A study of stochastic network with concurrent resources occupancy

ANG TECK MENG, MARCUS

(MSc,Singapore-MIT-Alliance(NUS))

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF DECISION SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2007

Acknowledgement

First, I would like to thank my supervisor A/P Ye Hengqing, whose unwavering support and solid guidance were instrumental in the initiation of this thesis. He introduced me to the beauty of stochastic modeling and personally nurtured me during the initial stage of my research. I would not have come this far without him spending so much of his time mentoring me on the difficult and challenging aspects of stochastic modeling. I have enormous respect for his expertise, and I have benefited in no small ways. He is indeed a scholar in every sense of the word, always prompting me to search for the most elegant way (if it exists) of solving a problem or proving a theorem. Thank you A/P Ye Hengqing. Working with you has been a deep and sobering intellectual experience, an experience that I thoroughly enjoyed though at times a whit frustrating. Your patience with me has been exemplary and I will always look towards you for intellectual guidance and stimulation.

I would also like to express my gratitude to Dr Cao Chengxuan who has helped me at the later stage of my research, especially on control in stochastic networks with concurrency resource occupancy and batch arrival.

Last but not least, I would like to thank National University of Singapore for providing me with the financial support to see me through my years as a doctoral student.

ANG Teck Meng, Marcus

14th November 2007

Summary

Network models with resources that are utilized concurrently to process jobs are considered in this thesis. The research on such models is motivated by issues in logistics management and communication systems.

The first part of the thesis studies the stability of network with random job arrival and service. In particular, each job upon arrival will be routed to a route that consists of a set of links (resources). We suppose that the network allows routing of jobs to achieve more flexibility in the allocation. The allocation of capacities of the link in the network is dynamically determined by some allocation policy, which is derived by solving a optimization problem that maximizes some utility function. A network is said to be stable under a given capacity allocation policy if roughly speaking the number of ongoing jobs in the network do not blow up over time. Using the fluid model approach, we show that the network is stable if the nominal workload offered to each link is within the link capacity.

The second part of the thesis is motivated by the work of Li and Yao (2004), in which a booking limit control policy based on a fixed point approximation was developed for a network with concurrent resources. When specific to the airline industry, the objective is to optimize the expected revenue subjected to the availability of seats on the flights. In our work, we allow batch passenger arrival. Our solving methodology involves deriving a fixed point approximation to express the network operating under a set of booking limits, and reformulating it into a linear program to solve for the booking limits. We show that the policy is optimal under certain limit. We also carry out extensive simulation studies, and draw interesting insights regarding the effect of the batch size on the expected revenue. Another contribution made is to study the updating mechanism for the booking limit, which turns the originally static policy to a dynamic one. Numerical analysis demonstrates significant improvement of dynamic policy.

Keywords:

concurrent resources, asymptotic optimality, batch size, booking limit, fluid limit

Contents

Chapter 1

Stability of stochastic network with routing

This chapter focuses on the study of the stability of a generic network which consists of a set of links and a set of possible routes which can be represented as fixed subsets of the links. The stability issue of a fixed routing network is studied by Ye, *et al.*(2005). We extend the result of the stability of the stochastic network models with fixed routing to the case with routing. The allocation of capacities of the link in the network is dynamically determined by some allocation policy, which is derived by solving a optimization problem that maximizes some utility function. A network model is said to be stable under a given capacity allocation policy if the number of ongoing jobs in the network does not "blow" up over time. We consider the stationary network model. The necessary stability condition (capacity constraint at each link) is clear, but the sufficient condition for stability requires a more rigorous proof. Our attempt to prove stability is via a fluid network approach.

1.1 Introduction/Outline

Our study is based on a class of stochastic networks with concurrent occupancy of resources shared by a number of different classes of jobs/customers. Such networks are present in many different applications. One example is the planning of a multi-leg flight on an airline reservation system. In order for a customer to book a 2 leg flight, seats on both legs must be reserved concurrently. Other examples include a make-to-order or assemble-to-order manufacturing system. When an order arrives, the production of all the components required will be processed simultaneously.

Analogously, the study of such a class of stochastic networks is closely related to the engineering design of Internet protocols. In modern data communication networks, digitized documents, like emails, files, images and sound, are transmitted from one source to another in packets. Often, there is no direct route from one source to another; hence the packets get routed to a series of transmission links before reaching its destination. Given today's technology, the speed of the packets is in the high range of 155Mbit/s to 2.5Gbit/s, hence a good approximation is to assume a concurrent usage of all the transmission links involved.

An extension to the model is the introduction of routing in the system. In the airline reservation system, often there is a choice for the planner to allocate to the customer on his choice of routes. We introduce the notion of routing in our stochastic network. Suppose a customer can go to his destination via two routes, say route A and route B. The planner will decide if it is more profitable or feasible to route the customers via route A or route B. Constraints like availability of seats in routes, cost and distance of the both routes and any interference from other airline using the same routes have to be considered. In the example of Internet protocols, the notion of routing gives the transmission of data more flexibility and robustness.

An abstract mathematical model of this class of network consists of a set of transmission links, and a set of possible routes with each route traversing a subset of links. It is straightforward to assume that the arrivals follow a Poisson process, and we build our model from there. Certain generalization can be made to the arrival process. It can be assumed to be a stationary renewal process. The arrival process can also be modeled in a bursty model introduced by Cruz (1991a,b). The service rate is assumed to be exponentially distributed for ease of technical analysis. One of the main concerns in such application is to derive a policy/protocol to control the routing of connections/job allocations. We assume that the routing of the connections in the network is determined according to some protocol/policy. The maximum throughput, proportionally fair and the minimum potential delay are some examples of such policies. The real-time allocation of the capacity of the links to each class of jobs/customers is derived from solving an optimization problem for each network state. Our study involves the macroscopic behavior of the network, i.e. the asymptotic convergence of the network. The microscopic study of how the jobs/connections are being established dynamically is beyond the content of this chapter. Essentially, we assume that the allocation is adapted accordingly and immediately.

Our main concern for the network is its stability, that is, given a allocation policy, will the queue of the network builds up to infinity over time. One obvious necessary condition for the stability of the network is that the average offered traffic on each link must be within the link's capacity. Subsequently, we will see that this condition is not a sufficient one. The use of the fluid model approach to analyze such networks is widely accepted, since there are results that state that a queueing network is stable if its corresponding fluid model (a continuous analog of the queueing network) is stable. Consequently, in order to use this result, our next task is to identify the corresponding fluid network model, followed by the establishment of the stability of the fluid network model. One technique in proving the stability is via the use of the Lyapunov function. This will be shown in the subsequent section of this chapter.

Outline

The outline of the chapter is as follows. We review some of the relevant papers related to this field of studies in section 1.1.1. In section 1.2, we introduce the mathematical model for the stochastic network and present some common policies. The notion of routing will be incorporated into the mathematical model. One contribution is to generalize the properties of the utility function. Thereafter, the stationary network model will be introduced in section 1.3. The stability results of the stationary network model will be given in the respective subsections. We also describe a bursty network model, and give the model and the similar results in the appendix. The fluid model for the network models used to prove the stability of the actual network will be given in section 1.4. In conclusion, we will close this chapter by consolidating our results in section 1.5.

1.1.1 Literature Review

This work is closely related to Ye, *et al.*(2005). They studied the stability of the network via fluid modelling. The conditions of the network are relaxed, and the results focused on a stationary network and a bursty network. However, an additional property, the partial radial homogeneity property, has to be assumed for the U-utility function in order to prove the stability results. The model presented in this chapter is similar to theirs. Assumptions like the service rate of jobs/arrivals following a exponential distribution remains unchanged. Properties like concurrent resource occupancy (See Whitt (1985)) is preserved. The major addition to the model is the feature of routing, which adds to the complication of the stochastic network. We seek to prove the stability of the routing network with respect to the U-utility allocation policy introduced in their paper, using the fluid model approach. We assume that the allocation allocated to each job converges to the solution of an optimization problem that maximizes some utility objective function after a short transitional period. Thus we assume that the allocated jobs are established dynamically, and the allocation is set up instantaneously upon solving the optimization problem for the optimal allocation. We referred such a equilibrium property as a microscopic stability property of the rate control and allocation of networks. For the microscopic aspect of such network, Kelly (1991) studied the problem of routing of such queueing networks. The multi-class flow model, which includes queueing and road traffic networks and telecommunication networks, was studied in that paper. One of the results is that the microscopic behavior of a telephone network, in terms of random arrival streams and rules for accepting and routing calls, cause the network to behave as if it is attempting to minimize some potential objective function. Analyzing such issues is beyond the scope of this chapter. Our concern is the macroscopic stability of networks. Another explanation for the focus on the macroscopic aspect of such network is the "separation of time scales". To be more precise, we treat the queueing of packets at the links and the bandwidth allocation to be set up immediately. Hence, we treat the time scale of the packet level rate control, which refers to the queueing of packets and bandwidth allocation of network, is small compared with the time scale of the connection level dynamics, which refers to the transmission duration for a connection.

Review of some allocation policies

With today's technology, research in loss networks has developed into area called the bandwidth sharing networks. The service capacity or the bandwidth on each link/server is shared at any time among all related jobs in the process at the link. These networks used to be focus on the study of internet protocols (e.g. TCP). Now it leads to the studies of new allocation schemes with applications to other areas like manufacturing and servicing industries. There are many studies on allocation schemes. We give a few schemes studied. Bertsekas and Gallager (1992) studied the classical max-min allocation policy, which gives the greatest possible allocation to the most poorly treated jobs. In short, an allocation is called a max-min fair allocation if the allocation to a job cannot be increased without decreasing that of another job having a smaller or equal allocation. There are many variations of the max-min allocation policies. To name a few, Cao and Zegura (1999) studied a bandwidth allocation scheme which can be viewed as a particular case of the bandwidth max-min allocation when the utility of all applications are equal. Fayolle *et al.*(2001) introduce the so-called min bandwidth sharing policy which is a conservative approximation to the classical max-min policy. The necessary and sufficient ergodicity conditions for best-effort networks under such a min policy is established.

The proportional fairness allocation policy by Kelly (1997), proposed an allocation which is determined by how much the user contributes. In the paper, it is shown that if each user is given the choice of charge per unit time that it is prepared to pay, and if the the allocated rates are determined by the network such that the rates per unit charge are proportionally fair, then the system optimum is achieved when the users' choices of charges and the network's choice of allocated rates are in equilibrium. Using such an allocation favor those poorly treated jobs but it is still not as much as the max-min allocation. In short, the objective can be interpreted as maximizing the overall utility of rate allocations. The logarithmic utility function is used to capture the characteristic of the law of diminishing return. This policy is further experimentally validated by Hurley, *et al.*(1999). A variation of the proportional fairness property is done by Mo and Walrand (2000). In their paper, the end-to-end window based congestion control protocols for packet switched networks with first come first served routers is studied. In their policy, the user controls its window size based on the total delay, whereas the user in Kelly's (1997) model controls the rate based on the feedback from the routers the connection goes through. Mo and Walrand's definition of fairness generalizes proportional fairness and strike a compromise between the user fairness and resource utilization. They went on deeper into the problem and further generalized to (p, α)-proportional fairness allocation policies. They have shown that the protocol converges at a fast rate and their proof is done using a Lyapunov function. Massoulie and Roberts (1999) introduced another criterion to the proportional fairness criterion, which is interpreted in terms of overall potential delay of the transfer of network flow in progress. Minimizing potential delay as a sharing objective provides an intermediate solution which is a compromise between max-min and proportional fairness. They also investigated the issue of deriving different possible bandwidth sharing allocations objectives and the design of flow control algorithms.

In particular for the case of the current Internet network, Kelly (2001) derived the $\arctan(\cdot)$ scheme that approximated the bandwidth allocation achieved by a type of TCP rate control protocol, called the Jacobson's TCP algorithm operating in the current Internet. The paper also address the issue on how mathematical models are being used to handle the problem of stability and fairness of rate control algorithms for the Internet. The models presented are a simplication of the complicated Internet, but nevertheless, it gives us a insight on how the Internet works. Such dynamic allocation takes the form of a solution to an optimization problem, with the objective being a utility function (of the state and the allocation), and the capacity constraints of the links (reflecting the concurrent resource occupancy). Low (2002) proposed a duality model of congestion control and applied the model to have a deeper understanding of the properties of the protocols used in Internet. Congestion control represent a distributed algorithm that optimally allocated network resources among competing sources that share the same type of resources. It consists of two components : a primal algorithm that determines the source rates in response to the congestion in its flow path, and a dual algorithm that updates a congestion measure and sends feedback to the sources that use that link. In the current Internet, the primal algorithm is carried out by the TCP algorithms, and the dual algorithm is carried out by the active queue management (AQM) schemes.

Stability of network

One of the concerns of such network is the stability of it. In other words, is the underlying Markov process positive recurrent? This forms the main thrust of this chapter. We have given a review of a number of allocation policies studied. We assume that the connections of the network is established according to some given protocol/allocation. Given an allocation policy, the network model is said to be stable if the flow in the network will not "blow up" over time. A necessary condition, called the normal offered load condition, which states that the average offered load on each link is within its link's capacity. However, Bonald and Massoulie (2001) have shown that this is not sufficient for the stability of the network with a counterexample.

For studies of stability of such networks, Bonald and Massoulie(2001) and de Veciana, *et al.*(2001) have shown the stability of network for a broad class of fair allocation under normal offered load conditions. Ye (2003) generalized their work and show that a number of common allocation schemes can be represented as a general utility function with certain properties. In the paper, it is shown that under the normal offered load condition, a network is stable using the bandwidth flow allocated according to the optimal solution when maximizing a class of general utility functions.

Fluid Model

With complications from the probabilistic behavior of such network even under Markovian assumptions, this motivated the use of fluid models of such networks, where the discreteness and randomness of the jobs are transformed via law of large numbers scaling, into continuous and deterministic values. Using the fluid model, one can obtain the stability of the network. The fluid model approach was first proposed by Rybko and Stolyar (1992), and has been an area of active research in the past decade. However, the converse is not necessarily true, that is, there exists queueing networks that are stable, but whose fluid models are instable. Bramson (1998) investigated this issue and presented a family of queueing networks with this characteristic. However, often we are interested in the stability of the original queueing network and not on the associated fluid model. We list some of the research work in this area.

Dai(1995) proved that a queueing network is positive Harris recurrent (which implies that the invariant measure is finite) if the corresponding fluid limit model converges to zero regardless the initial system configuration. To illustrate the result, the result was applied to a number of network like the single class network and multiclass network under the normal offered load condition. Chen (1995) extends the results and prove the stability of multiclass queueing network with general work-conserving disciplines. However, a queueing network (fluid network) may be stable under one service discipline, but proved to be unstable under another. Rybko and Stolyar (1992) provided a two-station queueing network which is stable under First-in-First-out (FIFO) but unstable under a priority service discipline. Dai and Meyn (1995) did a study on the open multiclass queueing network and one of their focus is on the moments on the queue length. Using the fluid approach, they provide sufficient conditions on the existence of bounds on long-run average moments of the queue lengths, and bound the rate of convergence of the mean queue length to its steady state value. Stolyar (1995) showed that the sequence of scaled (in space and time) underlying stochastic processes converges to a fluid process along some sample paths. The convergence together with the continuity and similarity properties of the sample paths of the fluid process shows that the original network is stable if each sample path of the fluid process with non-zero initial state falls below the initial value at least once. Bramson (1998) investigated the stability of two families of queueing network, namely the head-of-the-line network and the re-entrant network in a deterministic setting.

The application of these results require us to identify the corresponding fluid network model and use the Lyapunov function approach to prove the stability of the fluid network. This is one of the primary tools in establishing the stability of the fluid network. Following that, the stability of the original data network can be derived accordingly. Bonald and Massoulie (2001) prove the stability of of data network with (p, α)-proportionally fair bandwidth allocation with such a fluid model approach. Ye and Chen (2001) studied the use of the Lyapunov function and gave a theorem which facilitates the use of the fluid model approach. In their paper, they derive a necessary and sufficient condition for the stability of a generic fluid network, which is the existence of a Lyapunov function for its fluid level process. Chen and Ye (2002) utilize a piecewise Lyapunov function to obtain the sufficient conditions for the stability of a multiclass fluid network under priority service discipline. This work extends and generalizes the work by Ye and Chen (2002) that is based on a linear Lyanuov function.

1.2 Introduction to Network Infrastructure and Capacity Allocation

Consider a network with a set L of links/servers where each link $l \in L$ has a bandwidth of capacity C_l. Let S denote the set of sources. Each route in the network can be described by an index (s, r), where s is a source (describing a joint or distribution point) and r is a route directed from the source. We use the notation $r \in s$ where $s \in S$ denote a route r being one of the choice of route from source s. It follows that for $s \in S$ and $r \in s$, the index (s, r) represents a particular route.

The work allocation policies we consider depend only on the ongoing jobs in all routes. Suppose n_s is the number of ongoing jobs on source s and let $n = \{n_s : s \in S\}$. Hence, $n_s = \sum_{r \in s} n_{s,r}$ for some $n_{s,r} \geq 0$. $n_{s,r}$ represent the number of ongoing jobs going to route (s, r) from source s. We denote $a_s(n)$ as the work allocation(amount of work per unit time) allocated to each job on source s. Using this definition, the total work allocation to all jobs on source s is $\Lambda_s = n_s a_s(n)$. Thus $\Lambda_s = \sum_{r \in s} \Lambda_{s,r}$, for some $\Lambda_{s,r}(n) = n_{s,r} a_s(n) \geq 0$. $\Lambda_{s,r}$ is the total work allocation allocated to all jobs on route (s, r) from source s.

An allocation $\Lambda(n) = \{\Lambda_s(n) | s \in S\}$ is feasible if and only if the following feasible conditions are satisfied. For ease of presentation, we replace $\Lambda(n)$ by Λ. The feasible conditions are as follows:

$$\begin{aligned}
&\exists \Lambda_{s,r} \geq 0 \text{ s.t. } \Lambda_s = \sum_{r \in s} \Lambda_{s,r} \text{ for } s \in S \\
&\sum_{s \in S, l \in r, r \in s} \Lambda_{s,r} \leq C_l \text{ for } l \in L \\
&\Lambda_s = 0 \text{ if } n_s = 0 \text{ for } s \in S \\
&\Lambda_s \geq 0 \text{ for } s \in S.
\end{aligned} \tag{1.1}$$

Let M_r denote the feasible set of the routing model. Without routing, the feasible set

$$\begin{aligned}
&\sum_{l \in r} \Lambda_r \leq C_l \text{ for } l \in L \\
&\Lambda_r = 0 \text{ if } n_r = 0 \text{ for } r \in R \\
&\Lambda_r \geq 0 \text{ for } r \in R.
\end{aligned} \tag{1.2}$$

where the set R denotes the set of fixed routes in the network. Let M_{nr} denote the feasible set of the fixed route model. It is obvious that the feasible set for the fixed route model is smaller than that with routing. Hence, $M_{nr} \subset M_r$.

Remarks:

1. Note that the feasible region M_r for the routing case and M_{nr} for the non-routing case are convex polyhedral. The feature of routing increases the number of free variables in the feasible region and enlarges the size of the feasible region. To further generalize our results, our results hold as long as we restrict the feasible region to a convex polyhedral.

1.2.1 U-utility maximization allocation

As mentioned in the introduction, our main emphasis is to investigate the network stability. In this section, we introduce a generic class of utility maximization allocation policy, called the U-utility maximizing policy, which covers a number of allocation policies. The U-utility maximization allocation refers to the unique optimal solution of the following optimization problem:

$$\max_{\Lambda \in M_r} \sum_{s \in S} U_s(n_s, \Lambda_s) \tag{1.3}$$

where U_s satisfy the following properties:

1. $U_s(n_s, \Lambda_s)$ are second-order differentiable on $\Re_+ \times (0, \infty)$.

2. $U_s(0, \Lambda_s) = 0$ for $\Lambda_s > 0$.

3. $\partial_2 U_s(0, \Lambda_s) = 0$ for $\Lambda_s > 0$.

4. $\partial_2 U_s(n_s, \Lambda_s) > 0$ for $\Lambda_s, n_s > 0$

5. $\partial_1 \partial_2 U_s(n_s, \Lambda_s) > 0$ for $\Lambda_s, n_s > 0$

6. $U_s(n_s, \cdot)$ is strictly concave for fixed $n_s > 0$.

For technical requirement, we need the following condition.

Partial radial homogeneity property:

$$\Lambda_s^U(cn_s) = \Lambda_s^U(n_s) \tag{1.4}$$

for any $s \in S$ with $n_s > 0$ and any $c > 0$.

Remarks:

1. The first four assumptions are intuitively appealing. The fifth assumption states that increasing the allocation is more rewarding in the case of a higher n_s. The sixth constraints implies concavity. Intuitively, this means that adding an extra allocation is more beneficial when the allocated allocation is small than when it is large. (See Ye et al. (2005) for more details).

Technically, it can be shown that this generic class of utility policy leads to a unified treatment for the stability problem of some more specified policies. Some examples of such specified policies which fall under this category is as follows:

1. the proportionally fair allocation: $U_s(n_s) = n_s \log(\Lambda_s)$,

2. the minimum potential delay allocation : $U_s(n_s) = -\frac{n_s^2}{\Lambda_s}$,

3. the (p, α)-proportionally fair allocation : $U_s(n_s) = p_s n_s^\alpha \frac{\Lambda_s^{1-\alpha}}{1-\alpha}$ where $p_s, s \in S$ are fixed parameters.

The assumption on partial radial homogeneity is a technical requirement and it is not as restricted as it seems. It can be verified that utility functions for allocations like the proportionally fair allocation, the minimal potential delay allocation and the (p, α)-proportionally fair allocation satisfied all the above assumptions.

It may not seem obvious how we verify the partial radial homogeneity property for the allocation policy. A more 'convenient' form to verify the partial radial homogeneity property is to use the following lemma.

Lemma 1. *Suppose there exist a positive function $f : \Re^+ \to \Re^+$ such that $U_s(cn, \Lambda) = f(c)U_s(n, \Lambda)$, for any $s \in S$ with $n_s > 0$ and any $c > 0$, then the partial radial homogeneity property is satisfied.*

Proof. Let $\Lambda^U(n) = \{\Lambda(n) : \max_{\Lambda \in M_r} \sum_{s \in S} U_s(n_s, \Lambda_s)\}$. Then

$$\begin{aligned}
\Lambda^U(cn) &= \{\Lambda(n) : \max_{\Lambda \in M_r} \sum_{s \in S} U_s(cn_s, \Lambda_s)\} \\
&= \{\Lambda(n) : \max_{\Lambda \in M_r} \sum_{s \in S} f(c) U_s(n_s, \Lambda_s)\} \text{ (by assumption)} \\
&= \{\Lambda(n) : \max_{\Lambda \in M_r} f(c) \sum_{s \in S} U_s(n_s, \Lambda_s)\} \\
&= \{\Lambda(n) : f(c) \max_{\Lambda \in M_r} \sum_{s \in S} U_s(n_s, \Lambda_s)\} \\
&= \{\Lambda(n) : \max_{\Lambda \in M_r} \sum_{s \in S} U_s(n_s, \Lambda_s)\} \\
&= \Lambda^U(n) \qquad (1.5)
\end{aligned}$$

□

As an illustration, consider the proportionally fair allocation, $U_s(n_s) = n_s \log(\Lambda_s)$. Our objective function is $U(n, \Lambda) = \sum_{s \in S} n_s \log(\Lambda_s) = \sum_{s \in S} U_s(n_s, \Lambda_s)$. Let $f(x) = x$. Hence, $U_s(cn_s, \Lambda_s) = cn_s \log(\Lambda_s) = f(c)n_s \log(\Lambda_s) = f(c)U_s(n_s, \Lambda_s)$. Hence the proportionally fair allocation satisfies the partial homogeneity property.

Remarks:

1. One can verify that the arctan-utility maximization allocation ,$U_s(n_s) = w_s n_s \arctan(\frac{\Lambda_s}{w_s n_s})$ where w_s is a positive constant) is not a special case of the U-utility maximizing allocation. This is due to a violation of the partial radial homogeneity property or property 1.

2. The simplest utility function is the maximum throughput allocation, $U_s(n_s, \Lambda_s) = \Lambda_s$. But, in general, the maximum throughput allocation does not give a unique allocation for a fixed set of n jobs. It also does not fall under the category of the U-utility maximization allocation.

We see that the U-utility maximizing allocation is a representation of several common allocations. Hence, we use the U-utility maximizing allocation in our analysis for the rest of the paper. One of the drawbacks is that the U-utility maximizing allocation is unable to capture the characteristics of the arctan-utility maximization allocation, which is seen as a good approximation for the internet protocol.

1.3 Network Models

The allocation of service capacities takes place in each state n, and is determined by some optimizing problem (1.3). In the case of data transmission for internet protocols, connections for data transmission are established and terminated dynamically in real data networks according to some optimization problem. In order to ease the theoretical analysis and yet gain acceptable approximation to real problems, the assumption that the arrival processes of jobs are independent stationary renewal processes, for example, independent Poisson processes, is often used. However, such assumption can be unrealistic as arrival processes are often correlated and bursty, and this can affect the performance of the network. One approach to handle this is to use the bursty model introduced in Cruz (1991). To prevent further digress from the topic, we will present the bursty model in the appendix and focus our analysis on the stationary network model.

In this section, we present the stationary network model in detail and propose a complementary model for the stationary network model. The concept of stability in the network is defined. In an analogous way, the stability of the stationary network can be described as positive Harris recurrence of the underlying Markov process that captures the dynamics of the model. In order to set up the main result for this chapter, we first present the main theorem in this section which we will prove subsequently.

1.3.1 Stationary Network Model

In the stationary network, job arrivals to source $s \in S$ form independent stationary renewal process with mean arrival rate λ_s. Let $u_s(i)$ denote the interarrival time between $(i-1)$th and the ith job on source s. Hence $u_s(i)(i \geq 2)$ are i.i.d. random variables with mean $1/\lambda_s$, while the first residual arrival $u_s(1)$ is the residual arrival time. The work processed by the ith job on source s is denoted by $v_s(i)$, which are i.i.d. exponentials with mean ν_s.

We need two technical conditions on $u_s(i)$: an unbounded condition and a spread out condition.

Unbounded condition:

$$P\{u_s(1) \geq x\} > 0, \text{for any } x > 0 \tag{1.6}$$

Spread out condition:

There exist some integer j_s and some function $p_s(x) \geq 0$ for $x \geq 0$ with $\int_0^\infty p_s(x)dx > 0$ such that

$$P\{a \leq \sum_{i=1}^{j_s} u_s(i) \leq b\} \geq \int_a^b p_s(x)dx, \text{ for any } 0 \leq a < b \tag{1.7}$$

It is worth mentioning that the above two conditions are necessary for our results to hold, but we do not apply it directly. In the proof for the stability result for the stationary network model, the above conditions are relaxed by introducing the concept called the petite set, see Bramson(1998).

Let $\lambda = \{\lambda_s : s \in S\}$ and $\nu = \{\nu_s : s \in S\}$. Then the average offered traffic load to each source $s \in S$ in terms of amount of work per unit time is:

$$\rho_s = \lambda_s \nu_s \tag{1.8}$$

With routing, we see that there exists some average offered traffic load to each route from related source, represented by $\rho_{s,r} \geq 0$, such that:

$$\rho_s = \sum_{r \in s} \rho_{s,r} \tag{1.9}$$

for some $\rho_{s,r} \geq 0$ for all $r \in s, s \in S$. We interpret $\rho_{s,r}$ as the average offered traffic on route (s, r).

Given a state dependent allocation policy $\Lambda(.)$, the dynamics of the stationary model can be captured by a Markov process. Let N_s denote the number of jobs to be processed from source s. Then $N(t) = \{N_s(t) : s \in S\}$ is the ongoing job process. We now have for each $s \in S$

$$N_s(t) = \sum_{r \in s} N_{s.r}(t) \tag{1.10}$$

for some $N_{s,r}(t) \geq 0$.

If the job arrival processes are Poisson process, $N(t)$ is a continuous time Markov chain with transition rates given by

$$q(n,n') = \begin{cases} \lambda_s & \text{for } n' = n + e_s \\ \nu_s^{-1}\Lambda_s(n) & \text{for } n' = n - e_s, n_s \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $n, n' \in Z_+^{|S|}$.

In the more general case, i.e. general stationary renewal arrival processes, it is necessary to refine the structure of the network model by introducing more measures in order to capture the network dynamics accurately. Let $U_s(t)$ denote the remaining time before the next job arrival on source s at time t and $V_s(i,t)$ denote the amount of work of ith job on source s that has not been processed at time t. We treat the ongoing connections on a source to be lined up in the order of the arrival. Then $(N(t), U(t), V(t)) = \{N_s(t), U_s(t), V_s(t); s \in S\}$ is a strong Markov process describing the dynamics of the data network.(See Dai(1995) or Davis(1984) for a more comprehensive explanation.) The network model is said to be stable if the Markov process is positive Harris recurrent.

A necessary condition for stability is that the normal offered load condition holds, i.e. the average offered load to every link in the network is within the capacity of the link:

$$\sum_{s \in S, l \in r, r \in s} \rho_{s,r} < C_l, \text{for } l \in L \tag{1.11}$$

for some $\rho_{s,r} \geq 0$ where $\rho_s = \sum_{r \in s} \rho_{s,r}$.

We adopt a process sharing (PS) system for the stationary network in this section. In short, for this system, the capacity of a link (e.g. the processor) allocated to each route is shared equally by all ongoing jobs. The analysis using the PS system is in general difficult. Hence, we consider an alternative stationary model under the head-of-the-line processor sharing (HOLPS) system. Under the HOLPS, all the capacity allocated to a route goes to the ongoing job which is established first. The similarity for the two system is that it is sufficient to capture the network dynamics of the two system by a Markovian state descriptor, also denoted as $(N(t), U(t), V(t))$. The mathematical details for this is omitted since it is not used explicitly in the rest of the paper.

Under the exponential assumption of the processing workload, the ongoing job process is equal in distribution in the two systems. This is so because, in both systems, the rate at which ongoing connections in a route finish transmissions depends on the allocation allocated to the route, which in turns depends only on the number of ongoing jobs on each route. This leads to the deduction that the other two Markovian state descriptor, $U(t)$ and $V(t)$, are equivalent in distribution, since both depends on the ongoing connection process $N(t)$. Thus, we claim that the positive Harris recurrence of the HOLPS system implies the positive Harris recurrence of the PS since they have the same distribution. Therefore, in order to facilitate us in our technical analysis of the network, we assume that the stationary model is a HOLPS system for the rest of the paper.

1.3.2 HOLPS system

In the network model (HOLPS) we are working on, it is useful to introduce more performance measures to describe the network dynamics better.

Performance measures:

1. Queue length process : $X(t) = \{X_s(t) : s \in S\}$

2. Job arrival process : $E(t) = \{E_s(t) : s \in S\}$

3. Workload arrival process :$A(t) = \{A_s(t) : s \in S\}$

4. Capacity allocation process : $D(t) = \{D_s(t) : s \in S\}$

5. Job departure process : $S(t) = \{S_s(t) : s \in S\}$

$X_s(t)$ is the immediate remaining work load (in terms of the amount of work) embodied in the $N_s(t)$ ongoing jobs on source s at time $t \geq 0$ and is given by:

$$\begin{aligned} X_s(t) &= \sum_{i=1}^{N_s(t)} V_s(i,t) \\ &= \sum_{r \in s} \sum_{i=1}^{N_{s,r}(t)} V_s(i,t) \\ &= \sum_{r \in s} X_{s,r}(t) \qquad (1.12) \end{aligned}$$

where $X_{s,r}(t) = \sum_{i=1}^{N_{s,r}(t)} V_s(i,t)$. Recall that $V_s(i,t)$ is the amount of work of ith job on source s that has not been processed at time t. With routing, for each job i at the source s, it will be routed to one of the routes $r \in s$. Hence, $X_{s,r}(t)$ is the queue length at the route (s,r).

$E_s(t)$ is the total number of jobs that have arrived to source s during the time interval $[0, t]$ for $t \geq 0$, and is given by:

$$E_s(t) = \sup\{i : U_s(0) + U_s(1) + ... + U_s(i) \leq t\} \tag{1.13}$$

Recall that $U_s(t)$ is the remaining time before the next job arrival on source s at time t.

$A_s(t)$ is the total amount of workload embodied in all jobs that have been established at source s during time interval $[0, t]$ for $t \geq 0$, and is given by:

$$A_s(t) = \sum_{i=1}^{E_s(t)} v_s(i) \tag{1.14}$$

$D_s(t)$ is the total amount of work that has been processed via source s during the time interval $[0, t]$ for $t \geq 0$, and is determined by the capacity allocation process/policy Λ. Then at source s, we have:

$$D_s(t) = \int_0^t \Lambda_s(N_s(\tau))d\tau \tag{1.15}$$

and Λ_s is solved from the optimization problem:

$$
\begin{aligned}
& \max \sum_{s \in S} U_s(n_s, \Lambda_s) \\
s.t. \quad & \exists \Lambda_{s,r} \geq 0 \text{ that } \Lambda_s = \sum_{r \in s} \Lambda_{s,r} \text{ for } s \in S \\
& \sum_{s \in S, l \in r, r \in s} \Lambda_{s,r} \leq C_l \text{ for } l \in L \\
& \Lambda_s = 0 \text{ if } n_s = 0 \text{ for } s \in S \\
& \Lambda_s \geq 0 \text{ for } s \in S. \qquad (1.16)
\end{aligned}
$$

where $U_s(n_s, \Lambda_s)$ is some utility function. Note that Λ_s and $\Lambda_{s,r}$ are both decision variables for $r \in s, s \in S$. From (1.15), we have

$$
\begin{aligned}
D_s(t) &= \int_0^t \Lambda_s(N_s(\tau))d\tau \\
&= \int_0^t \sum_{r \in s} \Lambda_{s,r}(N_s(\tau))d\tau \\
&= \sum_{r \in s} \int_0^t \Lambda_{s,r}(N_s(\tau))d\tau \\
&= \sum_{r \in s} D_{s,r}(t) \qquad (1.17)
\end{aligned}
$$

where $D_{s,r}(t) = \int_0^t \Lambda_{s,r}(N(\tau))d\tau$.

$S_s(y)$ is the number of jobs that have been processed if amount of work that has been processed via source s is equal to y, and under HOLPS, is given by:

$$S_s(y) = \max\{i : v_s(1) + ... + v_s(i) \leq y\} \qquad (1.18)$$

Thus, $S_s(D_s(t))$ is equal to the number of jobs that have been processed up to time t.

Finally, the processes X, N, A, E, D, S are related by the following job flow balance equations and the job flow at each node:

$$\begin{aligned} X_s(t) &= X_s(0) + A_s(t) - D_s(t) && (1.19) \\ N_s(t) &= N_s(0) + E_s(t) - S_s(D_s(t)) && (1.20) \\ \sum_{r \in s} X_{s,r}(t) &= X_s(t) \text{ for some } X_{s,r} \geq 0 && (1.21) \\ \sum_{r \in s} N_{s,r}(t) &= N_s(t) \text{ for some } N_{s,r} \geq 0 && (1.22) \\ \sum_{r \in s} D_{s,r}(t) &= D_s(t) \text{ for some } D_{s,r} \geq 0 && (1.23) \end{aligned}$$

Before we proceed further, we list the main result of this chapter, which provides the necessary and sufficient conditions for the stability of the stationary network.

Theorem 1. *Suppose the normal offered load condition (*1.11*) is satisfied for the stationary network model $(L, C, R, M, \lambda, \nu)$. Then the allocations $\Lambda^{pp}, \Lambda^{pd}, \Lambda^{mm}, \Lambda^{\alpha}$ and Λ^{U} ensure the stability of the model.*

From the theorem, we see that the stability of the model depends heavily on the allocation policy we choose. An important observation is that the maximum throughput and priority based allocation policy does not fall under the conditions of theorem 1. We know that the normal offered load condition is not sufficient to guarantee the stability of the network in the fixed route case. The conclusion is similar for the routing case. We give a simple example to highlight our point.

Example 1 (Maximum throughput allocation in a priority network with routing)

Consider a network with three links $L = \{1, 2, 3\}$. We have three classes of jobs to be processed. Suppose that job class 1 is given a higher priorty than job class 2 and 3. Assuming the arrival processes are Poisson, then the dynamics of this network can be expressed by the ongoing job process $N(t)$, which is a continuous Markov chain with transition rates

$$q(n, n') = \begin{cases} \lambda_s & \text{for } n' = n + e_s \\ \nu_{s_1}^{-1} C_2 & \text{for } n' = n - e_{s_1}, n_{s_1} \geq 1 \\ \nu_{s_2}^{-1} C_1 & \text{for } n' = n - e_{s_2}, n_{s_1} = 0, n_{s_2} \geq 1 \\ \nu_{s_3}^{-1} C_3 & \text{for } n' = n - e_{s_3}, n_{s_1} = 0, n_{s_3} \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

In the routing case, assume we have a choice of allocation for job class 2. We can route some of the job of class 2 to link 3. Hence, the necessary and sufficient condition for the positive recurrence of the Markov chain $N(t)$ is:

$$\rho_{s2} = \rho_{s2,1} + \rho_{s2,3}, \frac{\rho_{s_3} + \rho_{s2,3}}{C_3} < 1, \frac{\rho_{s_1} + \rho_{s2,1}}{C_1} < 1, \frac{\rho_{s2,1}}{C_1} + \frac{\rho_{s_1}}{C_2} < 1. \tag{1.24}$$

But the normal offered load condition is:

$$\rho_{s2} = \rho_{s2,1} + \rho_{s2,3}, \frac{\rho_{s_3} + \rho_{s2,3}}{C_3} < 1, \frac{\rho_{s_1} + \rho_{s2,1}}{C_1} < 1, \frac{\rho_{s2,1}}{C_1} < 1. \tag{1.25}$$

By inspection, we see that the condition (1.24) is stronger than (1.25).

1.4 Fluid Network and its stability

In this section, we prove theorem 1 by a fluid model approach. First of all, we introduce a fluid network model and prove its stability. If properly scaled, we show that the stationary network model will converge to the limits that satisfy the fluid network model. The stability of the stationary model can be deduced from the stability of the fluid model.

1.4.1 Introduction of Fluid Network Model

We introduce a fluid network model corresponding to the stationary network model with the U-utility maximizing allocation. The fluid network model has the same infrastructure as the stationary network model. The difference is that in the fluid network, the routes carry continuous fluid flows. In particular, on source s, the fluid flows exogenously into the network at a rate less than or equal to p_s, and is transmitted through routes $r \in s$ at a rate subject to a given allocation policy.

We introduce the following fluid processes to describe the dynamics of the fluid network.

1. Fluid queue level process : $\bar{X}(t) = \{\bar{X}_s(t) : s \in S\}$

   $\bar{X}_s(t)$ is the amount of fluid waiting to be transmitted at source s at time t. With routing, at each source $s \in S$, we have $\bar{X}_s(t) = \sum_{r \in s} \bar{X}_{s,r}(t)$, where $\bar{X}_{s,r}(t)$ is the amount of fluid being assigned by the allocation Λ for indented processing at route (s, r).

2. Job level process : $\bar{N}(t) = \{\bar{N}_s(t) : s \in S\}$

   $\bar{N}_s(t)$ is the number of jobs at source s in fluid form at time t. With routing, at each source $s \in S$, we have $\bar{N}_s(t) = \sum_{r \in s} \bar{N}_{s,r}(t)$, where $\bar{N}_{s,r}(t)$ is the amount of fluid jobs being assigned by the allocation Λ at route (s, r).

3. Fluid arrival process :$\bar{A}(t) = \{\bar{A}_s(t) : s \in S\}$

   $\bar{A}_s(t)$ is the cumulative amount of fluid that has arrived at source s during the time interval $[0, t]$.

4. Capacity allocation process : $\bar{D}(t) = \{\bar{D}_s(t) : s \in S\}$

   $\bar{D}_s(t)$ is the total amount of fluid that has been transmitted via all routes (s, r) from source s during the time interval $[0, t]$.

5. Allocation process : $\bar{\Lambda}(n, q) = \{\bar{\Lambda}_s(n, q) : s \in S\}$

   $\bar{\Lambda}_s(n, q)$ is the allocation rate allocated to source s when the job level state is $\bar{N}(t) = n$ and the fluid inflow rate is $\dot{\bar{A}}(t) = q$. To be more precise, we define the allocation rate as

$$\bar{\Lambda}_s^U(n, q) = \begin{cases} \Lambda_s^U(n) & \text{if } n_s > 0 \\ q_s & \text{if } n_s = 0 \end{cases}$$

Using routing, we have $\bar{\Lambda}_s(n,q) = \sum_{r\in s} \bar{\Lambda}_{s,r}(n,q)$, where $\bar{\Lambda}_{s,r}(n,q)$ is the allocation rate allocated to route (s,r) from source s.

Fluid network model

Given the allocation $\bar{\Lambda}^U$, the dynamics of the fluid network model is characterized by the following system of equations:

For $s \in S$,

$$\bar{X}_s(t) = \bar{X}_s(0) + \bar{A}_s(t) - \bar{D}_s(t) \tag{1.26}$$

$$\bar{N}_s(t) = \bar{X}_s(t)/v_s \tag{1.27}$$

$$\bar{A}_s(t)\text{is Lipschitz continuous and } 0 \leq \dot{\bar{A}}_s(t) \leq \rho_s \text{ a.s.} \tag{1.28}$$

$$\begin{aligned}
&\bar{D}_s(t) = \int_0^t \bar{\Lambda}_s(\bar{N}(\tau), \dot{\bar{A}}_s(\tau))d\tau \\
&= \int_0^t \sum_{r\in s} \bar{\Lambda}_{s,r}(\bar{N}(\tau), \dot{\bar{A}}_s(\tau))d\tau \\
&= \sum_{r\in s} \int_0^t \bar{\Lambda}_{s,r}(\bar{N}(\tau), \dot{\bar{A}}_s(\tau))d\tau
\end{aligned} \tag{1.29}$$

$$\sum_{r\in s} \bar{X}_{s,r}(t) = \bar{X}_s(t) \text{ for some } \bar{X}_{s,r} \geq 0 \tag{1.30}$$

$$\sum_{r\in s} \bar{N}_{s,r}(t) = \bar{N}_s(t) \text{ for some } \bar{N}_{s,r} \geq 0 \tag{1.31}$$

$$\sum_{r\in s} \bar{A}_{s,r}(t) = \bar{A}_s(t) \text{ for some } \bar{A}_{s,r} \geq 0 \tag{1.32}$$

$$\sum_{r\in s} \bar{\Lambda}_{s,r}(n,q) = \bar{\Lambda}_s(n,q) \text{ for some } \bar{\Lambda}_{s,r} \geq 0 \tag{1.33}$$

where (1.26) and (1.27) are the flow balance equation. The job level process $\bar{N}(t)$ is introduced to maintain the similarity of fluid network to the stationary network. (1.28)is a regularity arrival processes, and (1.29) is self-explanatory.(1.30)-(1.33) captures the characteristic of routing at each source $s \in S$.

The above system of equations defines a fluid network model. Any solution satisfying the above system is called a fluid solution. By definition, the fluid model is said to be stable if there exists a time $\tau > 0$ such that $\bar{X}(\tau + \cdot) \equiv 0$ (or equivalently $\bar{N}(\tau + \cdot) \equiv 0$) for any fluid solution $\bar{X}(t)$ with initial condition $\|\bar{X}(0)\| = 1$. For the stationary network, it is clear that the normal offered load condition is necessary for the fluid network model to be stable under any allocation policy. For the normal offered load condition to be a sufficient condition, we need to impose conditions on the allocation policy. To state more formally, we show the stability of the fluid network model under the U-utility allocation policy under the normal offered load condition. We will need the following results in our proof.

Proposition 1. *Suppose the normal offered load condition (*1.11*) is satisfied for the fluid network model. Then the U-utility maximizing allocation $\bar{\Lambda}^U$ ensure the stability of the fluid network model.*

Proof. : Stability of the fluid network model with allocation $\bar{\Lambda}^U(.,.)$ follows from Theorem 2.3 (i) of Ye and Chen (2001) after the following claims (a)-(c) are verified.

(a) (Scale property) For any fluid solution $\bar{N}(.)$, the process $\frac{1}{z}\bar{N}(z\dot{)}$ is also a fluid solution for any fixed $z > 0$.

(b) (Shift property)For any fluid solution $\bar{N}(.)$, the process $\bar{N}(s + \cdot)$ is also a fluid solution for any fixed $s > 0$.

(c) (Lyapunov condition) For any fluid solution $\bar{N(.)}$, there is an absolutely continuous function $f(t)$ such that for almost all $t \geq 0$,

$$w_1(\|\bar{N}(t)\|) \leq f(t) \leq w_2(\|\bar{N}(t)\|), \tag{1.34}$$

$$\dot{f}(t) \leq -w_3(\|\bar{N}(t)\|) \tag{1.35}$$

where $w_i(.), i = 1, 2, 3$ are three strictly increasing continuous functions with $w_i(0) = 0, i = 1, 2, 3$.

Claims (a) and (b) are straightforward to verify. We prove claim (c). Let

$$f(t) = \sum_{s \in S} \int_0^{\bar{N}_s(t)} v_s \partial_2 U_s(y, \rho_s(1+\delta)) dy$$
$$\text{such that } \bar{N}_s(t) = \textstyle\sum_{r \in s} \bar{N}_{s,r}(t) \text{ for some } \bar{N}_{s,r}(t) \geq 0 \tag{1.36}$$

where δ is sufficiently small so that $\{\rho_s(1+\delta), s \in S\}$ still satisfies the normal offered load condition (1.11) with ρ_s replaced by $\rho_s(1+\delta)$. Then $f(t)$ is absolutely continuous because $\bar{N}(t)$ is Lipschitz continuous and the integrands are uniformly bounded on any compact set of y. This can be deduced from condition 2 and 5 of the property of U_s.

Define three strictly increasing continuous functions $w_i(.), i = 1, 2, 3$ as follows:

$$
\begin{aligned}
w_1(y) &= \frac{y}{2|S|}\underline{w}(\frac{y}{2|S|}) \\
w_2(y) &= y\bar{w}(y) \\
w_3(y) &= (\min_{s\in S}\lambda_s)\delta\underline{w}(\frac{y}{|S|})
\end{aligned}
\tag{1.37}
$$

where $\underline{w}(y) = \min_{s\in S}\{v_s\partial_2 U_s(y, \rho_s(1+\delta))\}$ and $\bar{w}(y) = \max_{s\in S}\{v_s\partial_2 U_s(y, \rho_s(1+\delta))\}$.

Then $w_1(0) = w_2(0) = w_3(0) = 0$. We first verify (1.34). Let $\hat{s} \in S$ such that $\bar{N}_{\hat{s}}(t) = \max\{\bar{N}_s(t) : s \in S\}$. Consider the left hand side of (1.34):

$$
\begin{aligned}
f(t) &\geq \sum_{s\in S}\int_{\frac{\bar{N}_s(t)}{2}}^{\bar{N}_s(t)} \underline{w}(y)dy \\
&\geq \sum_{s\in S}\frac{\bar{N}_s(t)}{2}\underline{w}(\frac{\bar{N}_s(t)}{2}) \\
&\geq \frac{\bar{N}_{\hat{s}}(t)}{2}\underline{w}(\frac{\bar{N}_{\hat{s}}(t)}{2}) \\
&\geq \frac{\|\bar{N}_{\hat{s}}(t)\|}{2|S|}\underline{w}(\frac{\|\bar{N}_{\hat{s}}(t)\|}{2|S|}) \\
&= w_1(\|N(t)\|)
\end{aligned}
\tag{1.38}
$$

We now verify the right hand side of (1.34). Assuming condition 5 of property of U_s, we have:

$$\begin{aligned} f(t) &\leq \sum_{s\in S}\int_0^{\bar{N}_s(t)} \bar{w}(y)dy \\ &\leq \sum_{s\in S} \bar{w}(\bar{N}_s(t))\bar{N}_s(t) \\ &\leq \bar{w}(\|\bar{N}(t)\|)\|\bar{N}(t)\| \\ &= w_2(\|\bar{N}(t)\|) \end{aligned} \tag{1.39}$$

The next step is to verify (1.35):

$$\begin{aligned} \dot{f}(t) &= \sum_{s\in S} \dot{\bar{N}}_s(t)\nu_s\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &= \sum_{s\in S} \dot{\bar{X}}_s(t)\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &= \sum_{s\in S}[\dot{\bar{A}}_s(t) - \sum_{r\in s}\bar{\Lambda}^U_{s,r}(\bar{N}(t), \dot{\bar{A}}(t))]\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \text{ for some } \bar{\Lambda}_{s,r} \\ &= \sum_{s\in S}[\dot{\bar{A}}_s(t) - \bar{\Lambda}^U_{s}(\bar{N}(t), \dot{\bar{A}}(t))]\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &= \sum_{s\in S, \bar{N}_s>0}[\dot{\bar{A}}_s(t) - \bar{\Lambda}^U_{s}(\bar{N}(t), \dot{\bar{A}}(t))]\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &\leq \sum_{s\in S, \bar{N}_s>0}[\rho_s - \bar{\Lambda}^U_{s}(\bar{N}(t), \dot{\bar{A}}(t))]\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \end{aligned} \tag{1.40}$$

Consider the following optimization problem:

$$\begin{aligned} \max \quad & \sum_{s\in S, \bar{N}_s>0} U_s(\bar{N}_s(t), \Lambda_s) \\ s.t. \quad & \sum_{l\in r, r\in s, \bar{N}_s>0, s\in S} \Lambda_{s,r} \leq C_l \text{ for } l \in L, \\ & \Lambda_s \geq 0 \text{ where } \Lambda_s = \textstyle\sum_{r\in s}\Lambda_{s,r} \text{ for some } \Lambda_{s,r} \geq 0 \end{aligned} \tag{1.41}$$

By our choice of δ, we see that $\rho_s(1+\delta)$ is feasible solution to the above optimization problem, and $\{\bar{\Lambda}_s^U(\bar{N}(t), \dot{\bar{A}}(t)) | \bar{N}_s > 0\} = \{\bar{\Lambda}_s^U(\bar{N}(t)) | \bar{N}_s > 0\}$ is optimal. Using the property that $\partial_2 U_s(n_s, \Lambda_s) > 0$ for $n_s, \Lambda_s > 0$ and the feasibility of $\rho_s(1+\delta)$, we have

$$\begin{aligned} \sum_{s\in S, \bar{N}_s > 0} \rho_s(1+\delta) &\leq \sum_{s\in S, \bar{N}_s>0} \bar{\Lambda}_s^U(\bar{N}(t), \dot{\bar{A}}(t)) \\ \Rightarrow \sum_{s\in S, \bar{N}_s>0} \rho_s - \bar{\Lambda}_s^U(\bar{N}(t), \dot{\bar{A}}(t)) &\leq - \sum_{s\in S, \bar{N}_s>0} \delta\rho_s \end{aligned} \tag{1.42}$$

Back to our equation (1.40):

$$\begin{aligned} \dot{f}(t) &\leq -\sum_{s\in S, \bar{N}_s>0} \rho_s\delta\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &= -\sum_{s\in S} \rho_s\delta\partial_2 U_s(\bar{N}_s(t), \rho_s(1+\delta)) \\ &\leq -\sum_{s\in S} \lambda_s\delta\underline{w}(\bar{N}_s(t)) \\ &\leq -(\min_{s\in S}\lambda_s)\delta\sum_{s\in S}\underline{w}(\bar{N}_s(t)) \\ &\leq -(\min_{s\in S}\lambda_s)\delta\underline{w}(\max_{s\in S}\bar{N}_s(t)) \\ &\leq -(\min_{s\in S}\lambda_s)\delta\underline{w}(\frac{\|\bar{N}(t)\|}{|S|}) \\ &= -w_3(\|\bar{N}(t)\|) \end{aligned} \tag{1.43}$$

Hence, we have prove the Lyapunov condition and conclude the stability of the fluid network model under normal offered load condition. □

At this point, we have proved the stability of the fluid model. In the next section, we will establish the relationship of the fluid model and the original network.

Remarks:

1. One alternative way to forumlate the problem is to consider the workload along the routes rather than at the source. However, the proof given here will not be applicable in this case. In this case, our Lyapunov function will be

$$\begin{aligned} f(t) = \sum_{s \in S} \int_0^{\bar{N}_{s,r}(t)} & v_s \partial_2 U_{s,r}(y, \rho_{s,r}(1+\delta))dy \\ \text{such that } \bar{N}_s(t) = \textstyle\sum_{r \in s} & \bar{N}_{s,r}(t) \text{ for some } \bar{N}_{s,r}(t) \geq 0 \end{aligned} \tag{1.44}$$

where δ is sufficiently small so that $\{\rho_{s,r}(1+\delta), s \in S\}$ still satisfies the normal offered load condition (1.11) with $\rho_{s,r}$ replaced by $\rho_{s,r}(1+\delta)$. In the proof, we require $\rho_{s,r}(1+\delta)$ to be feasible in the region for the optimization problem (1.41). But for $\delta > 0$,

$$\begin{aligned} & \sum_{l \in r, r \in s, \bar{N}_s > 0, s \in S} \Lambda_{s,r}(1+\delta) \leq C_l \text{ for } l \in L, \\ & \Lambda_s \geq 0 \text{ where } \Lambda_s = \textstyle\sum_{r \in s} \Lambda_{s,r}(1+\delta) \text{ for some } \Lambda_{s,r} \geq 0 \\ \Longleftrightarrow \quad & \sum_{l \in r, r \in s, \bar{N}_s > 0, s \in S} \Lambda'_{s,r} \leq C_l \text{ for } l \in L, \\ & \Lambda_s \geq 0 \text{ where } \Lambda_s = \textstyle\sum_{r \in s} \Lambda'_{s,r} \text{ for some } \Lambda'_{s,r} > 0 \end{aligned} \tag{1.45}$$

Hence, we cannot deduce that $\rho_{s,r}(1+\delta)$ is feasible in the region for the optimization problem (1.41).

2. To further generalize the result, note that no assumption is made on the feasible region:

$$\sum_{s\in S, l\in r, r\in s} \rho_{s,r} < C_l, \text{for } l \in L \tag{1.46}$$

for some $\rho_{s,r} \geq 0$ where $\rho_s = \sum_{r\in s} \rho_{s,r}$

Hence, we can generalize to any convex feasible region and the result will still hold.

1.4.2 Use of Fluid Network Model to prove Theorem 1

In this section, we will prove theorem 1. We first prove the convergence of a sequence of networks to a fluid solution of the fluid model. Given the convergence, we can work on the fluid solution. Using theorem 3 of Bramson (1998) and the results from proposition 1 and proposition 2, we can derive theorem 1.

Consider a sequence of such stationary models, indexed by $k = 1, 2, ...$ The superscript k described the kth model. Specially, for the k-th model, we have $\bar{X}^{(k)}(t)$, $\bar{N}^{(k)}(t)$, $\bar{E}^{(k)}(t)$, $\bar{A}^{(k)}(t)$, $\bar{D}^{(k)}(t)$ and $\bar{S}^{(k)}(t)$.

Let $\{z_k\}$ be an increasing sequence of the positive numbers with $z_k \to \infty$ and let

$$\begin{aligned}
\bar{N}_s^{(k)}(t) &= \frac{1}{z_k} N_s^{(k)}(z_k t) \\
\bar{X}_s^{(k)}(t) &= \frac{1}{z_k} X_s^{(k)}(z_k t) \\
\bar{A}_s^{(k)}(t) &= \frac{1}{z_k} A_s^{(k)}(z_k t) \\
\bar{D}_s^{(k)}(t) &= \frac{1}{z_k} D_s^{(k)}(z_k t)
\end{aligned} \tag{1.47}$$

In the fluid approach to show stability, we make use the limits of the processes above. Before presenting the main proposition needed to prove the theorem, we state a lemma which will be used in our proof.

Lemma 2. *(Partial continuity) Suppose a sequence of states $\{n^j, j = 1, 2, ...\} \subset \Re^{|S|}$ converges to $n \in \Re^{|S|}$ as $j \to \infty$. Then for the U-utility allocation policy Λ^U, we have*

$$\Lambda_s^U(n^j) \to \Lambda_s^U(n) \text{ as } j \to \infty \tag{1.48}$$

for any $s \in S$ such that $n_s > 0$.

Proof. Prove by contradiction. Suppose not, then there exist a sequence of states $\{n^j, j = 1, 2, ...\} \subset \Re^{|S|}$ converging to state $n \in \Re^{|S|}$ as $j \to \infty$, and for some $s \in S$ with $n_s > 0$, $\Lambda_s^U(n^j) \to \Lambda^* \neq \Lambda_s^U(n)$.

Define $Y = \{s \in S | n_s > 0\}$ and $Y_j = \{s \in S | n_s^j > 0\}$. Assume that $Y \subset Y_j$ for all $j = 1, 2, ...$

Let $\bar{\Lambda}_s = \Lambda_s^*$ for $s \in Y$ and $\bar{\Lambda}_s = 0$ otherwise. We can check that $\Lambda^U(n)$ and $\bar{\Lambda}$ are unique solution and a feasible solution respectively to optimization problem:

$$\begin{aligned} & \max \sum_s U_s(n_s, \Lambda_s) \\ s.t. \quad & \exists \Lambda_{s,r} \geq 0 \text{ that } \Lambda_s = \sum_{r \in s} \Lambda_{s,r} \text{ for } s \in S \\ & \sum_{s \in S, l \in r, r \in s} \Lambda_{s,r} \leq C_l \text{ for } l \in L \\ & \Lambda_s = 0 \text{ if } n_s = 0 \text{ for } s \in S \\ & \Lambda_s \geq 0 \text{ for } s \in S. \end{aligned} \tag{1.49}$$

Thus, we have

$$\sum_{s\in Y} U_s(n_s, \bar{\Lambda}_s) = \sum_{s\in Y} U_s(n_s, \Lambda_s^*) < \sum_{s\in Y} U_s(n_s, \Lambda_s^U(n)) \tag{1.50}$$

Note that convexity of the feasible region allows us to find a solution to the optimization problem and strict concavity of the objective function allows us to have uniqueness of the solution.

On the other hand, we can verify that $\Lambda^U(n^j)$ and $\Lambda^U(n)$ are the unique solution and feasible solution respectively, to the optimization (1.49) with n replaced n^j, noting that $Y \subset Y_j$ for all j.

Thus we have

$$\sum_{s\in Y} U_s(n_s^j, \Lambda_s^U(n^j)) \geq \sum_{s\in Y} U_s(n_s^j, \Lambda_s^U(n)) \tag{1.51}$$

for all j.

Let $j \to \infty$, from the joint continuity of $U_s, s \in S$, we have

$$\sum_{s\in Y} U_s(n_s, \bar{\Lambda}_s) \geq \sum_{s\in Y} U_s(n_s, \Lambda_s^U(n)) \tag{1.52}$$

since $n_s = 0$ for $s \in S\backslash Y$ and $U_s(0, .) = 0$ for $s \in S$. And hence contradiction. □

Remarks:

1. Note that the proof relies on the strict concavity of the objective function, U. Otherwise, the limit of convergence, $\Lambda_s(n)$ will not be unique.

2. The proof for the lemma is similar to Ye,Ou and Yuan (2003) because the strict concavity of the objective function is retained.

Proposition 2. *Given the allocation $\Lambda(.)$, and suppose that $\bar{X}^{(k)}(0)$ converges as $k \to \infty$. Then for almost all sample paths and any subsequence of $\{k\}$, there exists a further subsequence, also denoted by $\{k\}$, such that*

$$(\bar{N}^{(k)}(t), \bar{X}^{(k)}(t), \bar{A}^{(k)}(t), \bar{D}^{(k)}(t)) \to (\bar{N}(t), \bar{X}(t), \bar{A}(t), \bar{D}(t)) \text{ } u.o.c. \tag{1.53}$$

and $(\bar{N}(t), \bar{X}(t), \bar{A}(t), \bar{D}(t))$ is a fluid solution to the fluid network model.

Proof: The u.o.c. convergence of $\bar{A}^{(k)}(t)$ follows from the functional strong law of large numbers.

Note that at source s, we have:

$$0 \leq \bar{D}_s^{(k)}(t) \leq \min_{(s,r)}\{\sum_{l \in r, r \in s} C_l\}(t) \tag{1.54}$$

This is so because the capacity allocation process at source s is bounded above by the minimum of the set of links connected to source s via route (s, r). Hence the processes are pointwise bounded. Since

$$|\bar{D}_s^{(k)}(t) - \bar{D}_s^{(k)}(w)| \leq \min_{(s,r)}\{\sum_{l \in r, r \in s} C_l\}|t - w|, \tag{1.55}$$

the processes are also equicontinuous. Thus we deduce that the scaled processes $\bar{D}^{(k)}(t)$ are u.o.c. convergent. Consequently, convergence of $\bar{X}^{(k)}(t)$ follows from (1.26). Using the similar arguement, theorem 6.5 of Chen and Yao (2001) yields the u.o.c. convergence of

$\bar{N}^{(k)}(t)$. (1.27) can also be proven by similar arguement to the corresponding part in theorem 6.5 of Chen and Yao (2001).

We now verify (1.29). It suffices to show that

$$\dot{\bar{D}}_s(t) = \dot{\bar{A}}_s(t)(= \rho_s), \text{ if } \bar{N}_s(t) = 0 \tag{1.56}$$

and

$$\dot{\bar{D}}_s(t) = \bar{\Lambda}_s(\bar{N}(t), \dot{\bar{A}}_s(t)) = \Lambda_s(\bar{N}(t)), \text{ if } \bar{N}_s(t) > 0 \tag{1.57}$$

for any $t \geq 0$ such that all the related processes are differentiable.

When $\bar{N}_s(t) = 0$ or $\bar{X}_s(t) = 0$, we have $\dot{\bar{N}}_s(t) = 0$ since t is a local minimum of the function $\bar{X}_s(.)$. Thus (1.56) follows.

Consider the case $\bar{N}_s(t) > 0$. For any small positive h and a sufficiently large index k along a convergent subsequence, we have:

$$\begin{aligned}
& |\frac{1}{h}(\bar{D}_s^k(t+h) - \bar{D}_s^k(t)) - \Lambda_s(\bar{N}(t))| \\
= \; & |\frac{1}{h}\int_0^h \Lambda_s(N(z_k(\tau+t))) - \Lambda_s(\bar{N}(t)) d\tau| \\
\leq \; & \frac{1}{h}\int_0^h |\Lambda_s(N(z_k(\tau+t))) - \Lambda_s(\bar{N}(t))| d\tau \\
\leq \; & \frac{1}{h}\int_0^h |\Lambda_s(N(z_k(\tau+t))) - \Lambda_s(\bar{N}(t+\tau))| d\tau \\
& + \frac{1}{h}\int_0^t |\Lambda_s(\bar{N}(\tau+t)) - \Lambda_s(\bar{N}(t))| d\tau
\end{aligned} \tag{1.58}$$

Using the partial radial homogeneity and partial continuity of Λ_s and the convergence of a subsequence $\bar{N}^{(k)}(.)$, we get

$$\begin{aligned} &\Lambda_s(N(z_k(\tau+t))) - \Lambda_s(\bar{N}(t+s)) \\ &= \Lambda_s(\bar{N}^{(k)}(\tau+t)) - \Lambda_s(\bar{N}(t+s)) \\ &\to 0 \end{aligned} \tag{1.59}$$

as $k \to \infty$, for all $\tau \in (0,h)$.

Thus let $k \to \infty$ in (1.58), we have:

$$\begin{aligned} & |\frac{1}{h}(\bar{D}_s^k(t+h) - \bar{D}_s^k(t)) - \Lambda_s(\bar{N}(t))| \\ \leq & \frac{1}{h}\int_0^t |\Lambda_s(\bar{N}(\tau+t)) - \Lambda_s(\bar{N}(t))| d\tau \end{aligned} \tag{1.60}$$

By the Lipschitz continuity of $\bar{N}(t)$ and the partial continuity property of $\Lambda^U(\cdot)$, right hand side of the above inequality tends to 0 as $h \to 0_+$.

Hence $\dot{\bar{D}}_s(t+) = \bar{\Lambda}_s(t)$ and thus $\dot{\bar{D}}_s(t) = \bar{\Lambda}_s(t)$ a.e..

□

The proof of theorem 1 follows from proposition 1, proposition 2 and theorem 3 of Bramson (1998).

Having proved theorem 1, we have established the necessary and sufficient conditions for the stability of the network under some allocation, in particular, the U-utility maximizing allocation.

1.5 Conclusion

In this chapter, we are concerned with the stability of a stochastic network with a wide range of allocation policies. The policy is given in terms of solving an optimization problem with a unique objective function to represent a certain allocation policy. Given the wide collection of such policies, we seek to unify all the policies and use a U-utility allocation policy to represent the policies, and carry our analysis on it. The partial radial homogeneity property of the utility function is further generalized for easy identification of allocation policies.

The main result is to extend the result for the stochastic network with fixed routing to the case of routing. This enlarges the solution space of the problem. In order to work on such stochastic network (with or without routing), in particular the stationary network, we make the assumption that the processing workload is i.i.d. exponential. This serves to ease the technical difficulty in proving the stability of the network. The case for the bursty network model is presented in the appendix. The defined stationary network in this chapter is a PS system, but our analysis are based on a HOLPS system, which is sufficient for proving the positive Harris recurrence of the PS system.

The network models we introduce in the paper view the network at the higher level, and focus on the dynamic nature of work traffic in the network, but ignore some details on how the jobs are established and stabilized at the each instance. We proved the stability of the network via a fluid limit approach. The technique of Lynapunov function is used to help us in the proof for the stability.

1.6 Appendix: Bursty Network Model

In the bursty model (see Cruz (1991a,b)), work loads are injected by "an adversary" to the network, such that the arrivals do not depend on any particular probability assumptions. The arrivals to different routes may even be correlated. The results based on this is more robust in that they do not depend on any particular probability assumption. However, it only serves as an approximation to the network. Given this, the stationary model will not be good enough to describe the network dynamics of this network.

To overcome this, we consider a specific path realization of the network, and make use of three deterministic processes:

1. queue length process : $X(t) = \{X_s(t) : s \in S\}$
2. workload arrival process :$A(t) = \{A_s(t) : s \in S\}$
3. capacity allocation process : $D(t) = \{D_s(t) : s \in S\}$

The meaning of the above processes is similar to that of the stationary model except for some slight technical differences. In the bursty model, we also have the following flow balance equation:

$$X_s(t) = X_s(0) + A_s(t) - D_s(t) \text{ for } t \geq 0, s \in S \tag{1.61}$$

and constraints (1.21) to (1.23).

To relate the bursty model to the stationary model, we define an ongoing job process $N(t) = \{N_s(t) : s \in S\}$ with $N_s(t)$ given as:

$$N_s(t) = \frac{X_s(t)}{\nu_s}, \text{for } t \geq 0, s \in S, \tag{1.62}$$

which approximately represents the number of jobs on source s. The approximation can be justified for network with large number of arrivals and that the average workload on route r is ν_s.

We adopt the input model by Cruz (1991a) to capture the burstiness of inputs into the network. We assume that the arrival $A_s(t)$ at a source s need not have regularity properties like continuity or differentiability, but just follow the following bursty constraint:

$$0 \leq A_s(t_1) - A_s(t_2) \leq \rho_s(t_1 - t_2) + w, \text{ for } t_1 \geq t_2 \geq 0, s \in S \tag{1.63}$$

where ρ_s is a constant in units of work amount per unit time, and w is a constant in units of work amount. (This constraint is further explored by Borodin, et al. (2001).) Cruz (1991a) defines constraint (1.63) as a (w, ρ_s) regulator that controls the rate of a particular arrival session via a path in the network so that during any time interval $[t_1, t_2]$, the arrival traffic is bounded by $\rho_s(t_1 - t_2) + w$ of traffic. Such an arrival process can be viewed as the output of a "leaky bucket model" of a flow control, where input is rejected whenever the constraint (1.63) is violated.

Suppose the chosen state dependent allocation policy at source s is Λ_s. Then we have

$$D_s(t) = \sum_{r \in s} \int_0^t \Lambda_{s,r}(N(\tau)) d\tau, \text{for some } \Lambda_{s,r} \text{ and } t \geq 0, s \in S \tag{1.64}$$

where $\Lambda(N(t)) = \{\Lambda_s(N(t)) : s \in S\}$ with $\Lambda_s(N(t))$ being the bandwidth or transmission rate allocated to source s at time t when the network state is $n = N(t)$.

Thus, our bursty model can be represented by the octuple $(L, C, R, M, \lambda, \nu, \rho, w)$. With $A(t)$ and $\Lambda(.)$ specified, the network dynamics is characterized by (1.61), (1.62), (1.63) and (1.64).

Definition 1. *The bursty model is said to be stable if for any $X(0)$, there exists a constant $M_{X(0)}$ such that*

$$\sup_{t \geq 0} \|X(t)\| \leq M_{X(0)} \tag{1.65}$$

One must be careful when we use the bursty model. The constraint (1.63) is an approximation of the actual number of workload. Hence, we should bear in mind that the bursty model only serves as the workload approximation of the actual model. The necessary condition for the stability of the bursty model is the same as that of the stationary model, which is the capacity constraint. One can construct examples to illustrate this.

Theorem 2. *Suppose the normal offered load condition is satisfied for the bursty network model* $(L, C, R, M, \lambda, \nu)$. *Then the allocations* $\Lambda^{pp}, \Lambda^{pd}, \Lambda^{mm}, \Lambda^{\alpha}$ *and* Λ^{U} *ensure the stability of the model.*

1.6.1 Use of Fluid Network to prove Theorem 2

The steps involved here are similar to that for the stationary model. The difference is that in the bursty model, we need to scale the process and take the proper limits.

Let $\{t_k\}$ and $\{z_k\}$ be increasing sequences of the numbers with $t_k \to \infty$ and $z_k \to \infty$ respectively. t_k represent a sequence of times and z_k represent a sequence of positive numbers. Define:

$$\begin{aligned}
\bar{N}^{(k)}(t) &= \frac{1}{z_k} N(z_k t + t_k) \\
\bar{X}^{(k)}(t) &= \frac{1}{z_k} X(z_k t + t_k) \\
\bar{A}^{(k)}(t) &= \frac{1}{z_k} (A(z_k t + t_k) - A(t_k)) \\
\bar{D}^{(k)}(t) &= \frac{1}{z_k} (D(z_k t + t_k) - D(t_k))
\end{aligned} \tag{1.66}$$

We have the following proposition on the limits of these scaled processes.

Proposition 3. *Given the bandwidth allocation* $\Lambda^U(.)$*, and suppose that* $\{\bar{X}(0)\}$ *(or* $\{\frac{1}{z_k}X(t_k)\}$*) has convergent subsequences. Then there exists a subsequence of* $\{k\}$*, such that*

$$(\bar{N}^{(k)}(t), \bar{X}^{(k)}(t), \bar{A}^{(k)}(t), \bar{D}^{(k)}(t)) \to (\bar{N}(t), \bar{X}(t), \bar{A}(t), \bar{D}(t)) u.o.c. \tag{1.67}$$

and $(\bar{N}(t), \bar{X}(t), \bar{A}(t), \bar{D}(t))$ *is a fluid solution to the fluid network model.*

Proof. We give an outline of the proof. The u.o.c. convergence of the scaled processes $\bar{A}^{(k)}(t)$ and $\bar{D}^{(k)}(t)$ along a subsequence to $\bar{A}(t)$ and $\bar{D}(t)$ is a direct result from lemma 6.3 of Ye, *et al.*(2005). Hence, u.o.c. convergence of $\bar{N}^{(k)}(t)$ and $\bar{X}^{(k)}(t)$ follows consequently from (1.26) to (1.33). The case of u.o.c. convergence of $\bar{D}^{(k)}(t)$ is proved exactly the same as the stationary case, but with a change in time scale. i.e. $z_kt + t_k$ instead of z_kt. □

Hence, from the stability of the fluid model and proposition (3), we have prove theorem 2. (See Ye *et al.* (2005) for more details.)

Chapter 2

Control in Stochastic Networks with Concurrent Resource Occupancy and Batch Arrivals

This chapter is motivated by the work of Li and Yao (2004), in which a booking limit control policy based on a fixed point approximation was developed for a network with concurrent resources. It builds on their model and further generalize the conditions required for the model. There are many applications to this. We chose the airline industry to present our ideas. The objective is to optimize the expected revenue subjected to the availability of seats on the flights. In our work, we further generalize the arrival process to a batch arrival process. Our solving methodology involves deriving a fixed point approximation to express the network operating under a set of booking limits, and reformulating it into a linear program to solve for the booking limits. We show that the policy is optimal under certain limit. In order to show the accuracy of our approximation, we carry out extensive simulation studies. Another contribution made is to study the updating mechanism for the booking limit, which turns the originally static policy to a dynamic one. Numerical analysis demonstrates significant improvement of dynamic policy.

2.1 Introduction

We have seen that concurrent resource occupancy is a prevalent feature in many engineering and service systems. A multi-leg airline booking seat allocation, which involves concurrent seat reservations on several connecting flights, is one example. In the shipping, one has to ensure the concurrent availability of empty containers, the lift capacity at the terminal, and the space slots for containers on board the vessel. Yet another example involves hotel occupancy. A customer often books a room for a certain time duration and that room is blocked off from sale during that duration. Other examples include a make-to-order processing company, which requires concurrent processing of all its components, and file transfer on the internet, which involves the utilization of bandwidths on all links along its route from the source to the destination. These are real-life situations and the development of a robust stochastic network to accommodate concurrent resource occupancy is apposite. The stochastic and dynamic nature of the demand process makes such problems challenging to solve. To address this, we derive a control policy for these problems, and this forms the motivation of this chapter.

In this part of the thesis, we develop a static stochastic model with concurrent resources with an appropriate control policy. The idea of concurrent resource occupancy was called to attention by Whitt (1985) who mooted the notion of concurrent resource occupancy when he studied the blocking phenomenon in loss networks. Kelly (1988) went further and studied the behaviour of large loss network using a fixed point approximation for blocking probabilities. In addition, he examined properties of fixed point mappings in these loss networks. Kelly's work focuses on applications concerning circuit-switched communication networks. One of the similarities is the characteristic of concurrent occupancy of more than one type of resources. Our model of interest is the logistics of a network model that represents the airline transportation hub which displays the similar feature.

Online airline seat reservation is now a common feature. As such, this makes our problem more significant since we want to develop a robust stochastic network that can benefit the airlines. If one books a flight from Singapore to Hong Kong via Bangkok, the reservation system will process the seat reservation from Singapore to Bangkok and then from Bangkok to Hong Kong. This is a clearly a case of concurrent resource allocation, and the reservation process affects the concurrent allocation of resource almost instantaneously. This underlies the need for a computationally fast method to compute the control policy.

A fixed point approximation scheme is proposed to address the problem. Li and Yao (2004) studied a stochastic network with simultaneous resource occupancy and introduced a threshold control policy based on a fixed point approximation. Ye and Yao (2006) extended Li and Yao by establishing the asymptotic optimality of the control policy under fluid and diffusion scaling. Li and Yao optimize the revenue for flows through the network by resource control (i.e., developing policies for resource allocation amongst the various job classes). Our aim is to generalize the results of their paper. Though the model can be applied to a wide range of scenarios, we use the airline seat reservation system to present our model and its accompanying algorithm.

To generalize the model, we introduce batch arrival process rather than single arrival process. The underlying rationale is that customers book more than one seat. As such, a Poisson process for the arrival will not capture this feature accurately. A travel agency books in batches, and each batch depends on the number of customers who purchase their package. Instead of a Poisson process to represent the customer arrival, we assume that the arrivals follow a compound Poisson process. The main difficulty lies with the distribution of the batch size. Assuming that the batch sizes are independent and identical random variables, we can then generalize the results by Li and Yao (2004) and demonstrate that the single batch arrival studied is a special case. However, the analysis pertaining to a general batch size is much more complicated. Assuming a discrete distribution for the batch size proved to be computationally difficult to solve. Hence, we seek an approximation using a continuous distribution, and claims that the approximation is accurate at least in a limiting regime. Although our results focus on the limiting regime, it also suggests that the policies derived provides good results when implemented on a system with large mean demand and large capacity of the resources. This forms the main contribution.

Static policies are easy to implement but at the expense of the performance of the policy. Dynamic policy are known to be more accurate if applied in a “smart” way but the explosion of dimension of the problem due to updating of the state of the system prevents one from implementing it successfully, especially so in large-scale network problems. Our derived control policy is a static one. We introduce a booking limit policy to set a fixed threshold on the utilization of the resources in the network. It will be interesting to extend our static policy to a dynamic policy, one which includes the updating of the current state of the system. We analyze a few updating strategies numerically, and proposed a updating mechanism. By simulation, the performance of our updating policy is tested numerically against the hindsight optimum and the static case.

Outline

In this section, we formulate a revenue optimization problem with the concurrent resource occupancy and compound Poisson arrival process. We then derive a control policy with the use of a booking limit. We introduce a fixed point approximation for a network operating under a set of thresholds that control the access of jobs from each class. Solving it is computationally difficult. We therefore propose an approximation via a continuous distribution. In particular, we use the standard normal distribution to approximate the general batch. We justify such an approximation and claim that it is accurate at least in the asymptotic sense. We outline the proof for the asymptotic optimality of the method in the appendix.

Finally, we conclude the paper by demonstrating the results using some numerical examples. In reality, the actual distribution of the batches is unknown. Given this, it is interesting to see what impact the distribution assumed for the batch has on the revenue generated by the network. We investigate this further numerically.

A possible extension to our control policy is to introduce a new booking limit during the time horizon via some effective method of updating the system. We suggest a updating mechanism and show the effectiveness of it numerically.

2.2 Literature Review

Control or management policies are closely related to revenue management in airline, hotel and manufacturing industries. McGill and Van Ryzin (1999) provided a comprehensive survey of the research done on revenue or yield management. It is clear that revenue management will continue to interest academics because it is a rich area for research and for practitioners because they can learn how to increase their corporate revenue.

Adopting a sound control policy is important. Simulation results have demonstrated gains from using proper control policies. While the precise improvement is dependent on factors like the type of network, the demand distribution, the load factor, etc., improvements of up to 1.5% under moderate load conditions and up to 3% or more under high load conditions can be expected (Talluri and van Ryzin (1999)). These gains make the implementation of such policies relevant in practical applications. While the benefits of such policies are high, the complexity of the implementation is often difficult given that the actual distribution of the demand is unknown. The policies can be derived in two main ways, namely, using a static model or a dynamic model. It is impossible to list all the papers done till date. Hence, we list the more relevant ones in this section.

2.2.1 Dynamic Model

A control policy derived from a dynamic model takes into account the state of the process throughout the planning horizon. The downside is that it is usually computationally expensive to derive a control policy using the dynamic model.

Peng (1999) considered a multiple booking class airline-seat inventory control problem that takes into account either a single flight leg or a multiple flight-leg case. Peng formulated a dynamic model, in which the demand for the arrival is modeled as a discrete time stochastic process. The computation time of the dynamic model was reduced by consolidating the decision-making period into sets of critical decision periods. Another example of dynamic programming (Markov decision process) being applied to revenue management is Subramanian, Stidham and Lautenbacher (1999). They analyzed the airline seat allocation on a single-leg flight multiple fare class using a Markov decision process (dynamic programming) model. Contributing factors such as cancellation, no-shows and overbooking are taken into consideration for the control policy. There are many works that focus on a Markov-decision-process-type model. Lee and Hersh (1993) studied such model based on a single-leg flight, and uses discrete time dynamic programming to develop the optimal policy. Other works include Lautenbacher and Stidham (1999), Liang (1999) and Zhao and Zheng (2001). Most of the work focused on a single-leg scenario and provided a deep insight into the problem. However, practical problems often involve a network of flight legs, and this makes such implementation almost impossible due to an increase in dimension of the problem when applying dynamic programming approach.

An interesting approach using dynamic programming is the use of the bid price control, which is a method that sets a control policy based on the dual prices in the model. This technique was first suggested by Simpson (1989) and was later studied by Williamson (1992). Talluri and van Ryzin (2004) analyzed the use of bid-price as a form of threshold control. In short, using this approach, a class of customer is accepted only if its fare exceeds the sum of the bid price along the affected route. Their model takes account the demand uncertainty and allows the flexibility of random prices set within a fare class. The problem was analyzed via a dynamic programming approach. Using their general model, they proved that the bid-price control was suboptimal and investigated further the conditions that the bid-price scheme would fail to produce the right solution. They were able to prove that when the capacities and the arrival process were scaled by the same factor, the bid price control was shown to be optimal in the asymptotic sense. Feng and Gallego (2000) studied the problem of a fixed number of items to be sold over a finite horizon and used pricing as a tool to maximize the expected revenue. Extensions of Feng and Gallego's work involving dynamic pricing were carried out by Feng and Xiao (2000a) and Feng and Xiao (2000b).

The dynamic model is more accurate in that it takes into account the state of the problem at that time interval into consideration. However, a weakness of the dynamic approach is that it is computationally expensive to compute the optimal policy. This motivates us to take one step back to find some static control policies that are computationally feasible and yet yield good performance. Next, we summarize the literatures on the static model.

2.2.2 Static Model

In a static model, we treat the booking period as a single interval, and the aim is to set a booking limit for every booking class at the start of the booking process. Most static policies are derived by solving a deterministic version of the actual problem. The weakness in such an approach is that it is unable to consider the actual state during the process. Hence, the tradeoff is the accuracy of the policy. In contrast to the dynamic approach, it is generally computationally cheaper to compute the optimal solution, and large problems can be handled more readily. Hence, it is possible to compute a control policy for a general network. The control policy derived from static policies are simpler when compared to their dynamic counterparts. However, results have shown that such static policies are asymptotically optimal, in the sense that the policy is close to being an optimal one under appropriate scaling.

Glover et al. (1982) were amongst the first to study a deterministic network flow. They formulated the problem as a minimum cost (maximum profit) network flow problem to find the optimal allocation of seats between passenger itineraries and fare classes. The execution time on a 16bit microcomputer used then was linearly proportional to the number of arcs and nodes in the network. They kept computation cost at a manageable level. This renders further analysis on the system, which would otherwise be impossible on a dynamic policy, manageable.

Cooper (2002) studied similar problems in a stochastic framework, and derived a management policy from a deterministic optimization problem, which is asymptotically optimal in the sense that the normalized optimal revenue converges in distribution to a constant upper bound. Their focus centers around the use of a LP-based allocation to generate a simple policy which produces good results. Cooper made a counter-intuitive observation that a standard updating practice may not necessarily brings better performance than the case of no updating. This suggests that much research has to be put into this area to find a more effective method of introducing updating of the system.

Gallego and Van Ryzin (1994) considered varying the price to maximize the expected revenue. They computed an upper bound for the expected revenue based on a deterministic version of the problem. They proved that a static fixed price policy is asymptotically optimal as the volume of expected sales gets larger. In fact, a simple fixed price policy works well in many situations. This is encouraging since optimal dynamic pricing policies are difficult to implement since they require more monitoring and adjusting according to the state of the system.

A deterministic network model based on origin-destination pairs is studied by Talluri (1993). His model was based on the airline revenue management problem, that manages passenger routing using seat inventory control. Most revenue management models emphasize sales control of low fare class on a high demand situation. Talluri worked on the low demand alternative route to increase the expected revenue further. Numerical studies in large scale problems have showed improved revenue without sacrificing the service level.

The computational cost is also kept at a manageable level. Another work which is based on a deterministic network flow is by Gallego and Van Ryzin (1997). They investigated the problem of pricing finished products in a firm, and aim to maximize total expected revenue over a finite sales horizon. The problem was analyzed via deterministic version of the problem. The heuristics proposed were shown to be asymptotically optimal as the expected sales volume increases.

Most of the proposed static policies are asymptotically optimal on the fluid scale. Reiman and Wang (2006) studied a control policy for a revenue management problem in a network setting and proposed an accurate policy which is optimal on the more sensitive diffusion scale. Their policy consists of two stages and it is a mix of a static and dynamic policies. In the first stage, it begins by solving the deterministic version of the problem and using the computed results to form a probabilistic admission rule. In the second stage, the actual state of the system is tracked by a trigger function, which is the difference between the actual realized acceptance and the expected customer acceptance. Certain thresholds are defined and, if violated, they trigger a reoptimization of the problem with the parameters updated. Hence, the second stage is a dynamic one. They demonstrated that their policy is optimal on the diffusion scale.

Li and Yao (2004) proposed the policy. Ye and Yao (2006) studied the asymptotic optimality of the policy under both fluid and diffusion scaling. The class of the routes are categorized into two groups, namely, the H-class (the high revenue or premium class) and the L-class (low revenue or lower priority). They managed to show the results derived from

Li and Yao (2004) are asymptotically optimal on the diffusion scale, but required a condition that states that each bottleneck link in the system contains at most one H-class route.

2.3 Introduction of Network

Suppose a network consisting of a set of links, L, with each link connecting a pair of nodes. Each link $l \in L$ has capacity limit C_l.

Let R denote the set of routes/classes.(The term route and class is used interchangeably here, but this should not cause any ambiguity in the presentation.) Each route $r \in R$ is a subset of links connecting a source node to a sink node. Denote $l \in r$ if link l is part of route r. In the airline revenue problem, we can interpret R as the classes of customers or the set of routes. Hence, if a class r_1 customer uses link l_1, l_3 and l_5, $r_1 = \{l_1, l_3, l_5\}$. For the rest of the paper, we will use the route and the class of customer interchangeably and this should not cause any confusion. For a more general framework, we assume the demand on each route r follow a Compound Poisson process with rate independent of the demand on all other routes.

2.3.1 Revenue Management problem

Our main objective is to maximize the expected revenue over a finite time horizon. Let w_r be the price charged to each route r, and let A_r denote the number of orders that are accepted during the time horizon. Consider a time period of 1. To motivate, consider the following revenue optimization problem:

$$
\begin{aligned}
&\max_{A_r} \sum_{r \in R} w_r A_r \\
s.t. \quad &\sum_{s \neq r, s \ni k} A_s \leq C_l \text{ for } l \in r, r \in R \\
&0 \leq A_r \leq N_r(1) \qquad (2.1)
\end{aligned}
$$

where $N_r(t)$ denotes the number of class r arrivals in the period $[0, t]$. The optimal solution to (2.1) is defined to be the hindsight optimum. Solving for the hindsight optimal solution is impossible because we do not know the number of class arrival beforehand. Even the most sophisticated method for forecast will not allow us to predict the exact number of the class arrival. Hence, in a real life application, we can at best seek an approximation to it. However, we can treat the hindsight optimum as a benchmark by which we judge the accuracy of our approximation scheme.

The main focus is to construct a control policy on the arrival with the objective of maximizing expected revenue. Our control policy involves setting a threshold ("booking limit") y_r on the number of orders that will be accepted on each route r for a given time horizon. In the planning process, we have to optimize the decision variables y_r at time zero, using the model formulation as proposed. Once the decision variables are decided, the orders from all the routes are accepted on a first-come-first-served basis. The acceptance of the orders from the route is determined purely by the booking limit and the capacity available. In short, an order for a route r will be rejected once y_r is reached or the capacity of the links affected are all utilized, whichever happens first.

Let the arrival process $\{N_r(t), t \geq 0\}$ denote the compound Poisson process. Thus we have

$$N_r(t) = \sum_{i=1}^{\hat{N}_r(t)} B_{i,r} \tag{2.2}$$

where $\{\hat{N}_r(t), t \geq 0\}$ is a Poisson process with rate λ_r and $\{B_{i,r}, i \geq 1\}$ be the independent and identically distributed random variables denoting the batch size of the arrival, which is independent of $\{\hat{N}_r(t).t \geq 0\}$. We can interpret $N_r(t)$ as the total number of orders requesting route r up to t. To ease the notations, let N_r denote the total number of arrivals for the indented time horizon.

At a particular route r, the sum of the number of orders accepted by all the class must be less than the capacity of the links affected. Hence, we have the constraint:

$$A_r \leq C_l - \sum_{s \neq r, s \ni l} A_s, \forall l \in r \tag{2.3}$$

Therefore, the number of accepted orders is the minimum value of the booking limit, the arriving process and the remaining capacity of the link affected. Hence, the fixed point model representing this is

$$E[A_r] = E[N_r \wedge y_r \wedge \min_{l \in r}\{C_l - \sum_{s \neq r, s \ni l} A_s\}] \tag{2.4}$$

which says that the expected value of the accepted orders for route r is the expected value of the minimum of the arrival process, the booking limit and the remaining capacity available on the link which route r uses.

Expression (2.4) is difficult to solve, since there is random variables present on both side of the equation. An approximation for $E(A_r)$ was proposed by Li and Yao (2004) and it is as follows:

$$E(A_r) \approx E[N_r \wedge y_r \wedge \min_{l \in r}\{C_l - \sum_{s \neq r, s \ni l} E(A_s)\}] \tag{2.5}$$

Denote $x_r = E[A_r]$ and $m_r = \min_{l \in r}\{C_l - \sum_{s \neq r, s \ni l} x_s\}$. We let $N(\lambda, \hat{N})$ denote a compound Poisson variate where $\hat{N}$ denotes the Poisson variate with mean λ and i.i.d. batch sizes $B_{i,r}$. We denote $h(\lambda, n) = E[N(\lambda, \hat{N}) \wedge n]$. By conditioning on the Poisson process for the arrival, we have:

$$\begin{aligned}
& h(\lambda, n) \\
= \quad & E[N(\lambda, \hat{N}) \wedge n] \\
= \quad & n - E[n - N(\lambda, \hat{N})]^+ \\
= \quad & n - \sum_j E[n - N(\lambda, \hat{N}) | \hat{N} = j]^+ P(\hat{N} = j)
\end{aligned} \tag{2.6}$$

At this point, we assume that the batch size follow some distribution. Under the approximation (2.5) and using the function h derived from (2.6), we have

$$x_r = E[A_r] = h(\lambda, y_r \wedge m_r), r \in R \tag{2.7}$$

Hence, given y_r, the system of equations from (2.7) defined a fixed point model to approximate the expected number of expected orders on each route r. In order to compute (2.6) and (2.7), we need to assume a discrete distribution for the batch size. We present some of the common distributions for the batch size in the following section.

2.3.2 Assuming some common distribution for the batch

By assuming a distribution for the batch, we can compute (2.6). The batch size can be categorized into a single batch case and the non-single batch case.

Single batch Poisson arrival

A particular case of distribution for the batch size is assuming that all $B_{i,r}$ are constant. In particular, let $B_{i,r} = 1$ for all i. This is exactly the standard Poisson process for the arrival. From (2.6), we have

$$\begin{aligned} & h(\lambda, n) \\ = \;& n - E[n - N(\lambda, j)]^+ \\ = \;& n - \sum_{j \geq 0} E[n - N(\lambda, \hat{N}) | \hat{N} = j]^+ P(\hat{N} = j) \\ = \;& n - \sum_{j \geq 0} \sum_{k=0}^{n} (n-k) P[\sum_{i=1}^{j} B_{i,r} = k] e^{-\lambda} \frac{\lambda^j}{j!} \end{aligned} \tag{2.8}$$

From above, we see that $P[\sum_{i=1}^{j} B_{i,r} = k] = 1$ if $j = k$ and $P[\sum_{i=1}^{j} B_{i,r} = k] = 0$ otherwise. Thus, we can simplify it to:

$$\begin{aligned} & h(\lambda, n) \\ = \;& n - \sum_{j \geq 0} (n-j) e^{-\lambda} \frac{\lambda^j}{j!} \\ = \;& n - \sum_{j=0}^{n} (n-j) e^{-\lambda} \frac{\lambda^j}{j!} \end{aligned} \tag{2.9}$$

which is exactly the form given in Li and Yao (2004). Hence, the formulation given by them is a special case of our formulation. In the next section, we will discuss on the more general case, the non-single batch size.

Non-single batch size

In this section, we will introduce the formulation of the non-single batch size. Having a general distribution for the batch size, $B_{i,r}$ makes the result more interesting, but complicates the formulation. Hence we give the formulation of the case when batch sizes are Poisson random variables and Geometric random variables. In this section, to simply our formulations, we drop the subscript r and we consider the batch size at a route r.

Poisson batch size

A more general assumption is that the batch sizes, B_i are all independent Poisson random variables with mean β for all i. i.e. $P(B_i = k) = e^{-\beta}\beta^k/k!$.

We know that$\sum_{i=1}^{n} B_i$ is Poisson distributed with mean $n\beta$. Hence

$$\begin{aligned} & E[n - N(\lambda, \hat{N})|\hat{N} = j]^+ \\ = \;& E[n - N(\lambda, j)]^+ \\ = \;& \sum_{k=0}^{n}(n-k)P[\sum_{i=1}^{j} B_i = k] \\ = \;& \sum_{k=0}^{n}(n-k)e^{-j\beta}\frac{(j\beta)^k}{k!} \end{aligned} \tag{2.10}$$

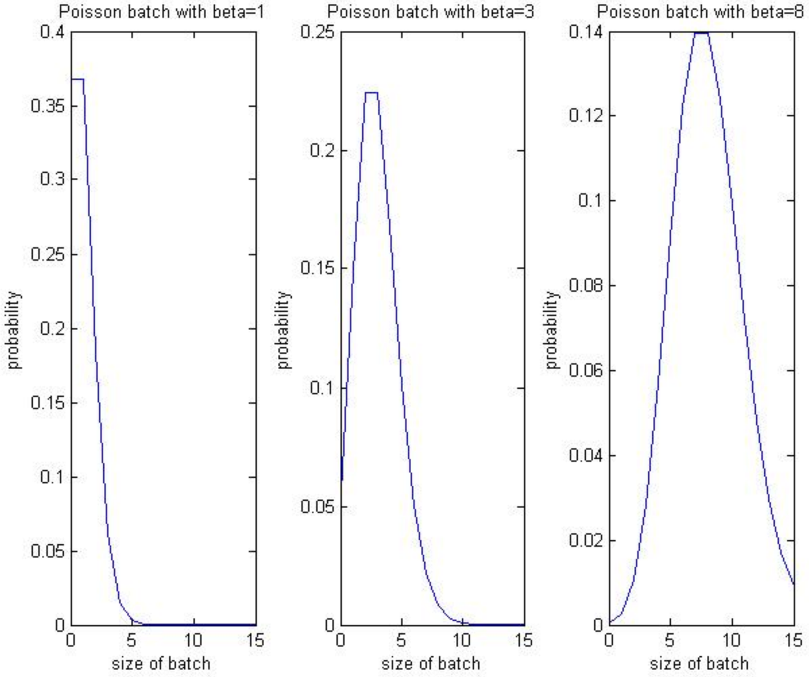

Figure 2.1: Probability distribution of Poisson batch with different β

Hence,

$$\begin{aligned} & h(\lambda, n) \\ = \;& n - \sum_{j\geq 0} E[n - N(\lambda, \hat{N})|\hat{N} = j]^{+} P(\hat{N} = j) \\ = \;& n - \sum_{j\geq 0} \sum_{k=0}^{n} (n-k) P[\sum_{i=1}^{j} B_i = k] P(\hat{N} = j) \\ = \;& n - \sum_{j\geq 0} \sum_{k=0}^{n} (n-k) e^{-j\beta} \frac{(j\beta)^k}{k!} e^{-\lambda} \frac{\lambda^j}{j!} \end{aligned} \tag{2.11}$$

The function h in (2.11) will be used in the fixed point approximation as shown later. Figure 1 shows the distribution of the Poisson batch size.

Geometric batch size

In this section, we assume that the batch size B_i follow a Geometric distribution with parameter p, $0 < p < 1$. Then we have

$$P(B_i = k) = (1-p)^k p, \text{ for } k = 0, 1, 2, 3, ... \tag{2.12}$$

Having a high value of p puts more weight on a smaller value of the batch size.

From (2.6), in order to evaluate $h(\lambda, n)$, we need to find the value of $P[\sum_{i=1}^{j} B_i = k]$. If B_i are independent geometrically distributed variables with parameter p, then $\sum_{i=1}^{j} B_i = k$ follows a negative binomial distribution with parameters j and p. See figure 2 for the graph for the Geometric batch size. The probability mass function of a random variable with a negative binomial distribution (NegBin(j, p))takes the following form:

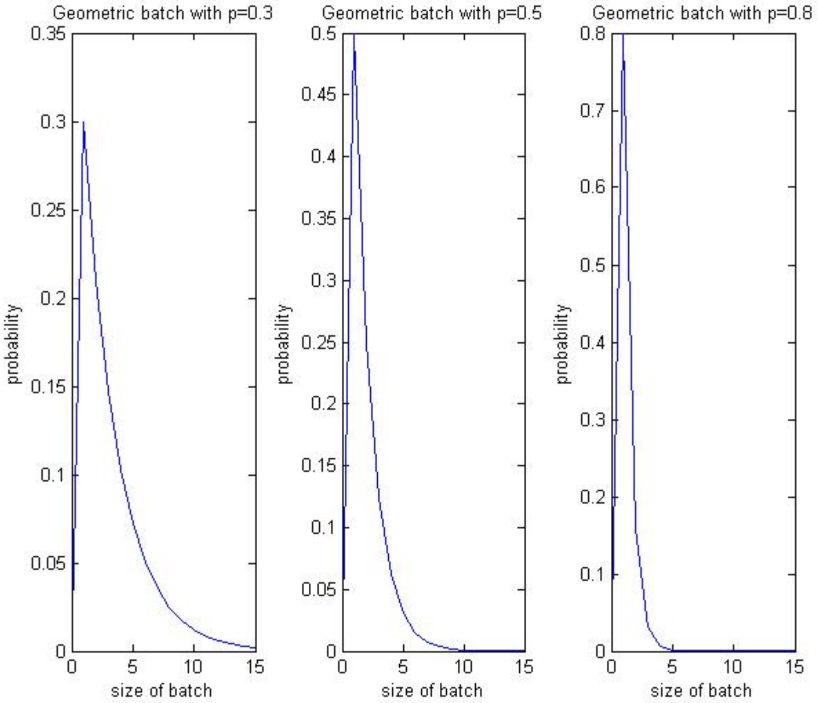

Figure 2.2: Probability distribution of Geometric batch with different p

$$f(k;j,p) = \frac{\Gamma(j+k)}{k!\Gamma(j)} p^j (1-p)^k \text{ for } k = 0,1,2,... \tag{2.13}$$

where Γ is the gamma function.

Thus, we have

$$\begin{aligned} & h(\lambda, n) \\ = \;& n - \sum_{j \geq 0} \sum_{k=0}^{n} (n-k) P[\sum_{i=1}^{j} B_i = k] P(\hat{N} = j) \\ = \;& n - \sum_{j \geq 0} \sum_{k=0}^{n} (n-k) f(k;j,p) e^{-\lambda} \frac{\lambda^j}{j!} \end{aligned} \tag{2.14}$$

where $f(k;j,p)$ is defined by (2.13).

In the rest of the chapter, we consider general batch size distribution. Recall that the main focus is to seek a control policy on the arrival with the objective of maximizing expected revenue. Our control policy involves finding a booking limit y_r, which acts as a threshold on the number of accepted orders. Hence, the booking limit problem is equivalent to solving the following optimization problem:

$$\max_{y_r} \sum_{r \in R} w_r x_r \tag{2.15}$$

where x_r is given by (2.7). This can be simplified and formulated as the following:

$$\begin{aligned} & \max_{y} \sum_{r \in R} w_r x_r \\ s.t. \;\; & y_r + \sum_{s \neq r, s \ni k} x_s \leq C_k \text{ for } k \in r, r \in R \\ & x_r = h(\lambda_r, y_r), r \in R \end{aligned} \tag{2.16}$$

Note that the final model (2.16) is similar to the model by Li and Yao (2004), but our model is based on a more general setting, which is assuming that the arrival process is a Compound Poisson process.

However, the linear program (2.16) is restricted to certain distribution for the batch size. We have presented a few distribution of the batch sizes which does not make the evaluation of (2.6) too computationally expensive. For more general distribution for the batch size B_i, it becomes more difficult to compute the expected value of the accepted orders, since it involves calculating the probability of sum of random variables which involves the convolution of probability space. Hence, we seek an approximation in the next section.

2.4 Approximation via continuous distribution for arrival

In this section, our aim is to get a approximation for a Compound Poisson process with general batch size. We have defined the Compound Poisson process as $\{N_r(t), t \geq 0\}$ as

$$N_r(t) = \sum_{i=1}^{\hat{N}_r(t)} B_{i,r} \tag{2.17}$$

where $\{\hat{N}_r(t), t \geq 0\}$ is a Poisson process with mean $\lambda_r t$ and $\{B_{i,r}, i \geq 1\}$ be the independent and identically distributed random variables denoting the batch size of the arrival, which is independent of $\{\hat{N}_r(t), t \geq 0\}$. Let b_r and $\sigma^2_{b,r}$ be the mean and variance of the batch size for route (or class) r respectively. Hence, $E[N_r(t)] = \lambda t E[B_r] = \lambda_r t b_r$ and $Var[N_r] = \lambda t E[B_r^2] = \lambda t(\sigma^2_{b,r} + b_r^2)$.

Given that using a general discrete distribution for the batch size proves to be a difficult problem and computationally difficult to solve, our aim is to approximate the batch arrival process using a continuous distribution. We first state a theorem from R*é*nyi,1970.

Theorem 3. *Let* $X_1, X_2, ..$ *be independent, identically distributed random variables with mean 0 and positive, finite variance* σ^2*. Set* $S_n = \sum_{k=1}^{n} X_k, n \geq 1$*. Suppose that* $\{\hat{N}(t), t \geq 0\}$ *is a family of positive, integer-valued random variables, such that, for some* $0 < \lambda < \infty$,

$$\frac{\hat{N}(t)}{t} \rightarrow_p \lambda \ \textit{as}\ t \rightarrow \infty.$$

Then,

$$\frac{S_{\hat{N}(t)}}{\sigma\sqrt{\hat{N}(t)}} \rightarrow_d \boldsymbol{N}(0,1) \ \textit{as}\ t \rightarrow \infty,$$

$$\frac{S_{\hat{N}(t)}}{\sigma\sqrt{\lambda t}} \rightarrow_d \boldsymbol{N}(0,1) \ \textit{as}\ t \rightarrow \infty$$

where $\boldsymbol{N}(0,1)$ *denote the standard normal distribution.*

Definition : Let Y and Y_n be random variables for $n = 1, 2, ...$ Y_n is said to converge to Y in probability, written as $Y_n \rightarrow_P Y$ if $\lim_n P[|Y_n - Y| \geq \epsilon] = 0$ holds for all $\epsilon > 0$. If Y_n is said to converge in distribution to Y, written as $Y_n \rightarrow_d Y$, then $\lim_n P[Y_n \leq y] = P[Y \leq y]$ holds for every y such that $P[Y = y] = 0$.

Back to our problem, we first consider the Poisson process, $\hat{N}(t), t \geq 0$. Let T_n denote the elapsed time between $(n-1)$st and the nth event and let $\hat{S}_n = \sum_{i=1}^{n} T_i, n \geq 1$. We know that $\hat{N}_r(t) \geq n \Longleftrightarrow \hat{S}_n \leq t$ and T_n are independent identically distributed exponential random variables with mean $1/\lambda$. As $t \to \infty$, $n \to \infty$ and using the strong law of large numbers,

$$P[\lim_{n\to\infty} \frac{\hat{S}_n}{n} = \frac{1}{\lambda}] = 1.$$

Hence,

$$\begin{aligned}
P[\frac{\hat{S}_n}{n} = \frac{1}{\lambda}] &= P[\hat{S}_n = \frac{n}{\lambda}] \\
&= P[\hat{N}(\frac{n}{\lambda}) - \epsilon < n < \hat{N}(\frac{n}{\lambda}) + \epsilon] \text{ for some arbitrary } \epsilon > 0 \\
&= P[\hat{N}(t) - \epsilon < \lambda t < \hat{N}(t) + \epsilon] \text{ (let } t = \tfrac{n}{\lambda}) \\
&= P[\frac{\hat{N}(t)}{t} - \frac{\epsilon}{t} < \lambda < \frac{\hat{N}(t)}{t} + \frac{\epsilon}{t}]
\end{aligned} \tag{2.18}$$

Let $t \to \infty$ on both sides, we have

$$P[\lim_{t\to\infty} \frac{\hat{N}(t)}{t} = \lambda] = 1 \tag{2.19}$$

which implies $\hat{N}(t)/t \to_p \lambda$ as $t \to \infty$. Denote $\tilde{S}_{r,n} = \sum_{i=1}^{n} B_{i,r}$ where $\{B_{i,r}, i \geq 1\}$ are independent and identically distributed random variables denoting the batch size of the arrival, with mean and variance of b_r and $\sigma_{b,r}^2$ respectively. Hence, $N_r(t) = \tilde{S}_{r,\hat{N}_r(t)}$. We standardize the random variable $\tilde{S}_{r,n}$ into the random variable $\tilde{S}^*_{r,n} = (\tilde{S}_{r,n} - E[\tilde{S}_{r,n}])/(\sqrt{var[\tilde{S}_{r,n}]}) = (\tilde{S}_{r,n} - nb_r)/(\sqrt{n}\sigma_{b,r})$, which have 0 expectation and standard deviation 1. Thus, using theorem 3, we can deduce that as $t \to \infty$, we have $(\tilde{S}_{r,\hat{N}_r(t)} - \lambda t b_r)/(\sigma_{b,r}\sqrt{\lambda t}) \to_d \mathbf{N}(0, 1)$.

Denote $z(N_r(t), t) = (N_r(t) - \lambda_r t b_r)/(\sigma_{b,r}\sqrt{\lambda t})$. Therefore,

$$
\begin{aligned}
\lim_{t\to\infty} P[N_r(t) \le x] &= P[z(N_r(t),t) \le \frac{z-\lambda_r t b_r}{\sigma_{b,r}\sqrt{\lambda t}}] \\
&= \int_0^x \Phi(\frac{u-\lambda_r t b_r}{\sigma_{b,r}\sqrt{\lambda t}})d(u) \\
&= \int_0^x \Phi_r(z(u,t))d(u) \qquad (2.20)
\end{aligned}
$$

where $\Phi_r(\cdot)$ is the density probability function for the normal random variable associated to route r.

We consider the limiting region, that is, when t is large. For ease in illustration, we drop the term t in our steps, and let $\tilde{N}_r$ denote the arrival process in some limiting regime. Let $F_r(x)$ be the cumulative distribution function for the corresponding normal distribution function. Denote:

$$
\begin{aligned}
E[\tilde{N}_r \wedge n] &= h_c(\Phi_r, n) \\
&= \int_0^n k\Phi_r(z(k))dk + \int_n^\infty n\Phi_r(z(k))dk \\
&= kF_r(z(k))|_0^n - \int_0^n F_r(z(k))dk + nF_r(z(k))|_n^\infty \\
&= nF_r(z(n)) - \int_0^n F_r(z(k))dk + n - nF_r(z(n)) \\
&= n - \int_0^n F_r(z(k))dk \\
&= \int_0^n 1 - F_r(z(k))dk \\
&= \int_0^n \bar{F}_r(z(k))dk \qquad (2.21)
\end{aligned}
$$

Note that we introduce a new notation h_c to denote the continuous case. Similarly as the discrete case, under the approximation (2.5) and using the function h_c derived from (2.21), we denote:

$$x_r = E[A_r] = h_c(\Phi, y_r \wedge m_r), r \in R \tag{2.22}$$

Hence in the limiting region, given y_r, the system of equations from (2.22) defined a fixed point model to approximate the expected number of expected orders on each route r.

Remark:

1. The expression for $h_c(f, n)$ in (2.21) holds for any probability distribution function, f, and corresponding cumulative distribution function, F. Hence, if we approximate the batch size using some arbitrary continuous distribution, we can use (2.21) to evaluate $E[\tilde{N}_r \wedge n]$.

2.5 Revenue Management Problem

Assume the setting of the network we described in the previous section. Our aim is to set the booking limit y_r by maximizing the expected revenue. Hence, the booking limit problem is equivalent to solving the following optimization problem:

$$\max_{y_r} \sum_{r \in R} w_r x_r \tag{2.23}$$

where $x_r = E[A_r]$ is given by (2.22). Then the revenue optimization problem becomes

$$\begin{aligned} & \max_{y_r} \sum_{r \in R} w_r x_r \\ s.t. \quad & y_r + \sum_{s \neq r, s \ni l} x_s \leq C_l \text{ for } l \in r, r \in R \\ & x_r = h_c(\Phi_r, y_r), r \in R \end{aligned} \tag{2.24}$$

Using (2.21),

$$\begin{aligned} & \max_{y_r} \sum_{r \in R} w_r \int_0^{y_r} \bar{F}_r(z(k)) dk \\ s.t. \quad & y_r + \sum_{s \neq r, s \ni l} \int_0^{y_s} \bar{F}_s(z(k)) dk \leq C_l \text{ for } l \in r, r \in R \end{aligned} \tag{2.25}$$

One can check that the objective function $\sum_{r \in R} w_r \int_0^{y_r} \bar{F}_r(z(k)) dk$ is an increasing concave function with respect to y_r. Hence, this is a convex linear programming problem. Let η_{rk} be the Lagrangian multipliers. Using K.K.T. conditions, we have the following optimality equations:

$$w_r \bar{F}_r(z(y_r)) - \sum_{l \in r} \eta_{rl} - (\sum_{s \neq r} \sum_{l \in s \cap r} \eta_{sl}) \bar{F}_r(z(y_r)) = 0 \tag{2.26}$$

To simplify the notations, let $\pi_r = \sum_{l \in r} \eta_{rl}$ and $\gamma_r = \sum_{s \neq r} \sum_{l \in s \cap r} \eta_{sl}$. Thus,

$$\begin{aligned} & w_r \bar{F}_r(z(y_r)) - \pi_r - \gamma_r \eta_{sl} \bar{F}_r(z(y_r)) = 0 \\ \Rightarrow \quad & \bar{F}_r(z(y_r)) = \frac{\pi_r}{w_r - \gamma_r} \\ \Rightarrow \quad & F_r(z(y_r)) = \frac{w_r - \pi_r - \gamma_r}{w_r - \gamma_r} \end{aligned} \tag{2.27}$$

Hence, the optimal booking limit for route r should be set at y_r, such that $P(N_r \leq y_r) = F_r(z(y_r))$ which is the proportion of orders accepted for that route is the ratio as derived above. The term π_r represents the penalty cost from using the links, and the term γ_r represents the indirect penalty cost which represents the impact on other routes that share the link with route r. We see that the numerator of the ratio represents the net profit from each accepted order for route r and the denominator of the ratio represent the net profit plus the indirect cost penalty.

2.5.1 Solving Methodology

From (2.21), we see that $E[N_r \wedge n]$ is increasing with respect to n. Assume n is an integer. Then, we can write:

$$\begin{aligned} E[N_r \wedge n] &= \int_0^n \bar{F}_r(z(k))dk \\ &= \sum_{j=0}^{n} \theta_{r,j} \int_j^{j+1} \bar{F}_r(z(k))dk \end{aligned} \tag{2.28}$$

where $\theta_{r,j} = 1 \iff \theta_{r,j+1} > 0$, $\theta_{r,j} \leq \theta_{r,j+1}$ and $\theta_{r,j} \in [0,1]$ for all $j \geq 0$ and $r \in R$.

From the constraint in (2.24), we see that $y_r \leq C_l - \sum_{s \neq r, s \ni l} E[A_s] = m_r$ for $l \in r, r \in R$. Thus $E[A_r] = E[N_r \wedge y_r \wedge m_r] = E[N_r \wedge y_r]$. Let $\hat{C}_r = min_{l \in r} C_r$. Note that $y_r \leq \hat{C}_r$ for all $r \in R$. In order to keep the feasible region as small as possible, we can write:

$$E[A_r] = E[N_r \wedge y_r] = E[N_r \wedge \hat{C}_r] = \sum_{j=0}^{\hat{C}_r} \theta_{r,j} \int_j^{j+1} \bar{F}_r(z(k))dk \tag{2.29}$$

$$y_r = \sum_{j=0}^{\hat{C}_r} \theta_{r,j} \tag{2.30}$$

Using (2.29) and (2.30), we express (2.24) as a optimization problem. Our decision variables becomes $\theta_{r,j} \in [0,1]$.

$$
\begin{aligned}
&\max_{\theta_{r,j}} \sum_{r\in R} w_r \sum_{j=0}^{\hat{C}_r} \theta_{r,j} \int_j^{j+1} \bar{F}_r(z(k))dk \\
s.t. \quad & \sum_{j=0}^{\hat{C}_r} \theta_{r,j} + \sum_{s\neq r, s\ni l} \sum_{j=0}^{\hat{C}_s} \theta_{s,j} \int_j^{j+1} \bar{F}_s(z(k))dk \le C_l \text{ for } l \in r, r \in R \\
& \theta_{r,j} = 1 \Longleftrightarrow \theta_{r,j+1} > 0, \forall j, r \in R \\
& \theta_{r,j} \le \theta_{r,j+1}, \forall j, r \in R \\
& 0 \le \theta_{r,j} \le 1, \forall j, r \in R
\end{aligned}
\qquad (2.31)
$$

The program (2.31) is not easy to solve since it involves a integer constraint. One can show that the function $h_c(\Phi, n)$ is a increasing concave function in n. Using this property, we can remove the integer constraint $\theta_{r,j} = 1 \Longleftrightarrow \theta_{r,j+1} > 0$ and constraint $\theta_{r,j} \le \theta_{r,j+1}, \forall j, r \in R$. The linear program becomes

$$
\begin{aligned}
&\max_{\theta_{r,j}} \sum_{r\in R} w_r \sum_{j=0}^{\hat{C}_r} \theta_{r,j} \int_j^{j+1} \bar{F}_r(z(k))dk \\
s.t. \quad & \sum_{j=0}^{\hat{C}_r} \theta_{r,j} + \sum_{s\neq r, s\ni l} \sum_{j=0}^{\hat{C}_s} \theta_{s,j} \int_j^{j+1} \bar{F}_s(z(k))dk \le C_l \text{ for } l \in r, r \in R \\
& 0 \le \theta_{r,j} \le 1, \forall j, r \in R
\end{aligned}
\qquad (2.32)
$$

Solving the above linear program (2.32), the optimal booking limit is

$$
y_r^* = \sum_{j=0}^{\hat{C}_r} \theta_{r,j}^*, r \in R \qquad (2.33)
$$

where $\theta_{r,j}^*, r \in R$ is the optimal solution to (2.32).

2.5.2 Asymptotic Optimality under Fluid Scaling

In this section, we discuss the optimality of our control policy. We know that the stochastic model corresponding to the single Poisson arrival process is asymptotically optimally under fluid scaling. The same conclusion can be drawn from the same stochastic model but with the single Poisson arrival process replaced by the more general compound Poisson process. We first state the result and highlight the outline for the proof, which follows from the idea of Ye and Yao (2006), in the appendix.

We introduce a sequence of networks, indexed by the superscript k. Each of the network in the sequence is exactly the same as the network introduced, but with its arrival rates and link capacities replaced by $k\lambda_r, r \in R$ and $kC_l, l \in L$ respectively. We denote the superscript k for the variables in the k-th network. The variables concerned are $A_r^k(t), N_r^k(t), y_r^k$. The price w_r for each sequence of network remains unchanged. Recall that $z(N_r^k(t), t) = (N_r^k(t) - k\lambda_r t b_r)/(\sigma_{b,r}\sqrt{k\lambda t})$. Let the time horizon t be 1. Note that there is no loss in generality when we drop the time argument, t, and assume $t = 1$. Omitting the time argument does not change the proof, and it can be easily generalize to the case of an arbitrary t. To ease the notation, we omit the variable t. From (2.25), the k-th network becomes:

$$\begin{aligned} &\max_{x^k, y^k} \sum_{r \in R} w_r x_r^k \\ s.t. \quad & y_r^k + \sum_{s \neq r, s \ni l} x_s^k \leq kC_l \text{ for } l \in r, r \in R \\ & x_r^k = \int_0^{y_r^k} \bar{F}_r(z(u))du \end{aligned} \tag{2.34}$$

Under fluid scaling, we denote:

$$(\bar{A}_r^k(t), \bar{N}_r^k(t), \bar{y}_r^k, \bar{x}_r^k) = (\frac{1}{k}A_r^k(t), \frac{1}{k}N_r^k(t), \frac{1}{k}y_r^k, \frac{1}{k}x_r^k) \tag{2.35}$$

Using the notations from fluid scaling, we transform the problem (2.34) into

$$\begin{aligned} &\max_{\bar{x}^k, \bar{y}^k} \sum_{r\in R} w_r \bar{x}_r^k \\ s.t. \quad &\bar{y}_r^k + \sum_{s\neq r, s\ni l} \bar{x}_s^k \leq C_l \text{ for } l \in r, r\in R \\ &\bar{x}_r^k = \int_0^{\bar{y}_r^k} \bar{F}_r(\frac{u-\lambda_r b_r}{\sqrt{\lambda t \sigma_{b,r}^2/k}})du \end{aligned} \tag{2.36}$$

We first state the result in the following theorem.

Theorem 4. *Suppose for each k, $(\bar{x}_r^k, \bar{y}_r^k)_{r\in R}$ is an optimal solution to the k-th network in (2.36).*

1. *Then, for any subsequence of this solution sequence, there exists a further subsequence that converges to a limit, $(\bar{y}_r)_{r\in R}$, and the limit $(\bar{y}_r)_{r\in R}$ is an optimal solution to the following problem:*

$$\begin{aligned} &\max_{\bar{x}} \sum_{r\in R} w_r \bar{x}_r \\ s.t. \quad &\sum_{r\in l} \bar{x}_r \leq C_l \textit{ for } l \in r, r\in R \\ &\bar{x}_r = \bar{y}_r \wedge \lambda_r b_r \end{aligned} \tag{2.37}$$

2. *Under the threshold control with $y_r^k = k\bar{y}_r^k, r \in R$, the (fluid scaled) accepted orders converge to the optimal solution of (2.37), i.e.,*

$$\bar{A}_r^k \rightarrow \bar{x}_r^* \textit{ as } k\rightarrow\infty, r\in R, \tag{2.38}$$

along the convergent subsequence.

3. *The threshold control is asymptotically optimal (under fluid scaling) in the following sense:*

   Let $(A_r^{k,G})_{r\in R}$ *denote the number of accepted order associated with any (general) threshold control scheme* G *for the* k*-th network; and let* $(\bar{A}_r^{k,G})_{r\in R} = (\frac{1}{k}A_r^{k,G})_{r\in R}$*. Then we have*

$$\limsup_{k\to\infty} \sum_{r\in R} w_r \bar{A}_r^{k,G} \leq \sum_{r\in R} w_r \bar{x}_r^* \tag{2.39}$$

   where $\bar{x}_r^*$ *solves* (2.37)*.*

Proof : See appendix.

Theorem 4 shows that the original problem (2.24) is asymptotically optimal in the fluid sense. To be more precise, the original problem can be approximated by solving (2.37) for the optimal booking limit. The approximation is more accurate when the arrival rate $\lambda_r, r \in R$ and the capacity of the link $C_l, l \in L$ are scaled up by the same factor.

2.6 Numerical Studies for Static policies

In this section, we illustrate our results by some numerical examples. In the first subsection, we focus on the accuracy of the static policies which we have introduced. Our aim is to show the accuracy of our proposed approximation using (2.32) and compare it with the approximation using exact discrete distribution for the batch size. We first show numerically that our proposed approximation is close to that of the discrete approximation proposed by Li and Yao (2004). Although our proposed allocation is based on the usage of the normal distribution to approximate the allocation method given by Li and Yao (2004), it provides more generality in the assumption of the batch size distribution for the compound Poisson process, yet maintain the gap between the two approximations. Given that our proposed allocation is still an approximation, it is interesting to see how it fare with the actual optimal allocation. We compare the expected result computed from our approximation with results from the hindsight optimum.

Generalizing the arrival process allows us to explore deeper into the problem such as the effect of the batch size has on the expected revenue obtained. In example 3, we compare the accuracy of our approximation policy with the hindsight optimum. The accuracy of our approximation policy is compared with the policy using discrete distribution for the batch size. In reality, it is unrealistic to assume a certain distribution for the batch size. Hence, our approximation assumes a general batch size distribution and the results obtained by the policy is compared with the hindsight optimum.

2.6.1 Implementation of Static Policies

In this subsection, we will deal with the implementation of the static policies and test the accuracy of our proposed approximation with the hindsight optimal solution.

Example 1 (Comparison of the result from the normal distribution approximation with the result using the discrete distribution)

We first compare the numerical result from our proposed approximation with that of the approximation from using the discrete distribution for the batch size. For ease of presentation, we refer the approximation by Li and Yao (2004) as approximation 1.

We consider an example with 2 classes of customers, r_1 and r_2, using the same resources from 1 link with capacity $C = 10$. Classes r_1 and r_2 follow a Compound Process with rates $\lambda_{r1} = 5$ and $\lambda_{r2} = 10$. Revenue earned from accepting one entry from classes r_1 and r_2 are $w_1 = 2$ and $w_2 = 1$ respectively. We assume a Poisson batch size. Batch size from class r_1 and r_2 are independent Poisson random variables with mean 2 and 1 respectively.

We test the two approximations using the same set of parameters. The arrival rates and the capacities are scaled by a factor of $k, k = 1, ...10$. For each k, we compute the expected revenue under the two approximations. The result is presented in table 2.1.

The difference and the absolute difference in the expected revenue, represented by the gap and absolute gap respectively, is monitored and presented in figure 2.4.

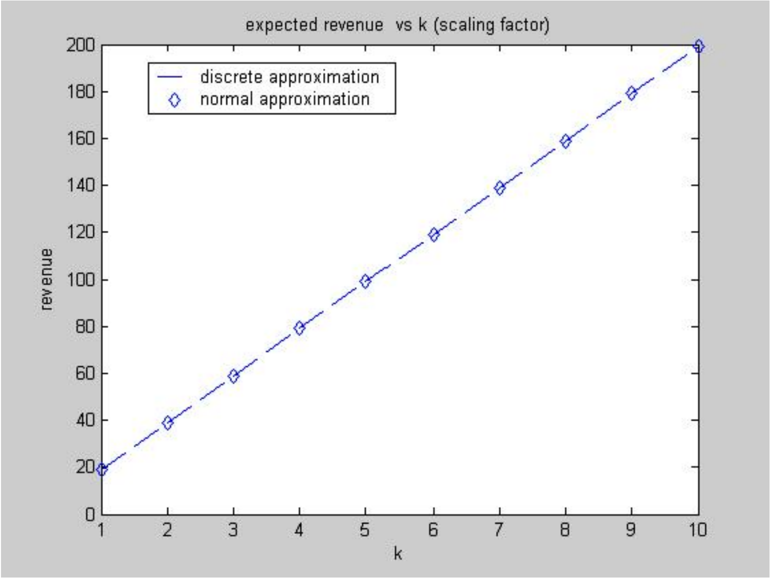

Figure 2.3: Expected revenue vs the scaling factor k

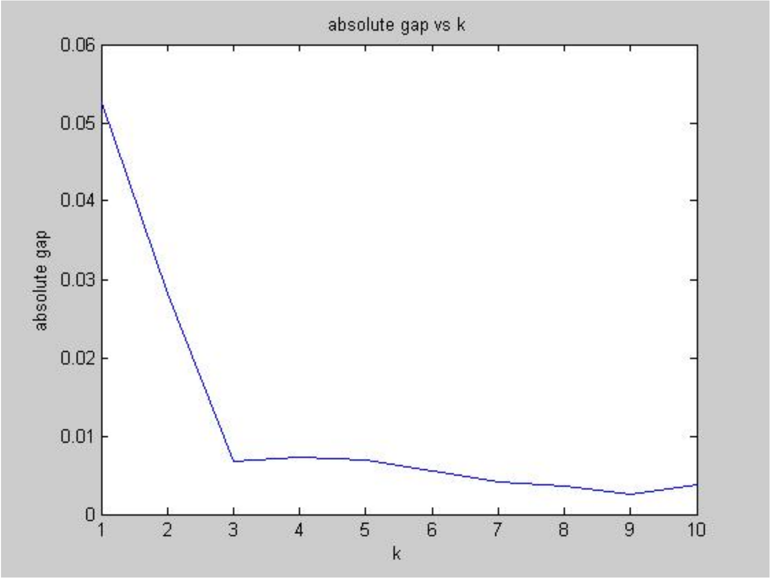

Figure 2.4: Optimal gap vs the scaling factor k

k	revenue (discrete)	revenue (normal)
1	19.08	19.17
2	39.05	39.07
3	59.03	59.05
4	79.03	79.04
5	99.02	99.04
6	119.02	119.03
7	139.02	139.03
8	159.02	159.02
9	179.01	179.02
10	199.01	199.02

Table 2.1: Expected revenue

From figure 2.3, we see that the expected revenue computed under the two approximations are very close. As expected, the expected revenue for both cases increases with the problem size. Figure 2.4 shows that the absolute gap between the two approximations is small. The absolute gap decreases with the problem size and remain consistent at about 0.05 with a scaled factor of 10. Hence, our approximation gets closer to the approximation 1 with the problem size. This shows that our method approximates the discrete case well.

Comparing with simulated results

We know that our proposed allocation is not optimal. Hence, we seek to see how large is the optimal gap with the hindsight optimum of simulated results.

Similarly, we have 2 classes of customers, r_1 and r_2, using the same resources from 1 link with capacity $C = 20$. Classes r_1 and r_2 follow a Compound Process with rates $\lambda_{r1} = 3$ and $\lambda_{r2} = 5$. Revenue earned from accepting one entry from classes r_1 and r_2 are $w_1 = 1$ and $w_2 = 2$ respectively. We suppose that the batch size from class r_1 and r_2 are random variables with mean 1 and 2 respectively. The variance of the batches are both 1.

In order to obtain the simulated results, we generate 500 sample paths for the arrival process. For each sample path, we solve for the hindsight optimum. With the booking limit, the revenue earned under each sample path arrival can also be computed. The mean revenue is estimated by averaging the revenues.

Using the approximation method by normal distribution, we can compute the booking limit under the set of parameters in this example. Likewise with each sample path, we apply the booking limit on the arrival process. In this way, we can calculate the revenue for each sample path and the mean revenue can be computed similarly.

We apply the same scaling as in example 1, to the arrival rates and the capacities by a factor of $k, k = 1, ...10$. For each k, we compute the mean revenue from the simulation and the average expected revenue from our policy derived from approximation implemented on the simulated results.

In figure 2.5, we see that the revenue from the simulated results and the approximation increases with the problem size. The results derived from the two methods are very close. Note that the simulated results is always larger than that of the approximated result. This is obvious since we applied the booking limit to the simulated data.

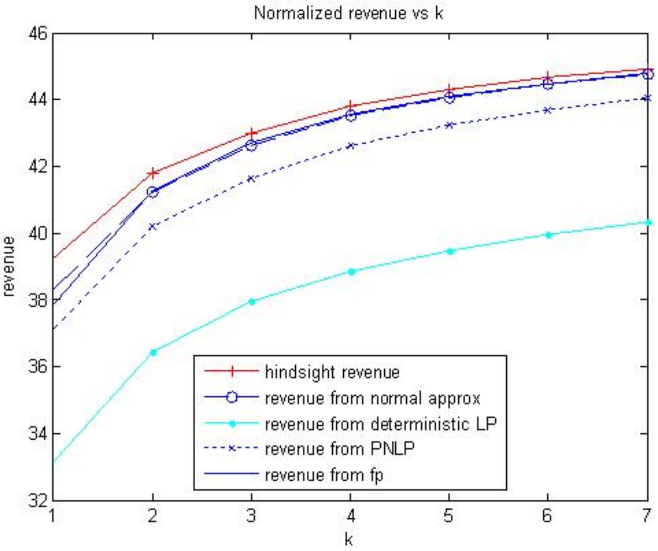

Figure 2.5: Revenue vs the scaling factor k

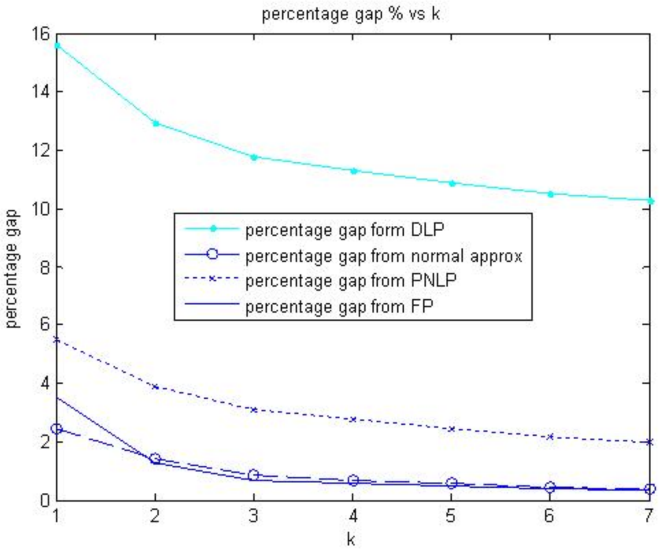

Figure 2.6: Optimal gap vs the scaling factor k

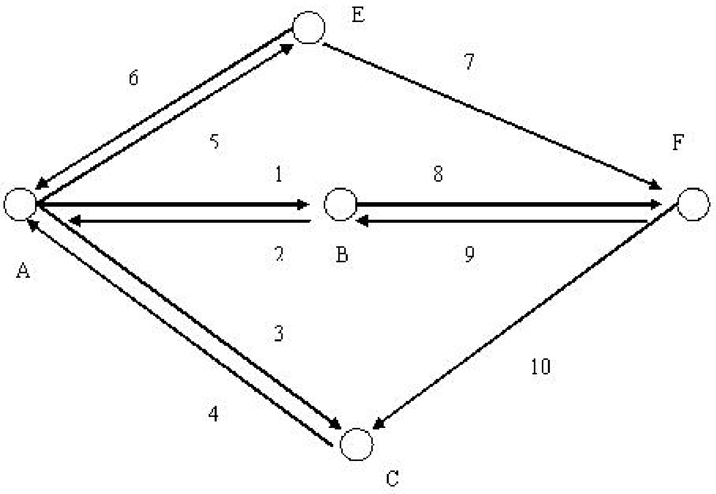

Figure 2.7: Network with 5 nodes and 11 links

For ease of comparison, we compute the gap which is the difference of the exact revenue from simulation and the revenue obtained from the proposed booking limit. The error can be better analyzed by taking the absolute value of the gap which is presented in figure 2.6 graph 2. A more accurate comparison is to compare the normalized gap, which is the absolute gap divided by the scale factor. The normalized gap will tell us the accuracy of the approximation with respect to the size of the problem. It is clear from figure 2.4 that the normalized gap decreases to 0 as the problem size increases. This further illustrates the asymptotical optimality of the proposed approximation.

Example 2 (Arbitrary network)

In this example, we want to show the accuracy of our approximation for a more complicated network as compared to example 1. We compare our approximation with the allocation from simulated results in a arbitrary network shown in figure 2.7. The parameters for this network is presented in table 2. There are 5 nodes and 10 links in the network. Each link has a capacity of 10. 22 routes are present and their arrival follow a compound Poisson process with arrival rate and the mean of the batch size given in the table. Each route brings a certain revenue for every acceptance of entry for that route and pass through certain links. For example, arrival from route 2 passes from node A to node B using link 1 and passes from node B to its destination node D using link 8, and each accepted arrival will fetch a revenue of 1 unit.

Similarly, we generate 1000 sample paths for the arrival process. For each sample path, we solve for the hindsight optimum. With the booking limit based on our approximation, the revenue earned under the sample path arrival can be computed. The mean revenue is estimated by averaging the revenues. Using our proposed approximation, we solve the network using the set of parameters in table 2.2. For each sample path, we apply the computed booking limit from our approximation. In this way, we can compute the revenue under our booking limit policy for each sample path. The average of the approximated revenue is being recorded. Hence, the gap(exact revenue - approximated revenue) can be calculated easily.

We apply the same scaling as in the previous example, i.e. the arrival rates and the capacities by a factor of $k, k = 1, ...10$. For each k, we compute the mean revenue from the simulated sample path and the average expected revenue from our approximation applied to

Route	Nodes	Link used	Price	Arrival rate	Mean of batch size
1	A-B	1	1	3	2
2	A-B-D	1,8	2	5	1
3	A-C	3	1	6	1
4	A-E	5	1	4	2
5	A-E-D	5,7	2	3	1
6	B-A	2	1	4	1
7	B-D-C	8,10	2	6	2
8	B-A-E	2,5	3	3	1
9	C-A	4	1	2	1
10	C-A-E	4,5	4	6	2
11	D-B	9	1	6	1
12	D-B-A	2,9	3	6	2
13	D-C	10	1	4	1
14	D-C-A	4,10	1	2	1
15	E-D	7	1	2	1
16	E-D-C	7,10	4	4	2
17	E-A	6	2	5	1
18	E-A-B	1,6	1	2	1
19	E-A-C	3,6	1	6	1
20	E-D-B	7,9	1	6	1
21	E-D-B-A	2,7,9	5	6	1
22	C-A-B	1,4	2	3	2

Table 2.2: Parameters for Network

k	Exact rev('000)	Approx rev('000)	Gap	Normalized gap	% of error
1	0.1300	0.1229	7.0906	7.0906	5.45
2	0.2689	0.2589	10.2010	5.1005	3.79
3	0.4060	0.3902	15.8404	5.2801	3.90
4	0.5480	0.5296	18.4139	4.6035	3.36
5	0.6879	0.6700	17.9160	3.5832	2.60
6	0.8279	0.8055	22.3724	3.7287	2.70
7	0.9659	0.9427	23.1526	3.3075	2.40
8	1.1043	1.0790	25.3246	3.1656	2.29
9	1.2406	1.2070	33.6060	3.7340	2.71
10	1.3851	1.3504	34.7566	3.4757	2.51

Table 2.3: Revenue from simulation and the revenue computed from approximation

the simulated results. The result is shown in table 2.3.

In order to give more insight, we plot the percentage of the gap i.e. gap/hindsight optimum. From figure 2.9, we see that the gap percentage decrease with the scale of the problem. The percentage of error decreases to an average of 2.5% when the problem is scaled up by a factor of of more than 5.

The numerical result illustrates that our approximation method works well with an arbitrary network in the asymptotical sense. i.e. the results gets more accurate as the scale of the problem gets scaled up.

Example 3 (Varying batch size)

With an extension from the assumption of single Poisson arrival process to a compound

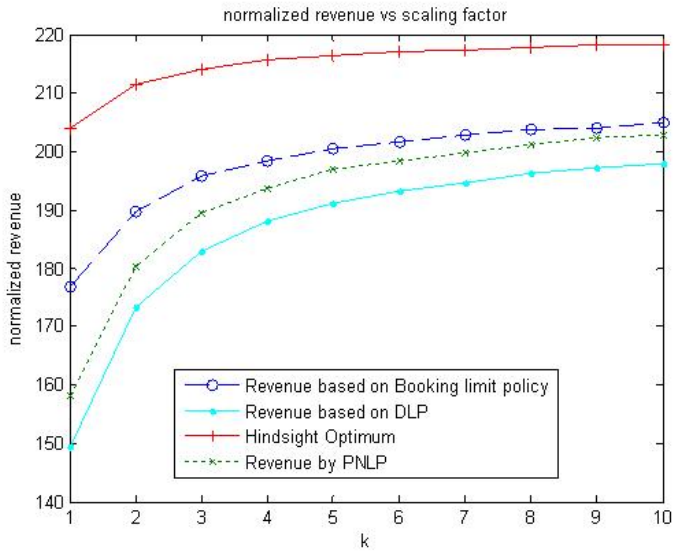

Figure 2.8: Normalized revenue vs the scaling factor k

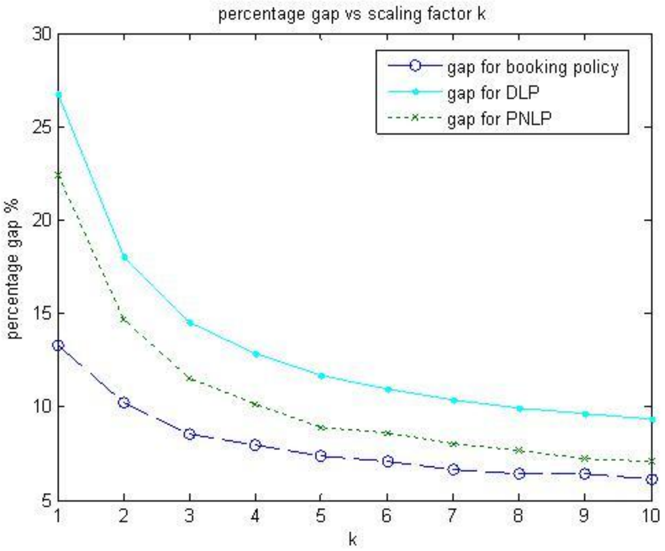

Figure 2.9: ratio of gap to exact revenue (percentage gap) vs the scaling factor k

Poisson process, we are able to analyze the impact of the distribution of the batch size has on the performance of the policy. In this example, we investigate the impact of the distribution of the batch size has on the accuracy of our approximation. Consider a single link of capacity 20, with 2 classes of customer, class 1 and class 2 with arrival rates $\lambda_1 = 3$ and $\lambda_2 = 8$ respectively. Both arrivals follow a Compound Poisson process. The batch sizes for class 1 and class 2 have mean $\beta_1 = 3$ and $\beta_2 = 2$ respectively. We assume that the batch sizes for both classes have variance of 2.

We know that mean of the class 1 and class 2 are $\lambda_1\beta_1$ and $\lambda_2\beta_2$. We vary the batch size by a factor k and reduce the arrival rate by a factor of k to keep the overall mean the same. i.e. Arrival mean of class 1 = $(\lambda_1/k)(k\beta_1)$. Hence, the larger the value k, the larger the batch size mean and the smaller is the arrival rate.

We generate 5000 sample paths for the arrival process. For every k, we solve for the hindsight optimum which we will use as our benchmark. In reality, we do not know the actual distribution of the batch size. Hence, we assume 3 cases, namely, the Poisson batch size, the Geometric batch size and the general batch size. The first two cases can be evaluated using the actual discrete distribution. The general batch size case will be computed using our proposed normal distribution approximation. Figure 2.10 shows the average revenue computed vs the factor k. It is clear that as k increases, the variation of the batch sizes gets larger and the revenue decreases. This is intuitive because with an increase in batch size and decrease in arrival rate, the problem becomes a 0-1 problem, accept all the arrival and reject all the arrival.

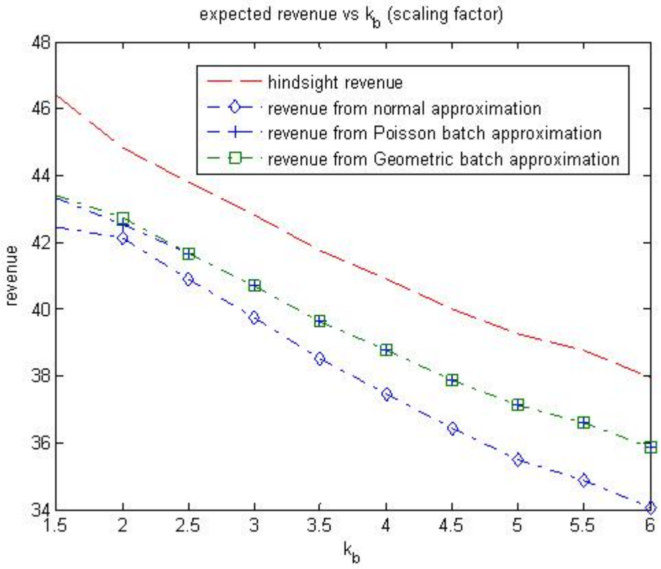

Figure 2.10: Graph of expected revenue vs factor k

Figure 2.11 shows the gap when computing using the discrete approximation and the normal distribution. Hence, using our approximation method to compute the expected revenue in reality is fairly accurate under varying batch size.

2.7 Implementation of Updating Policies

It is established that the static policies proposed approximates the optimal solution well. Static policies are easier to compute but the policies derived do not take into account the actual state of the system. In our case, a single booking limit is being applied to the entire time horizon. On the other hand, dynamic policies derived their policies by observing the states of the system produce a more realistic solution. However, the well-known curse of

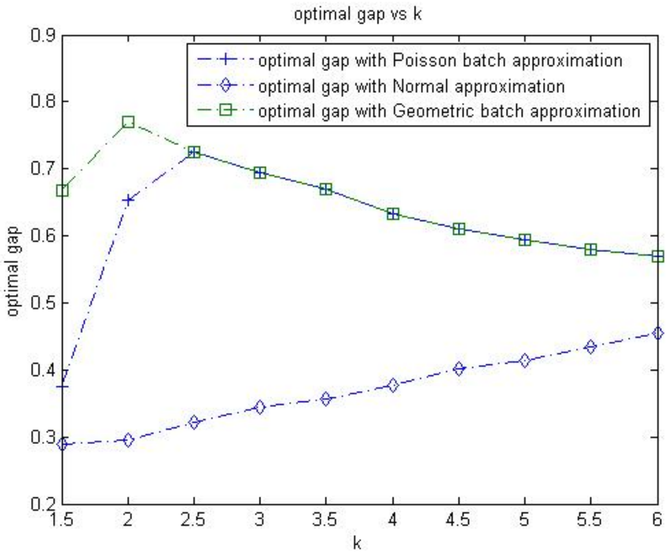

Figure 2.11: Graph of gap vs factor k

dimensionality of dynamic programming prevents it from being successfully implemented in large scaled applications.

We suggest some possible dynamic implementation of the policy. We show the accuracy of such dynamic implementation using numerical simulation. A updating mechanism is proposed and its performance is compared with that of a static policy (no updating). Our contribution involves deriving the policy and test it with some simulation results against the hindsight optimum and other updating strategies. Extensive simulation shows that our proposed updating policy produces better results.

2.7.1 Introducing updating policies

Solving revenue management problems using mathematical modeling often give rise to Markov decision process models. Such problems uses dynamic programming to produce the policy. However, in many cases, getting a exact solution from these models is close to impossible given the curse of dimension phenomenon. The computation cost is magnified greatly by the scale of the problem that it is difficult to solve even with today's supercomputers. Hence, practitioners have turned to approximate methods for their policies. Using the fixed point approximation as discussed is one of the tools available.

An intuitive way to improve the performance of the static policy is to introduce updating of the system to take account of the state of the problem into consideration before deriving the policy. This makes the policy a dynamic one. Dynamic policies are known to be more accurate, if implemented properly, than static policies because the former takes into account the state of the problem into consideration before a decision is made.

Secomandi(2007)studied the re-solving issue in a control algorithm for a class of revenue management problems. Their approach consists of heuristically solving a Markov decision process formulation of the problem and categorizes the different re-solving algorithm into different groups with distinct properties. However, their paper does not deal with the problem of selecting the re-solving time. The issue of selecting a re-solving time remains a topic for further research. Much consideration has to be put into deciding when the updating of the problem takes place. Cooper(2002) has investigated the updating procedure and stated that if the time to update is not chosen carefully, the end result for the case of updating will be less than that without updating. Secomandi (2007) analyze the problem further and attribute the inaccurate re-solving method to a lack of sequential consistency. The updating has to be done early enough for the change to be made to the existing policy to correct the deviation of the realized demand, and not too early to render the updating redundant. Deriving a protocol to determine the updating is not trivial. After the time to update is determined, the problem is resolved with the parameters of the existing model.

Before we go on, we present a numerical example which provide the motivation for the study of updating policy.

Example 4 (Comparison of the different updating strategies)

In this example, we investigate the accuracy of the different updating strategies. Consider a single link of capacity 6, with 2 classes of customer, class 1 and class 2 with arrival rates $\lambda_1 = 8$ and $\lambda_2 = 3$ respectively. Both arrivals follow a Poisson process. Revenue earned per arrival is 2 units and 5 units respectively.

Consider 2 updating strategies. Assume a time horizon of 1. The system is solved at time 0 and the optimal booking limit is implemented on the system until the next update time. The first strategy is to update the system every 0.2 time units (update 5 times) and the second strategy is to update the system every 0.5 time units (update 1 time). After updating the system (the arrival rate will be updated to the arrival mean of the remaining time and the capacity of the link will be updated according to the number of resources remaining), the problem is resolved for a new booking limit. The new booking limit will be implemented until the next updating time or the end of the planning time horizon. We compare this two updating strategy with that of the static policy (no updating). The hindsight optimum is used as a benchmark for comparison.

We generate 1000 sample paths for the arrival process. For every sample paths, we solve for the hindsight optimum which we will use as our benchmark,and apply the static policy and the dynamic policy. We apply the same scaling as in the previous examples, i.e. the arrival rates and the capacities by a factor of $k, k = 1, ...10$. For each k, we compute the revenue from the simulated sample path based on the booking limit policy computed from our approximation. The mean revenue from each of the policy is recorded and plotted against the scaling factor, k. For a better comparison, we plotted the percentage gap which is the ratio of the absolute gap of the policy (compared to the hindsight optimum) to the hindsight optimum.

Figure 2.12 shows comparison of the mean revenue derived from the hindsight optimum and the updating policies. We can see that the percentage gap of the 2 updating policies is marginally better than that of the static policy in this example in figure 2.13. Hence, arbitrary updating of the system in this case increase the expected revenue slightly. This further reemphasize on the fact that the updating time has to be carefully chosen to ensure the effectiveness of the policy. One possible explanation of the relatively poorer performance of the updating policy is that in the process of updating at many intervals without justifications, important information regarding the network system over the entire time horizon

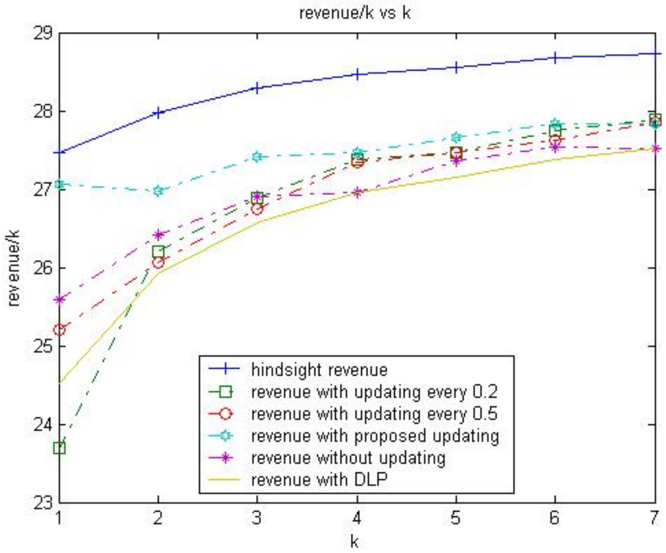

Figure 2.12: Graph of revenue/k vs k

may be lost. Hence, much care has to be taken into implementing a updating criteria.

Proposed updating mechanism

Consider a problem with a single general updating scheme. The problem is solved at time 0 and resolved at a updating time τ with the parameters of the system being updated. Let A_{r1} be the accepted orders for route $r \in R$ before the updating time. Let A_{r2} be the accepted orders for route $r \in R$ after the updating time for the remaining time horizon. Thus, the total expected orders for route $r \in R$ is $E[A_{r1} + A_{r2}]$. The accepted orders A_{r1} and A_{r2} are

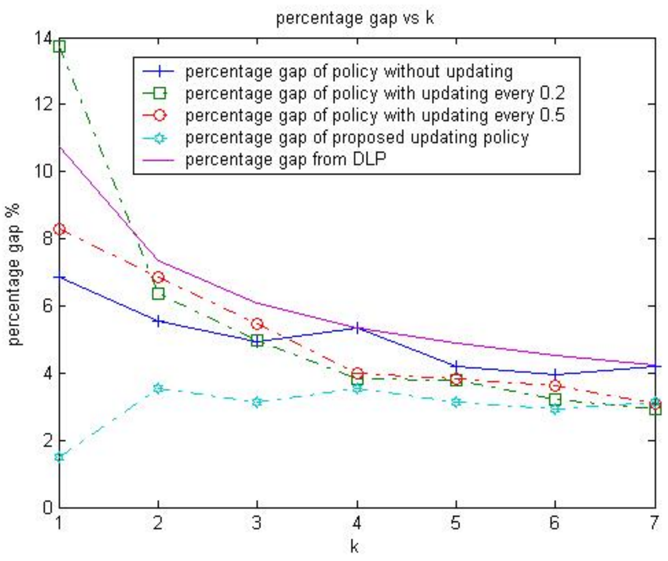

Figure 2.13: graph of percentage gap vs k

not independent. Hence,

$$E[A_{r1} + A_{r2}] = \sum_x \{x + E[A_{r2}|A_{r1} = x]\}P[A_{r1} = x] \tag{2.40}$$

Using the fixed point model, for $r \in R$,

$$E[A_{r1}] = E[N_r \wedge y_{r1} \wedge \min_{l \in r}\{C_l - \sum_{s \neq r, s \ni l} A_{r1}\}] \tag{2.41}$$

$$E[A_{r2}] = \sum_x E[\{N_r - x\} \wedge y_{r2} \wedge \min_{l \in r}\{C_l - x - \sum_{s \neq r, s \ni l} A_{r2}\}]P[A_{r1} = x] \tag{2.42}$$

Note that there are two stages to the problem now. The first stage is before the updating time and the second stage is after the updating time. We have not found the updating point, hence we have to solve for it. As a result, the revenue optimization problem is

$$\max_{(y_{r1}, y_{r2})_{r \in R}} E[A_{r1}] + E[A_{r2}] \tag{2.43}$$

The problem (2.43) is not easy to solve. Hence, we propose an approximation to it. Suppose the length of the time horizon is 1. Consider the single stage problem:

$$\begin{aligned} & \max_{y} \sum_{r\in R} w_r E[A_r] \\ s.t. \quad & y_r + \sum_{s\neq r, s\ni k} E[A_s] \leq C_l \text{ for } l \in r, r \in R \\ & E[A_r] = h(\lambda_r, y_r), r \in R \end{aligned} \tag{2.44}$$

Solving it at time 0 will give us a booking limit $y^*_{r1}, r \in R$. Implement the booking limit to the arrival process. Find the updating time, τ which is defined as $\min\{t|N_r(t) \geq y^*_{r1}\}$. In short, τ is the time when the booking limit of any class is first exceeded by the arrival process. Using τ as our resolving time, we have for stage 2,

$$E[A_{r2}] = E[\{N_r(1) - N_r(\tau)\} \wedge y_{r2} \wedge \min_{l\in r}\{C_l - \sum_{r\in l} N_r(\tau) - \sum_{s\neq r, s\ni l} A_{r2}\}] \tag{2.45}$$

Using the similar method for formulating, we have

$$\begin{aligned} & \max_{y_{r2}} \sum_{r\in R} w_r E[A_{r2}] \\ s.t. \quad & y_{r2} + \sum_{s\neq r, s\ni k} E[A_{s2}] \leq C_l - \sum_{s\ni l} N_r(\tau) \text{ for } l \in r, r \in R \\ & E[A_{r2}] = h(\lambda_r(1-\tau), y_{r2}), r \in R \end{aligned} \tag{2.46}$$

Solving (2.46) will gives us the optimal booking limit, $y^*_{r2}, r \in R$ for the remaining time horizon. We accept the arrivals according to the booking limit, $y^*_{r2}, r \in R$ for the remaining time horizon. We summarize the updating mechanism in the following algorithm.

Summary of Updating Algorithm:

1. Solve the revenue optimization at time 0 for the booking limit, $y_{r1}^*, r \in R$.

$$\begin{aligned} & \max_y \sum_{r\in R} w_r x_r \\ s.t. \quad & y_r + \sum_{s\neq r, s\ni k} x_s \leq C_k \text{ for } k \in r, r \in R \\ & x_r = h(\lambda_r, y_r), r \in R \end{aligned} \tag{2.47}$$

2. Let the realized demand be denoted as $N_r(t)$ up to time t. Find the updating time, τ which is defined as $\min\{t|N_r(t) \geq y_{r1}^*\}$.

3. We have two cases:

   (a) If $\tau \geq 1$, then no updating is done. Accept the arrivals according to the booking limit, $y_{r1}^*, r \in R$, for the whole time horizon.

   (b) If $\tau < 1$, resolve the problem (2.47) with the updated parameters, $\lambda_r \rightarrow \lambda_r(T-\tau)$ and $C_k \rightarrow C_k - \sum_{s\neq r, s\ni k} N_r(\tau)$.

$$\begin{aligned} & \max_y \sum_{r\in R} w_r x_r \\ s.t. \quad & y_r + \sum_{s\neq r, s\ni k} x_s \leq C_k - \sum_{s\neq r, s\ni k} N_r(\tau) \text{ for } k \in r, r \in R \\ & x_r = h(\lambda_r(T-\tau), y_r), r \in R \end{aligned} \tag{2.48}$$

   Solving (2.48) will gives us the optimal booking limit, $y_{r2}^*, r \in R$ for the remaining time horizon.

   Accept the arrivals according to the booking limit, $y_{r2}^*, r \in R$ for the remaining time horizon.

We implement the proposed updating policy on the same network as above to obtain the expected revenue and the percentage gap. From figure 2.12, the performance of the proposed updating policy is better than that of the static policy throughout the scaling factor k. The percentage gap for the proposed updating policy is also lower than the other policies. Hence, the updating policy is effective in this problem. With just one updating from our updating mechanism, the performance of the proposed updating policy is significantly better than the static policy and the two policies which requires more than one updating.

However, it may not be the most efficient method of updating. Future challenge will be to provide a more accurate form of updating the problem.

2.8 Conclusion

In this part of the thesis, we have formulated a general stochastic framework for network problems, in particular, the revenue management problem for airline industry. Our concern is to derive a control policy by setting a booking limit on classes of customers, with the aim of maximizing revenue. The feature of concurrent resources occupancy makes solving such problems complicated, given the randomness of the arriving flows. Hence, we seek an approximation to such problems.

Assuming the arrival process as Compound Poisson processes is a more realistic assumption to real life situations, since orders for airline tickets often come in batches. Hence, we are able to generalize the assumptions of Li and Yao (2004). Maintaining the assumption that the distribution for the batch size is discrete proves to be difficult to solve when we assume the distribution is general. We are able to give computations for batch size with Poisson distribution and Geometric distribution. We argue that in the limiting regime, that is when the time horizon is infinite, we can use the normal distribution, to approximate the batch size. While there are no accurate ways to derive a optimal control policy under the randomness of the unrealized demand, we can at best compute a approximate one. Under such the fixed point approximation, we proved that the solution is optimal in the asymptotic sense under fluid scaling. Our numerical studies have verified the accuracy of the proposed method.

From the numerical examples, we can see that the approximation does act as a good booking limit control policy, especially so in a scaled up problem. Our normal approximation approximates the general batch well. Using our proposed approximation technique, we are able to consider the general case when the batch size are random variables. The approximated results from the numerical examples are very promising. With our approximation, it opens up the option of exploring the impact of the batch size. Hence, we are able to investigate further the effect of the varying batch sizes has on the control policy. In the case of the varying batch sizes, the gap from using the normal approximation is reasonable, considering that as the batch size increase and the arrival rate decreases, the number of accepted orders becomes more difficult to compute.

Dynamic policies are better at considering the exact state of the problem before making a decision. However, the curse of dimension in the former prevents the implementation of it on large-scaled problems. We proposed a extension of our static policy to a dynamic policy, which involves updating of the system based on a stopping criteria. The results computed from simulation is encouraging, hence it suggests that more research work can be done in this direction.

2.9 Appendix : Proof for Theorem 4

Proof:

The existence of a convergent subsequence is assured since the sequence $(\bar{x}_r^k, \bar{y}_r^k)_{r\in R}$ is positive and bounded by $C_l, l \in L$.

Note that the batch size $B_{i,r}$ has finite mean. Hence, when $k \to \infty$, using the functional strong law of large numbers (See Chen and Yao(2001), Chapter 5),

$$\bar{N}_r^k \to \lambda_r b_r \text{ u.o.c.} \tag{2.49}$$

(u.o.c. stands for "uniformly on compact intervals.")

Hence, $z(\bar{N}_r^k(t)) = (\bar{N}_r^k(t) - \lambda_r b_r)/(\sigma_{b,r}\sqrt{\lambda/k}) \to 0$ as $k \to \infty$.

1. If $\bar{y}_r^k \leq \lambda_r b_r$, then $\bar{x}_r^k = \int_0^{\bar{y}_r^k} \bar{F}_r(0) d(\bar{N}_r^k) = \int_0^{\bar{y}_r^k} d(\bar{N}_r^k) = \bar{y}_r^k$.
2. If $\bar{y}_r^k > \lambda_r b_r$, then $\bar{x}_r^k = \int_0^{\bar{y}_r^k} \bar{F}_r(0) d(\bar{N}_r^k) = \int_0^{\lambda_r b_r} d(\bar{N}_r^k) = \lambda_r b_r$.

Thus, at the limit, $\bar{x}_r^k = \bar{y}_r^k \wedge \lambda_r b_r$. Using this value, from the constraint $\bar{y}_r^k + \sum_{s\neq r, s\ni l} \bar{x}_s^k \leq C_l$, we have $\sum_{r\in l} \bar{x}_r^k \leq C_l$. Thus, when $k \to \infty$, the limit of any convergent sequence feasible solution in (2.36) will be a feasible solution to the problem (2.37):

$$\begin{aligned} &\max_{\bar{x},\bar{y}} \sum_{r\in R} w_r \bar{x}_r \\ s.t. \quad &\sum_{r\in l} \bar{x}_r \le C_l \text{ for } l \in r, r \in R \\ &\bar{x}_r = \bar{y}_r \wedge \lambda_r b_r, r \in R \\ &\bar{y}_r \ge 0, r \in R \end{aligned} \tag{2.50}$$

Let $(\bar{x}^*_r, \bar{y}^*_r)_{r\in R}$ be the limit of the convergent subsequence. Then $(\bar{x}^*_r, \bar{y}^*_r)_{r\in R}$ is a feasible solution to (2.37).

Claim: $(\bar{x}^*_r, \bar{y}^*_r)_{r\in R}$ is an optimal solution to (2.37).

Suppose not. There exists $(\tilde{x}_r, \tilde{y}_r)_{r\in R}$ such that

$$(\tilde{x}^k_r, \tilde{y}^k_r)_{r\in R} \to (\tilde{x}_r, \tilde{y}_r)_{r\in R} \text{ as } k \to \infty, \tag{2.51}$$

and $(\tilde{x}_r, \tilde{y}_r)_{r\in R}$ is another feasible solution to (2.37) with a greater objective function. i.e. $\sum_{r\in R} w_r \bar{x}^*_r < \sum_{r\in R} w_r \tilde{x}_r$. Thus, we have

$$\sum_{r\in R} w_r \tilde{x}^k_r \to \sum_{r\in R} w_r \tilde{x}_r > \sum_{r\in R} w_r \bar{x}^*_r \text{ as } k \to \infty. \tag{2.52}$$

We can choose a k large enough such that $\sum_{r\in R} w_r \tilde{x}^k_r > \sum_{r\in R} w_r \bar{x}^k_r$. Note that we have assumed that $(\bar{x}^k_r, \bar{y}^k_r)_{r\in R}$ is an optimal solution to the k-th network in (2.36), thus there is a contradiction and hence our claim.

The solution for each k-th network may not be unique. As a result, the solution sequence $(\bar{x}_r, \bar{y}_r)_{r\in R}$ may not converge as $k \to \infty$. But, for any subsequence of the solution sequence, there will always be a convergent subsequence such that its limit is a solution to the problem (2.37), which is the limit of the sequence of problems, with the values of the variables at their limits.

We now prove the second part of the theorem. Note that $a \wedge b = a - (a-b)^+$. Using this identity, problem (2.37) can be converted to

$$\begin{aligned} &\max_{\bar{x},\bar{y}} \sum_{r\in R} w_r \bar{x}_r \\ s.t. \quad &\sum_{s\in l} \bar{x}_s \leq C_l - \max_{r\ni l}(\bar{y}_r - \lambda_r b_r)^+ \text{ for } l \in r, r\in R \\ &\bar{x}_r = \bar{y}_r \wedge \lambda_r b_r, r \in R \\ &\bar{y}_r \geq 0, r \in R. \end{aligned} \tag{2.53}$$

From the optimal solution computed from (2.37), we can find a set of bottleneck link. A bottleneck link is defined to be the links which contain a binding constraint in the optimal solution, i.e., $\sum_{r\ni l} \bar{x}_r = C_l$. Let L^* denote the set of all the bottleneck links associated with the optimal solution, and let R^* denote the set of all bottleneck routes. We define a route as a bottleneck route if it contains at least one bottleneck link.

Consider any $r \in R^*$. Choose any bottleneck link $l \in L^*$ such that $r \ni l$. Thus,

$$(\bar{y}_r^* - \lambda_r b_r)^+ \leq C_l - \sum_{s\ni l} \bar{x}_s^* = 0.$$

Note that the inequality is from the capacity constraint in (2.53) and the equality is due to the choice of a bottleneck link. Thus, $\bar{y}_r^* \leq \lambda_r b_r$. Hence, we can deduce that $\bar{x}_r^* = \bar{y}_r^*$ from the second constraint of (2.53).

For $r \in R \setminus R^*$, we claim that $\bar{x}_r^* = \lambda_r b_r$. Suppose not. Then $\bar{x}_r^* = \bar{y}_r^* < \lambda_r b_r$. We define $(\bar{x}_r^* = \epsilon, \bar{y}_r^* + \epsilon)$ and $(\bar{x}'_s, \bar{y}'_s) = (\bar{x}_s^*, \bar{y}_s^*)$ for $s \neq r$, where $\epsilon > 0$. Since ϵ is arbitrary, we can find another feasible solution $(\bar{x}'_s, \bar{y}'_s), s \in R$ to (2.53) but with a greater objective value than the optimal $(\bar{x}_s^*, \bar{y}_s^*), s \in R$. This is a contradiction, and hence our claim.

Recall that from the definition of accepted orders $A_r(t)$ from our model, we have

$$A_r = N_r \wedge y_r \wedge \min_{l \in r}\{C_l - \sum_{s \neq, s \ni l} A_s\}.$$

We assume that $\bar{y}_r^k \to \bar{y}_r^*$ as $k \to \infty$ along a convergent subsequence. Call the subsequence K. Applying fluid scaling to the above equation for each network. It can be shown that the u.o.c. limit of $\bar{A}_r, r \in R$ of any convergent subsequence of $(\bar{A})_{r \in R}, k \in K$ satisfies

$$\bar{A}_r = \lambda_r b_r \wedge \bar{y}_r^* \wedge \min_{l \in r}\{C_l - \sum_{s \neq, s \ni l} \bar{A}_s\}.$$

We are done if we can show that the limit $\bar{A}_r = \bar{x}_r^*$. We can rewrite the above equation as $\bar{A}_r \leq \lambda_r b_r \wedge \bar{y}_r^* = \bar{x}_r^*$. Hence,

$$\min_{l \in r}\{C_l - \sum_{s \neq, s \ni l} \bar{A}_s\} \geq \min_{l \in r}\{C_l - \sum_{s \neq, s \ni l} \bar{x}_s^*\} \geq \bar{y}_r^*$$

Thus, we can write $\bar{A}_r = \lambda_r b_r \wedge \bar{y}_r^* = \bar{x}_r^*$. Since $\bar{A}_r$ is unique, it implies that $\bar{A}_r^k \to \bar{A}_r$ uniformly along the whole subsequence K, and we are done.

The results in the third part follows from the second part and the fact that the limit of any convergent subsequence of $(\bar{A}_r^{k,G})$ is a feasible solution to the problem in (2.37). □

Part I

References

Bibliography

[1] Andrew, M. and Zhang, L. The effects of temporary sessions on network performance, Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms, 448 - 457.

[2] Bertsekas, D. and R. Gallager. (1992). Data networks. Prentice-Hall, Englewood Cliffs, NJ.

[3] Bonald, T. and L. Massoulie. (2001). Impact of fairness on internet performance. Proc.ACM SIGMETRICS 2001, Boston, MA, June 2001.

[4] Borodin, A., J. Kleinberg, P. Raghavan, M. Sudan, and D. Williamson. (2001). Adversarial queuing theory. Journal of the ACM, 48(1), 13-38. Earlier version appeared in Proceedingsof the Twenty¨CEighth Annual ACM Symposium on Theory of Computing, 376-385, 1996.

[5] Bramson, M. (1998). Stability of two families of queueing networks and a discussion offluid limits. Queueing Systems: Theory and Applications, 23, 7-31.

[6] Bramson, M. (1999) A stable queueing network with unstable fluid model. The Annals od Applied Probability, 9, 818-853.

[7] Bu, T. and Towsley, T.,2001. Fixed point approximations for TCP behavior in an AQM network, Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, 216-225.

[8] Cao,Z. and Zegura, E.W. (1999). Utility max-min: an application-oriented bandwidth allocation scheme. INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2, 793-801

[9] Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks:Work-conserving discipline. Annals of Applied Probability, 5, 637-655.

[10] Chen, H., O. Kella and G. Weiss. (1997). Fluid approximations for a processor-sharingqueue. Queueing Systems, Theory and Applications 27, 99-125.

[11] Chen, H. and A. Mandelbaum. (1991). Discrete flow networks: bottlenecks analysis andfluidf fluid approximations. Mathematics of Operations Research, 16, 408-446.

[12] Chen, H. and D.D. Yao. (2001). Fundamentals of queueing networks: performance, asymp-totics and optimization, Springer-Verlag New York, Inc.

[13] Chen, H. and Ye, H.Q. (2002). Piecewise linear lyapunov function for the stability of multiclass priority fluid network. IEEE Transactions on Automatic Control, 47, 564-575.

[14] Cooper, W.L., 2002. Asymptotic Behavior of An Allocation Policy for Revenue Management, Operations Research, Vol.50, No.4, 720-727.

[15] Cruz, R.L. (1991a). A calculus for network delay, part I: network elements in isolation.IEEE Trans. Information Theory, 37, 114-131.

[16] Cruz, R.L.(1991b). A calculus for network delay, part II: network analysis. IEEE Trans. In-formation Theory, 37, 132-141.[11] Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: a unifiedapproach via fluid models. Annals of Applied Probability, 5, 49-77.

[17] Dai, J.G. and G. Weiss. (1996). Stability and instability of fluid models for re-entrant lines.Mathematics of Operations Research, 21, 115-134.

[18] Dai, J.G. and S.P. Meyn. (1995). Stability and convergence of moments for multiclassqueueing networks via fluid models. IEEE Transactions on Automatic Control, 40, 1899-1904.32

[19] Davis, M.H.A. (1984). Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models . Journal of Royal Statist. Soc. series B, 46, 353-388.

[20] De Veciana, G., T. J. Lee and T. Konstantopoulos. (2001). Stability and performanceanalysis of networks supporting elastic services. IEEE/ACM Transactions on Networking,9, 2-14. Earlier version appeared in the Proceedings of 18th IEEE INFOCOM¡¯99, NewYork, March 1999.

[21] Fayolle, G., A. de la Fortelle, J.-M. Lasgouttes, L. Massoulie, and J. Roberts. (2001).Best effort networks: modeling and performance analysis via large network asymptotics.Proceedings of 20th IEEE INFOCOM, Anchorage, Alaska, April 2001.

[22] Feng, Y. and Xiao, B.,2000a. A Continuous-Time Yield Management Model with Multiple Prices and Reversible Price Changew, Management Science, Vol.46, No.5, 644-657

[23] Feng, Y. and Xiao, B.,2000b. Optimal Policies of yield management with multiple predetermined prices, Operations Reseach, Vol.48, No.2, 332-343.

[24] Gallego, G. and Van Ryzin, R.,1997. A Multiproduct Dynamic Pricing Problem and its applications to Network Yield Management, Operations Research, Vol. 45, No.1, 24-41.

[25] Gallego,G. and Van Ryzin,G.,1994. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons, Management Science, Vol. 40, No.8, 999-1020.

[26] Gamarnik, D. (2000). Using fluid models to prove stability of adversarial queueing networks. IEEE Transactions on Automatic Control, 45, 741-747.

[27] Glover, F., Glover, R., Lorenzo, J., McMillan, C., 1982. The passenger-mix problem in the scheduled airlines, Interfaces, Vol. 12, No.3, 73-79.

[28] Gromoll, H. C., A. L. Puha and R. J. Williams. (2002). The Fluid Limit of a Heavily-Loaded Processor Sharing Queue. Annals of Applied Probability, 12, 797-859.

[29] Hurley, P., J.Y. Le Boudec and P. Thiran. (1999). A note on the fairness of additive increase and multiplicative decrease. Proceedings of 16th International Teletraffic Congress (ITC-16), Edinburgh, UK, June 1999.

[30] Jacobson, V. (1988). Congestion avoidance and control. Proceedings of the ACM SIGCOMM 88 Conference, 314-329.

[31] Kelly, F. P.,1988. Routing in Circuit-Switched Networks: Optimization, Shadow Prices and Decentralization. Advances in Applied Probability, Vol.20 , 112–144.

[32] Kelly, F.P (1991). Network routing. Phil. Trans. Roy. Soc. Ser. A337 , 343-367.

[33] Kelly, F.P. (1997). Charging and rate control for elastic traffic. European Transactions onTelecommunications, 29, 1009-1016.

[34] Kelly,F.P. (2001). Mathematical modeling of the Internet, in B. Engquist and W. Schmid(ed.), Mathematics Unlimited - 2001 and Beyond, 685-702, Springer-Verlag, Berlin.

[35] Kelly, F.P., A. Maulloo and D. Tan. (1998). Rate control in communication networks:shadow prices, proportional fairness and stability. Journal of the Operational ResearchSociety, 49, 237-252.

[36] Kelly, F.P. and R. J. Williams. (2003). Fluid model for a network operating under a fairbandwidth-sharing policy. Annals of Applied Probability, to appear.

[37] Lautenbacher, C. J., Stidham, S., 1999. The underlying Markov decision process in the single-leg airline yield management problem. Transportation Science. Vol.33, 136-146.

[38] Lee, T. C., Hersh,M., 1993. A model for dynamic airline seat inventory control with multiple seat bookings. Transportation Science. Vol.27, 252-265.

[39] Li, X. and Yao, D.D.,2004. Control and Pricing in Stochastic Networks with Concurrent Resource Occupancy, ACM SIGMETRICS Performance Evaluation Review, Vol.32 , Issue 2,50 -52.

[40] Liang, Y., 1999. Solution to the continuous time dynamic yield management model. Transportation Science. Vol.33, 233-256.

[41] Low, S. (2003). A Duality Model of TCP and Queue Management Algorithms. IEEE/ACMTransactions on Networking, to appear.

[42] Luenberger, D.G. (1984). Linear and nonlinear programming, 369-371. Addison-WesleyPublishing Company, Reading, Massachusetts.

[43] Massoulie, L. and J.W. Roberts. (1999). Bandwidth sharing: objectives and algorithms.Proceedings of IEEE INFOCOM 99.

[44] Massoulie, L. and J.W. Roberts. (2000). Bandwidth sharing and admission control forelastic traffic. Telecommunication Systems, 15, 185-201. Earlier version appeared in Proceedings of ITC Specialist Seminar, Yokohama, 1998.

[45] Mathis, M., J. Semke, J. Mahdavi, and T. Ott. (1997). The macroscopic behavior of theTCP congestion avoidance algorithm. ACM Computer Communication Review, Vol. 27,No. 3.33

[46] Mcgill, J.I. and Van Ryzin, G., 1999. Revenue Management: Research Overview and Prospects, Transportation Science, Vol.33, No.2, 233-256.

[47] Mo, J. and J. Walrand. (2000). Fair end-to-end window-based congestion control.IEEE/ACM Transactions on Networking, 8, 556-567.

[48] Paxson, V. and S. Floyd. (1995). Wide-area traffic: the failure of Poisson modeling.IEEE/ACM Transactions on Networking, Vol. 3, No. 3, pp. 226-244. Also appeared inSIGCOMM 94, pp. 257-268, August 1994.

[49] Peng, S.Y. 1999. Dynamic Pricing in Airline Seat Management for Flights with Multiple Flight Legs, Transportation Science, Vol.33, No.2, 192-206.

[50] Rényi, A 1970. Probability Theory. North-Holland Publishing Company, Amsterdam.London.

[51] Rieman, M.I., Wang, Q., 2006. An asymptotically optimal policy for a quantity-based network revenue management problem. Working Paper.

[52] Rudin, W. (1987). Real and complex analysis (3rd ed.). McGraw-Hill Inc., New York, USA.

[53] Rybko, A. N. and A.L. Stolyar. (1992). Ergodicity of stochastic processed describing theoperations of open queueing networks. Problemy Peredachi Informatsii, 28, 2-26.

[54] Simpson, R.W., 1989. Using network flow techniques to find shadow prices for market and seat inventory control. MIT Flight Transportation Laboratory Memorandum M89-1, Cambridge, MA.

[55] Stolyar, A. L. (1995). On the stability of multiclass queueing network: a relaxed sufficientcondition via limiting fluid processes. Markov Process and Related Fields, 1, No. 4, 491-512.

[56] Subramanian, J., Stidham, S. Lautenbacher, C., 1999. Airline Yield Management with Overbooking, Cancellations and No-Shows, Transportation Science, Vol.33, No.2, 147-167.

[57] Talluri, T.K., 1993. Airline revenue management with passenger routing control: a new model with solution approaches, International Journal of Services Technology and Management, Vol. 2, No.1-2, 102-115.

[58] Talluri, T.K., Van Ryzin, G., 1999. The theory and practice of revenue management. Kluwer Academic Publishers.

[59] Williamson, E. L., 1992. Airline network seat control. Ph.D Thesis, MIT, Cambridge, MA.

[60] Whitt, W.,1985. When Service is required from several facilities simultaneously. AT&T Technical Journal, 64, 1807-1856.

[61] Whitt,W.,1995. When Service is required from several facilities simultaneously. AT&T Technical Journal, 64, 1807-1856.

[62] Yao, D.D. and Ye, H.Q.,2006. Asymptotic Optimality of Threshold Control in a Stochastic Network based on a Fixed-Point Approximation, Proceedings of the Eighth Workshop on Mathematical Performance Modeling and Analysis.

[63] Ye, H.Q. (2003). Stability of data networks under an optimization-based bandwidth allocation. IEEE Transactions on Automatical control, 48, No. 7, 1238-1242.

[64] Ye, H.Q. and Yao, D.D (2006). Heavy-Traffic Optimality of a stochastic network under Utility-Maximizing Resource Control, Working paper.

[65] Ye, H.Q. and H. Chen. (2001). Lyapunov method for the stability of fluid networks. Operations Research Letters, 28, 125-136.

[66] Ye, H.Q., Ou, J. and Yuan,X.M. (2005). Stability of Data Networks: Stationary and Bursty Models, Operations Research, No.1, 107-125.

[67] Zhao, W., Zheng, Y.S., 2001. A dynamic model for airline seat allocation wih passenger diversion and no-shows. Transportation Science, Vol.35, 80-98.