CHAPTER 1

INTRODUCTION

1.1 A SURVEY OF WORKS ON PARALLEL QUEUES

Parallel queues prevail in manufacturing systems, communications systems, computer systems, and also in our daily life.

There are a number of interesting problems which are related to parallel queues. One of the classic problems is the shortest queue (SQ) problem. Here one has two parallel queues and an arrival stream. A new arrival will be sent to the shortest queue. If both systems have equal occupancy, the arrival joins either with probability ½. The model is often called the symmetric shortest queue problem if two servers are working at same rate μ. For the non-symmetric case, two servers are allowed to work at different rates μ_1 and μ_2.

Haight (1958) used differential-difference equations to study the SQ system with two heterogeneous servers. Flatto and McKean (1977) studied a symmetric model for two parallel queues to obtain the steady-state joint probability distribution using generating functions and complex variable arguments. Before the works done by Flatto and McKean, some asymptotic results under symmetric conditions have been obtained by Kingman (1961) using the generating functions to study the behavior of the stationary solution for two similar queues in parallel. Later, a linear programming method was used by Halfin (1985) to study the SQ problem in a system of two single-server queues. In this method, upper and lower bounds for the steady state probabilities which are asymptotically tight in heavy traffic were derived in getting the numerical solutions. Adan et al (1990 & 1991) studied the SQ problem for the symmetric and

asymmetric case. The stationary queue length distribution was obtained by using the "compensation approach".

Yao and Knessl (2005 & 2006) investigated the infinite server shortest queue problem in two parallel $M/M/\infty$ queues for a symmetric and non-symmetric case. The asymptotic approach was then used to obtain the joint steady state queue length distribution. Yao and Knessl (2008) did similar analyses of the SQ problems in two parallel *M*/*M*/*N*/*N* queues, while Knessl and Yao (2011) dealt with the case of two parallel *M*/*M*/*K* queues. SQ problem based on a comparison of the accumulated workload was analyzed by Wu and Posner (1997) using a level-crossing approach. The stationary waiting time and queue length distributions for the asymmetric two-server SQ system were then generalized to many-server systems (*n* queues and *n* servers).

Another SQ problem is one in which there are m parallel queues and $m + 1$ arrival streams. Arrival stream i will be routed only to queue i, $i = 1, 2, \ldots, m$ and arrival stream $m + 1$ will be routed to the shortest queue at arrival time (see Turner (2000) and Fleming and Simon (1999)). In paper by Fleming and Simon (1999), an infinite server SQ model was used to describe Code Division Multiple Access (CDMA) cellular systems. The shortest queue problem with jockeying was analyzed by Adan, Wessels and Zijm (1993). Here the system consists of c parallel servers and the arrival joins the shortest queue. When there are multiple shortest queues, the arrival is assigned to one of these queues. Also if the maximum difference between the lengths of the c queues is more than some threshold value T, then a job is switched from the longest to the shortest queue. In the case of multiple longest queues, one of these queues is selected to lose an arrival. A matrix-geometric approach was used to find the equilibrium queue lengths probabilities. SQ problems with jockeying were also analyzed by Zhao and Grassman (1990) using generating functions and Sakuma (2011) using the matrix analytic approach for a system of two parallel MArP/PH/2 queues. Recently, Tarabia

(2008 & 2009) studied the SQ problems in two parallel queues with jockeying and finite capacity. In the 2008 paper, matrix-analytical techniques were used to find the steady-state probabilities and in the 2009 paper, he analyzed transient-state probabilities numerically using both Runge-Kutta and randomization methods.

Puhalskii and Vladimirov (2007) studied the tail asymptotic for *k* parallel queues in which there are multiple classes of customers who can only choose the SQ among queues assigned to them. Sakuma (2010) studied *k* parallel queues in which the join SQ policy is implemented and jockeying is permitted. Using the matrix analytic approach, he derived the tail decay rate of the stationary distribution for the longest queue. Recently, Kobayashi (2011) investigated SQ problem with *k* parallel queues using quasi-birth-and-death (QBD) and reflecting random walk process formulation.

Movaghar (2011) studied a system of *s* parallel queues each with a given capacity and a given number of servers. Customers arrive according to a Poisson process with a rate which depends on the queue sizes of the *s* queues. Each incoming customer has a deadline and may not stay in the system indefinitely. Two kinds of stationary policies for assigning incoming customers were discussed. First is the dynamic (state-dependent) policy and second is the static (state-independent) policy. In dynamic policy, an incoming customer is assigned to the shortest non-full queue (SNQ) whereas in static policy, the customer is assigned to join each parallel queue with equal probability (RANDOM). It was found that the state process of the system in the long run converges in distribution to a Markov process.

Very often in a cellular system, there is a backbone network consisting of a number of fixed base stations interconnected through a fixed network (usually wired), and of mobile units that communicate via wireless links with the base stations. For each base station, there is a geographic area called a cell within which mobile units can

communicate with the base station. Neighbouring cells overlap with each other. The mobile units communicate with each other, as well as with other networks, through the base stations and the backbone network. The user releases the channel under two conditions. First is when the user completes the call. Second is when the user moves to another cell before the call is completed. The switch to another cell while the call is in progress is called handoff.

When a base station has no free channel to allocate to a mobile user, two types of blocking occurs. *New call blocking* refers to the incident in which a new call is blocked when all the channels are busy. *Handoff blocking* occurs when a handoff is performed and there is no channel available in the new cell.

The above cellular system may be modelled as a system of m dependent queues (cells) of which the i-th queue has a fixed number c_i of servers (channels), a capacity of c_i, and a dedicated arrival stream of customers given by the mobile units which are still in cell i (see for example Sidi and Starobinski (1996), Tamba Kortequee. et al (2006)).

1.2 INTRODUCTION TO THE THESIS

In the systems of dependent parallel queues, the distributions of the arrival streams of customers are usually assumed in the literature to be Poisson while the service times are considered to have exponential distributions. The thesis attempts to deal with the more general situations in which the distributions of the arrival streams of customers are assumed to have 2-phase hypoexpeonential distributions and the service times are also assumed to have 2-phase hypoexpeonential distributions.

In the thesis we introduce the following two interaction schemes to specify the dependence relation of the parallel queues:

First Interaction Scheme:

"The customer who arrives at queue m will stay back in queue m with probability q_{mm} or cross over to another queue m' ($m' \neq m$) with probability $q_{mm'}$."

Second Interaction Scheme:

"The customer who arrives at queue m will stay back in queue m with probability q_{mm} or cross over to one of the I_s shortest queues (among the remaining $M-1$ queues) with probability $(1-q_{mm})/I_s$."

The stationary queue length and waiting time distributions in the system of M dependent queues are derived for the cases when M is small. The generalization of the proposed method for other large values of M is also given.

1.3 LAYOUT OF THE DISSERTATION

In Chapter 2, a method to find the joint queue length distribution in a system of M Hypo(2)/Hypo(2)/1 queues which follow the First Interaction Scheme is proposed.

In Chapter 3, the method for finding the queue length distribution in Chapter 2 is adapted to find the queue length distribution in a system of M Hypo(2)/Hypo(2)/1 queues which follow the Second Interaction Scheme.

Chapter 4 is devoted to the derivation of the waiting time distribution in a system of M Hypo(2)/Hypo(2)/1 dependent queues.

In Chapter 5, the method proposed in Chapter 2 is modified to find the queue length distribution in a system of two dependent Hypo(2)/Hypo(2)/c/c queues.

The thesis is concluded by some concluding remarks.