widetilde{R}_{A_i}(\widetilde{T}_{D_i}[n])$. Let us use mathematical induction and show that if the above property holds for n, it holds for $n+1$, as well. Consider $t_k[n] = \widetilde{T}_{D_k}[n+1] - \widetilde{S}_k[n]$, $\forall\, k \in \{f, i\}$, which is equal to $\widetilde{T}_{D_k}[n]$ if and only if $\widetilde{Q}_{D_k}[n] > 0$. If $\widetilde{Q}_{D_k}[n] = 0$, $\widetilde{Q}_k(t_k[n]) = 1$ and $R_{A_k}(t_k[n]) = \widetilde{A}_k[n+2]$. It is straightforward to show that the above property holds at $t = t_k[n]$. Thus, for any $0 \leq t \leq \widetilde{S}_k[n+1]$ we have either $\widetilde{Q}_f(\widetilde{T}_{D_f}[n] + t) > \widetilde{Q}_i(\widetilde{T}_{D_i}[n] + t)$ or $\widetilde{Q}_f(\widetilde{T}_{D_f}[n] + t) = \widetilde{Q}_i(\widetilde{T}_{D_i}[n] + t)$ and $\widetilde{R}_{A_f}(\widetilde{T}_{D_f}[n] + t) \leq \widetilde{R}_{A_i}(\widetilde{T}_{D_i}[n] + t)$ because $\widetilde{A}_f[n] \leq \widetilde{A}_i[n]$ and $\widetilde{S}_f[n] = \widetilde{S}_i[n]$. Hence, the property holds for $n+1$, as well. Therefore, $\widetilde{Q}_{D_f}[n] \geq \widetilde{Q}_{D_i}[n]$ for all n. Thus, $\mathbb{1}\{\widetilde{Q}_{D_f}[n] \leq j\} \leq \mathbb{1}\{\widetilde{Q}_{D_i}[n] \leq j\}$. From the finiteness of the queues and the strong law of large numbers, we conclude that $\mathbb{P}\{\widetilde{Q}_{D_f}[n] \leq j\} \leq \mathbb{P}\{\widetilde{Q}_{D_i}[n] \leq j\}$; meaning that equation (2.1.6) holds. Furthermore for $m \geq 0$, we have

$$\int_{\widetilde{T}_{D_k}[n]}^{\widetilde{T}_{D_k}[n+1]} \mathbb{1}\{\widetilde{Q}_k(s) > m\} ds = \int_{\widetilde{T}_{D_k}[n]}^{t_k[n]} \mathbb{1}\{\widetilde{Q}_k(s) > m\} ds + \int_{t_k[n]}^{\widetilde{T}_{D_k}[n+1]} \mathbb{1}\{\widetilde{Q}_k(s) > m\} ds$$

Where $\int_{\widetilde{T}_{D_k}[n]}^{t_k[n]} \mathbb{1}\{\widetilde{Q}_k(s) > m\} ds = 0$, $\int_{\widetilde{T}_{D_f}[n]}^{t_f[n]} \mathbb{1}\{\widetilde{Q}_f(s) > m\} ds \geq 0$, and

[1]The constructed queueing systems will be distinguished from the primal ones by the use of a $\sim$ sign.

$\int_{t_i[n]}^{\widetilde{T}_{D_i}[n+1]} \mathbb{1}\{\widetilde{Q}_i(s) > m\}ds \leq \int_{t_f[n]}^{\widetilde{T}_{D_f}[n+1]} \mathbb{1}\{\widetilde{Q}_f(s) > m\}ds$. Therefore,

$$\int_{\widetilde{T}_{D_i}[n]}^{\widetilde{T}_{D_i}[n+1]} \mathbb{1}\{\widetilde{Q}_i(s) > m\}ds \leq \int_{\widetilde{T}_{D_f}[n]}^{\widetilde{T}_{D_f}[n+1]} \mathbb{1}\{\widetilde{Q}_f(s) > m\}ds, \ \forall n \geq 1$$

Hence,

$$\mathbb{E}_D[\int_0^{T_{D_i}} \mathbb{1}\{Q_i(s) > m\}ds] \leq \mathbb{E}_D[\int_0^{T_{D_f}} \mathbb{1}\{Q_f(s) > m\}ds]$$

Where T_{D_i} is the stationary inter-point time of Q_{D_i} and T_{D_f} is the stationary inter-point time of Q_{D_f}. Moreover, since $\widetilde{T}_{D_i}[n+1] - \widetilde{T}_{D_i}[n] \geq \widetilde{T}_{D_f}[n+1] - \widetilde{T}_{D_f}[n]$ for all n, $\mathbb{E}_D[T_{D_i}] \geq \mathbb{E}_D[T_{D_f}]$. Thus, $\lambda_{D_i} \leq \lambda_{D_f}$, where λ_{D_i} is the mean rate of Q_{D_i} and λ_{D_f} is the mean rate of Q_{D_f}. Using the Palm inversion formula introduced in [23], we obtain that

$$\mathbb{P}\{Q_i > m\} = \mathbb{E}[\mathbb{1}\{Q_i > m\}] = \lambda_{D_i} \cdot \mathbb{E}_D[\int_0^{T_{D_i}} \mathbb{1}\{Q_i(s) > m\}ds]$$
$$\leq \lambda_{D_f} \cdot \mathbb{E}_D[\int_0^{T_{D_f}} \mathbb{1}\{Q_f(s) > m\}ds] = \mathbb{E}[\mathbb{1}\{Q_f > m\}] = \mathbb{P}\{Q_f > m\}$$

Thus,

$$\mathbb{P}\{Q_i \leq m\} = 1 - \mathbb{P}\{Q_i > m\} \geq 1 - \mathbb{P}\{Q_f > m\} = \mathbb{P}\{Q_f \leq m\}$$

Hence, equation (2.1.7). 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 [3], 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 [22, 23, 33] 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 [31].

Definition 2.1.2.

1. *The random variable X is IFR (Increasing Failure Rate) 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 (Decreasing Failure Rate) if and only if, $[X - t|X \geq t] \leq_{st} [X - t_1|X \geq t_1]$ whenever $t \leq t_1$.* [2]

2. *The non-negative random variable X is NBU (New Better than Used) if and only if $X \geq_{st} [X - t|X \geq t]$ for all $t \geq 0$ and X is NWU (New Worse than Used) if and only if, $X \leq_{st} [X - t|X \geq t]$ for all $t \geq 0$.*

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

We will provide non-trivial examples of NBU random variables in Remark 2.1.2. Now, let us present the following result.

Theorem 2.1.3. *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.*

[2] The random variable X is IFR if and only if $1 - F(x)$, where $F(x)$ is the cululative distribution function of the random variable X, is logconcave. Similarly, X is DFR if and only if $1 - F(x)$ is logconvex.

Proof. Let function $F(\cdot)$ denote the cumulative distribution function, random variable B_A the backward recurrence times of the arrival process at the instants of interest, and random variable A the stationary inter-arrival times. Moreover, $F_{B_A}(x) = \mathbb{P}\{B_A \leq x\}$ will be used to denote the distribution of B_A. Then if $R_{A|Q_D^{K-1}} \leq_{st} A$,

$$\begin{aligned}
\mathbb{P}\{R_A \geq x\} &= \int_0^\infty dF\{R_A \geq x, B_A \leq \alpha\} \\
&= \int_0^\infty \mathbb{P}\{R_A \geq x | B_A = \alpha\} dF_{B_A}(\alpha) \\
&= \int_0^\infty \mathbb{P}\{A - B_A \geq x | A \geq \alpha\} dF_{B_A}(\alpha) \\
&\leq \int_0^\infty \mathbb{P}\{A \geq x\} dF_{B_A}(\alpha) \\
&= \mathbb{P}\{A \geq x\} \cdot \int_0^\infty dF_{B_A}(\alpha) = \mathbb{P}\{A \geq x\}
\end{aligned}$$

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 2.1.4. *Let us call $G_1/G_2/1/K$ as system 1 and the dual 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*

$$\begin{aligned}
\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) \ &\leq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\
= \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad &\forall\, 0 \leq i \leq K-1 \qquad (2.1.10)
\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) \ &\leq \sum_{j=K-i}^{K} \pi_2(j) \\ = \mathbb{P}\{Q_2(t) \geq K-i\}, \quad &\forall\, 0 \leq i \leq K \end{aligned} \tag{2.1.11}$$

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

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} = \sum_{j=1}^{i} \pi_{D_1}(j) \geq &\sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\ = \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad &\forall\, 0 \leq i \leq K-1 \end{aligned} \tag{2.1.12}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} = \sum_{j=1}^{i} \pi_1(j) \geq &\sum_{j=K-i}^{K} \pi_2(j) \\ = \mathbb{P}\{Q_2(t) \geq K-i\}, \quad &\forall\, 0 \leq i \leq K \end{aligned} \tag{2.1.13}$$

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 2.1.2 and 2.1.3 for the primal queue, we have

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} \quad &\leq \quad \mathbb{P}\{Q_{D_{1'}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \\ \mathbb{P}\{Q_1(t) \leq i\} \quad &\leq \quad \mathbb{P}\{Q_{1'}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \end{aligned}$$

Similarly, for the dual queue, we have

$$\begin{aligned}\mathbb{P}\{Q_{D_2}(t) \leq i\} \quad &\leq \quad \mathbb{P}\{Q_{D_{2'}}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \\ \mathbb{P}\{Q_2(t) \leq i\} \quad &\leq \quad \mathbb{P}\{Q_{2'}(t) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K, \end{aligned}$$

Moreover, Theorem 2.1.1 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 2.1.2 and 2.1.3 into the result of Theorem 2.1.1, concludes the proof. □

Corollary 2.1.1. *Theorem 2.1.4 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* $\mathbb{P}\{Q_1(t) \geq K-i\}$. *The same holds for the queue length distribution at departure instants.*

Corollary 2.1.2. *A special case of Theorem 2.1.4,* $i = 0$, *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.*

Remark 2.1.2. *Though the relations proposed in Theorem 2.1.4 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. Bimodal random variable is an example of a random variable that depending on the parameters of its distribution can be a NBU random variable or not. 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.*

2.2 $M/G/1/K$

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 2.1.2 for the special case of queues with Poisson arrivals.

Theorem 2.2.1. *Let us call the $M/G/1/K$ with FIFO discipline, system f and the $M/G/1/K$ with virtual service discipline, system v. Also, let us denote the service time of the first customer of each busy period in system v by $R_{S|Q_A^0}$ and let S be a random variable that is distributed as regular service times of system f. Then,*

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

$$\begin{aligned} \mathbb{P}\{Q_{D_f}[0] \leq i\} &\leq \mathbb{P}\{Q_{D_v}[0] \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \quad (2.2.14) \\ \mathbb{P}\{Q_f(0) \leq i\} &\leq \mathbb{P}\{Q_v(0) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \quad (2.2.15) \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}\{Q_{D_f}[0] \leq i\} &\geq \mathbb{P}\{Q_{D_v}[0] \leq i\} \quad , \quad \forall\, 0 \leq i \leq K-1 \quad (2.2.16) \\ \mathbb{P}\{Q_f(0) \leq i\} &\geq \mathbb{P}\{Q_v(0) \leq i\} \quad , \quad \forall\, 0 \leq i \leq K \quad (2.2.17) \end{aligned}$$

Due to PASTA property introduced in [38], a similar result holds for the queue length distribution seen at arrival instants.

Proof. We will prove part 1 and the proof of part 2 follows similarly. The argument is similar to Theorem 2.1.2 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 f* such that $\pi_{N_{A_f}} = \pi_{\widetilde{N}_{A_f}}$ and $\pi_{N_{D_f}} = \pi_{\widetilde{N}_{D_f}}$. We will propose a construction for *system v* that ensures that $\widetilde{Q}_{D_f}[n] \geq \widetilde{Q}_{D_v}[n]$ for all n. Let us use mathematical induction. We assume that $\widetilde{Q}_{D_f}[k] \geq \widetilde{Q}_{D_v}[k]$ for $k \leq m$

and we want to prove that $\widetilde{Q}_{D_f}[m+1] \geq \widetilde{Q}_{D_v}[m+1]$. If $\widetilde{Q}_{D_v}[m] > 0$, we choose the $(m+1)$-th service time of *system* v equal to the one in *system* f. Since arrivals are Poisson and exponential random variables are memoryless, we can construct the same arrivals for *system* v during the $(m+1)$-th service time. Hence, $\widetilde{Q}_{D_f}[m+1] \geq \widetilde{Q}_{D_v}[m+1]$. On the other hand, if $\widetilde{Q}_{D_v}[m] = 0$, 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* f for *system* v. Then, because exponential random variables are memoryless, the number of arrivals during the $(m+1)$-th service time of *system* v will be less than or equal to the number of arrivals in *system* f. Thus, $\widetilde{Q}_{D_f}[m+1] \geq \widetilde{Q}_{D_v}[m+1]$, meaning that equation 2.2.14 holds.
For proving equation 2.2.15, we will use a result proposed in *Theorem 1* of [8]. 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* f, there exists a sample path of *system* v that satisfies $A_{k_f} = A_{k_v}$. Moreover, *Theorem 1* of [8] suggests that for the above-mentioned systems, $B_{k_f} \geq B_{k_v}$. Hence, we have $P_{L_f} \geq P_{L_v}$. Therefore,

$$\begin{aligned} \mathbb{P}\{\widetilde{Q}_{A_f}(t) \leq i\} &= (1 - P_{L_f}) \cdot \mathbb{P}\{\widetilde{Q}_{D_f}(t) \leq i\} \\ &\leq (1 - P_{L_v}) \cdot \mathbb{P}\{\widetilde{Q}_{D_v}(t) \leq i\} = \mathbb{P}\{\widetilde{Q}_{A_v}(t) \leq i\} \end{aligned}$$

Using the PASTA property and the fact that the queue length distributions of the constructed and primal queues are the same, we obtain equation (2.2.15). □

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 2.1.4.

Corollary 2.2.1. *Let us call* $M/G/1/K$ *system* 1 *and its dual* $GI/M/1/K$,

system 2. *Then,*

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

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} &= \sum_{j=1}^{i} \pi_{D_1}(j) \leq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\ &= \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad \forall\, 0 \leq i \leq K-1 \end{aligned} \tag{2.2.18}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} &= \sum_{j=1}^{i} \pi_1(j) \ \leq \sum_{j=K-i}^{K} \pi_2(j) \\ &= \mathbb{P}\{Q_2(t) \geq K-i\}, \quad \forall\, 0 \leq i \leq K \end{aligned} \tag{2.2.19}$$

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

$$\begin{aligned}\mathbb{P}\{Q_{D_1}(t) \leq i\} &= \sum_{j=1}^{i} \pi_{D_1}(j) \ \geq \sum_{j=K-1-i}^{K-1} \pi_{D_2}(j) \\ &= \mathbb{P}\{Q_{D_2}(t) \geq K-1-i\}, \quad \forall\, 0 \leq i \leq K-1 \end{aligned} \tag{2.2.20}$$

and

$$\begin{aligned}\mathbb{P}\{Q_1(t) \leq i\} &= \sum_{j=1}^{i} \pi_1(j) \ \geq \sum_{j=K-i}^{K} \pi_2(j) \\ &= \mathbb{P}\{Q_2(t) \geq K-i\}, \quad \forall\, 0 \leq i \leq K \end{aligned} \tag{2.2.21}$$

Proof. We omit the proof since it is similar to Theorem 2.1.4. □

In this chapter, using the duality relations for queues with interrupted arrival and virtual service disciplines alongside stochastic orders, we found some new relations between the queue length distribution of a finite buffer FIFO queue 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.

Chapter 3

Fluid Flow Model

3.1 Preliminaries

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 [6, 23] for details.

Proposition 3.1.1. *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,*

$$\mathbb{P}_A\{W \leq x\} \leq \mathbb{P}_{A'}\{W \leq x\} = \mathbb{P}_D\{W \leq x\} \tag{3.1.1}$$

Proof. By changing equation 2.41 of [23] 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, $\lambda_{A'} = \lambda_D$. Thus,

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

The left hand side inequality follows from the definition of fluid Palm measure

in equation 1.22 of [23]. 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 3.1.1 resembles the relation between the queue length distribution at arrival and departure instants 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 instants, the term in the middle is the queue length distribution at the instants of arrivals that enter the queue, and the right-most term is the queue length distribution at departure instants.

3.2 General inputs and service processes

At this point, we can express the following duality results for general fluid flow queues.

Theorem 3.2.1. *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,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{3.2.2}$$

$$\mathbb{P}_D\{W_1(t) \leq \beta\} = \mathbb{P}_D\{W_2(t) \geq K - \beta\} \tag{3.2.3}$$

$$\mathbb{P}_A\{W_1(t) \leq \beta\} \leq \mathbb{P}_A\{W_2(t) \geq K - \beta\} \tag{3.2.4}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{3.2.5}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1} L_1(t) = \lim_{t\to\infty} t^{-1} U_2(t) = \Lambda_2 \tag{3.2.6}$$

Proof. Let us use the *role inversion* model introduced in [22] and later used in [17]. 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 the remaining workload of *systems* 1 and $1'$ 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 (3.2.2).
Note that when *system* 1 is full, the fluid input sees a full system with a rate equal to the instantaneous fluid input rate[1], 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 3.1.1 and the fact that *systems* 1 and 2 are stochastically equivalent, we obtain equation (3.2.3). Moreover, using equation (3.2.3) and the inequality of Theorem 3.1.1, we obtain equation (3.2.4).
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 equation (3.2.5). Equation (3.2.6) 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

[1] the rate is the derivative of C_t

fluid Palm measure associated with the output of its dual. Corollary 3.2.1 studies this relation.

Corollary 3.2.1. *For systems* 1 *and* 2 *defined in Theorem 3.2.1, we have*

$$\mathbb{P}_A\{W_1(t) \geq K - \beta\} \quad \geq \quad \mathbb{P}_D\{W_2(t) \leq \beta\} \tag{3.2.7}$$

Proof. Substituting the inequality proposed in Proposition 3.1.1 into equation (3.2.3) results in

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

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 $n \geq 1$. 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

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

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 the 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,

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

By interchanging β and $K - \beta$, we obtain equation (3.2.7). □

Remark 3.2.1. *In Theorem 2.1.4, duality relations for point process inputs are given by inequalities, while Theorem 3.2.1 relates the remaining workload distribution of a fluid queue to the remaining workload distribution of its dual via an equality.*

In this section, we provided a duality relation between the workload distribution of the primal fluid flow queue and the workload distribution of the dual queue and showed that the underflow probability of one can be related to the overflow probability of the other. Next, we will consider the special case of fluid queues with ON-OFF inputs and constant available processing.

3.3 ON-OFF fluid queues

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 in stationary queues are define as follows.

$$R_L \triangleq \lim_{t\to\infty} \frac{1}{C(t)} \int_0^t \mathbb{1}\{\omega(s) \in L\} d(C(s) - S(s)) = \lim_{t\to\infty} \frac{U(t)}{C(t)} \tag{3.3.8}$$

$$R_S \triangleq \lim_{t\to\infty} \frac{1}{S(t)} \int_0^t \mathbb{1}\{\omega(s) \in S\} d(S(s) - C(s)) = \lim_{t\to\infty} \frac{L(t)}{S(t)} \tag{3.3.9}$$

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 3.3.1. *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,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{3.3.10}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{3.3.11}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1} L_1(t) = \lim_{t\to\infty} t^{-1} U_2(t) = \Lambda_2 \tag{3.3.12}$$

$$R_{L_1} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \cdot R_{S_2} \tag{3.3.13}$$

Notice that a relation similar to equations (3.2.3) and (3.2.4) does not hold for the above queues.

Proof. From Theorem 3.2.1, the 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 (3.2.2), the relation (3.3.10) 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 (3.2.5), we obtain equations (3.3.11) and (3.3.12).
Based on the definition of *system* 2, we have

$$\frac{\mathbb{E}[S_2(t)]}{\mathbb{E}[S_3(t)]} = \frac{(H_1 - C_1) \cdot t}{H_1 \cdot \frac{\mathbb{E}[A_1]}{\mathbb{E}[A_1]+\mathbb{E}[B_1]} \cdot t} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \tag{3.3.14}$$

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

$$R_{S_3} = \frac{H_1 - C_1}{H_1} \cdot \frac{\mathbb{E}[A_1] + \mathbb{E}[B_1]}{\mathbb{E}[A_1]} \cdot R_{S_2} \tag{3.3.15}$$

Moreover, using equation (3.2.5) and the fact that $C_1(t) = S_3(t)$, because they are duals, we obtain $R_{L_1} = R_{S_3}$ and hence equation (3.3.13).

□

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 3.3.1. *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,*

$$\mathbb{P}\{W_1(t) \leq \beta\} = \mathbb{P}\{W_2(t) \geq K - \beta\} \tag{3.3.16}$$

$$\mathbb{P}\{U_1(t) \leq \alpha\} = \mathbb{P}\{L_2(t) \leq \alpha\} \tag{3.3.17}$$

$$\nu_1 = \lim_{t\to\infty} t^{-1} L_1(t) = \lim_{t\to\infty} t^{-1} U_2(t) = \Lambda_2 \tag{3.3.18}$$

Proof. Similar to Theorem 3.3.1, 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 (3.2.2), 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 (3.3.16). 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 (3.2.5), we obtain equations (3.3.17) and (3.3.18).

□

The results proposed in Corollary 3.3.1 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.

In this chapter, we presented a duality relation between the workload distribution of the primal fluid flow queue and the workload distribution of its dual. In section 3.3, we focused our attention on a very well-known subclass of such queues, namely queues with ON-OFF input and constant available processing, and presented a relationship between the rate of starvation of

the primal queue and the rate of overflow of its dual. Similar results were presented for queues with bounded fluid input processes.

Chapter 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 a bimodal random variable is not NBU for all values of its coefficient of variation. We will select its parameters such that the random variable is NBU. Tables 4.1 and 4.2 present the numerical results for the queue length distribution seen at departure instants of a finite buffer queue with a buffer size of 10 and Table 4.3 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 (2.2.20) and (2.2.21). 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 inequalities hold true. Moreover, the numerical results

Table 4.1: Queue length distribution seen by departures in the primal and the dual queue - U/B (U and B stand for uniform and bimodal 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	$\mathbb{P}\{Q_D^2(t) \geq 3\}$

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 2.1.4, 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 4.4 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.

Table 4.2: 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 4.3: 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	$\mathbb{P}\{Q^2(t) \geq 2\}$

Table 4.4: 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

4.2 Fluid Flow Model

Now, let us present the numerical results for a finite buffer fluid flow queue. Figure 4.1 depicts the normalized fluid Palm measure of the remaining workload associated with the input in a finite buffer fluid queue with ON-OFF input and constant available processing alongside the counterpart of its dual.[1] 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 (3.2.4), 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 fluid Palm measure associated with the input is the portion of the input that is lost. This portion is usually much 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 4.1.

[1] 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.

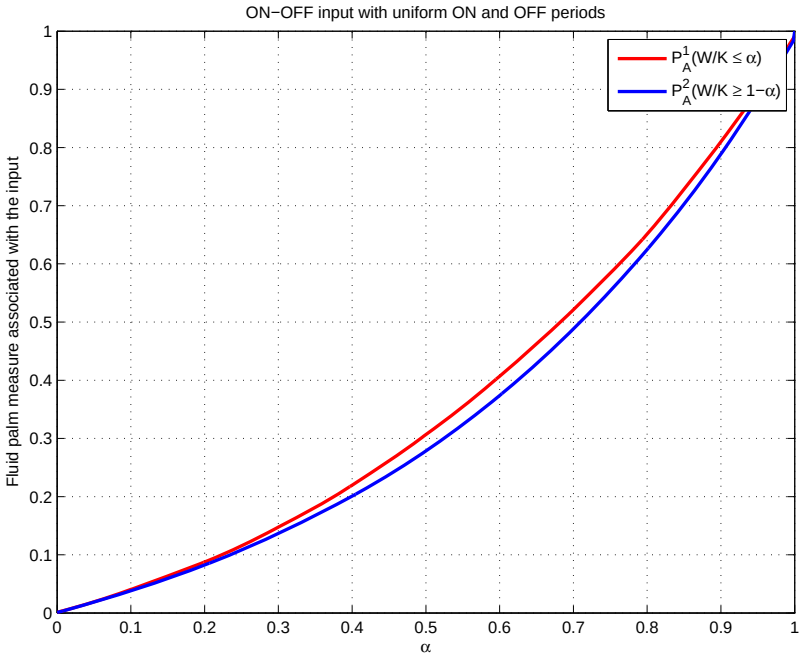

Figure 4.1: Workload distribution of the primal and dual fluid queues with an ON-OFF input

Chapter 5

Conclusion

In this work, 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 instants. Such results can be utilized in many applications including multimedia streaming and energy systems. For instance, Theorem 2.1.4 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 $\pi(0)$ for 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 were related to their counterparts in the dual queue in Theorem 3.2.1. 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.

Future work on this subject could investigate the notion of conditional residual life (for instance, the residual life of the arrival process given that a departure leaves i customers in the queue) to generalize the duality relations provided in this work to more general service time and arrival distributions.

Bibliography

[1] Le comité consultatif international des communications téléphoniques à grande distance. In *1946 Proceedings C.C.I.F.*

[2] Soohan Ahn and Vaidyanathan Ramaswami. Duality results for markov-modulated fluid flow models. *Journal of Applied Probability*, 48:309–318, 2011.

[3] N.K. Kim ans K.C. Chae. Transform-free analysis of the GI/G/1/K queue through the decomposed littles formula. *Computers & Operations Research*, 30:353365, 2003.

[4] O. Ardakanian, S. Keshav, and C. Rosenberg. On the use of teletraffic theory in power distribution systems. In *Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy), 2012 Third International Conference on*, pages 1–10, 2012.

[5] S. Asmussen and H. Albrecher. *Ruin Probabilities*. Advanced Series on Statistical Science & Applied Probability 14, 2 edition, 2010.

[6] Francois Baccelli and Pierre Brémaud. *Elements of Queueing Theory, Palm Martingale Calculus and Stochastic Recurrences*. Applications of Mathematics, Springer-Verlag, 2 edition, 2003.

[7] Jongho Bae and Sunggon Kim. The stationary workload of the G/M/1 queue with impatient customers. *Queueing systems*, 64(3):253–265, 2010.

[8] A.W. Berger and W. Whitt. Comparisons of multi-server queues with finite waiting rooms. *Communications in Statistics. Stochastic Models*, 8(4):719–732, 1992.

[9] U.N. Bhat. Customer overflow in queues with finite waiting spaces. *Australian Journal of Statistics*, 7:15–19, 1965.

[10] A.K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20(6):87–98, 1909.

[11] A.K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13:5–13, 1917.

[12] A.K. Erlang. On the rational determination of the number of circuits. *The life and works of A.K. Erlang*, pages 216–221, 1948.

[13] Demetrios Fakinos. Duality relations for certain single server queues. *Queueing Systems*, 4(1):77–83, 1989.

[14] F.N. Gouweleeuw. The loss probability in an overloaded queue using the dual queue. *Operations Research Letters*, 21:101–106, 1997.

[15] Irwin Greenberg. Some duality results in the theory of queues. *Journal of Applied Probability*, pages 99–121, 1969.

[16] C.R. Heathcote. On the maximum of the queue GI/M/1. *Journal of Applied Probability*, 2(1):206–214, 1965.

[17] M. Hlynka and T. Wang. Comments on duality of queues with finite buffer size. *Operation Research Letters*, 14:29–33, 1993.

[18] N.C. Hock. *Queueing Modeling Fundamentals*. John Wiley & Sons, 1996.

[19] Johannsen. Waiting times and number of calls. *Post Office Electrical Engineers Journal*, 1910.

[20] F. Karaesmen and S.M. Gupta. Duality relations for queues with arrival and service control. *Computer & Operatoins Research*, 24(6), 529-538 1997.

[21] Toshikazu Kimura. Duality between the M/G/s/s and GI/M/s/s queues. *Computers & Mathematics with Applications*, 24(1):35–40, 1992.

[22] Leonard Kleinrock. *Theory, volume 1, Queueing systems.* Wiley-interscience, 1975.

[23] R.R. Mazumdar. *Performance Modelling, Loss Networks, and Statistical Multiplexing.* Morgan & Claypool Publishers, 2009.

[24] Shun-Chen Niu and Robert B. Cooper. Duality and other results for M/G/1 and GI/M/1 queues, via a new ballot theorem. *Mathematics of Operations Research*, 14(2):281–293, 1989.

[25] Shun-Chen Niu and Robert B. Cooper. A duality relation for busy cycles in GI/G/1 queues. *Queueing systems*, 8(1):203–209, 1991.

[26] David Perry and Wolfgang Stadje. Duality of dams via mountain processes. *Operations Research Letters*, 31(6):451–458, 2003.

[27] N.U. Prabhu. *Queues and Inventories, a Study of Their Base Stochastic Processes.* John Wiley & Sons, 1965.

[28] V. Ramaswami. A duality theorem for the matrix paradigms in queueing theory. *Stochastic Models*, 6(1):151–161, 1990.

[29] V. Ramaswami and Marcel F. Neuts. A duality theorem for phase type queues. *The Annals of Probability*, pages 974–985, 1980.

[30] Vaidyanathan Ramaswami. Matrix analytic methods for stochastic fluid flows. *Teletraffic science and engineering*, pages 1019–1030, 1999.

[31] M. Shaked and J.G. Shanthikumar. *Stochastic Orders.* Springer Series in Statistics, 2007.

[32] D.N. Shanbhag. Letter to the editoron a duality principle in the theory of queues. *Operations Research*, 14(5):947–949, 1966.

[33] J.G. Skellam and L.R. Shenton. Distribution associated with random walk and recurrent events. *Journal of the Royal Statistical Society, Series B (Methodological)*, 19(1):64–118, 1957.

[34] David Sonderman. Comparing multi-server queues with finite waiting rooms ii: different numbers of servers. *Advances in Applied Probability*, 11(2):448–455, 1979.

[35] V. Schmidt T. Rolski, H. Schmidli and J. Teugels. *Stochastic Processes for Insurance and Finance*. John Wiley & Sons, 1999.

[36] Lajos M. Takács. *Combinatorial methods in the theory of stochastic processes*. Wiley New York, 1967.

[37] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. 2001.

[38] R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[39] Yuedong Xu, Eitan Altman, Rachid El-Azouzi, Majed Haddad, Salaheddine Elayoubi, and Tania Jimenez. Probabilistic analysis of buffer starvation in markovian queues. In *INFOCOM, 2012 Proceedings IEEE*, pages 1826–1834. IEEE, 2012.