CHANNEL SCHEDULING FOR OPTICAL COMMUNICATION NETWORK WITH FREQUENCY CONCURRENCY

by

ALBERT KAI-SUN WONG

S.B., Massachusetts Institute of Technology (1982)
S.M., Massachusetts Institute of Technology (1984)
E.E., Massachusetts Institute of Technology (1984)

Submitted to the Department of Electrical Engineering and
Computer Science in partial fulfillment of the
requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1988

Signature of Author ______________________________
Department of Electrical Engineering and Computer Science
February 25, 1988

Certified by ______________________________
Professor Robert S. Kennedy
Thesis Supervisor

Accepted by ______________________________
Professor Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

CHANNEL SCHEDULING FOR OPTICAL COMMUNICATION NETWORKS WITH FREQUENCY CONCURRENCY

by

ALBERT KAI-SUN WONG

Submitted to the Department of Electrical Engineering and Computer Science on February 25, 1988 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

ABSTRACT

This thesis addresses the problem when individual users each have access to only a subset of all frequency channels in an optical communication network. The channel scheduling problem refers to the problem of optimally assigning calls to channels when calls arrive or when channels become available. Both loss networks, where no queueing is allowed, and hold networks, where calls are held in queues, are considered in this thesis. The objective is to minimize blocking probability in loss networks and average queueing delay in hold networks.

A Markovian model is adopted and a number of theoretical results are obtained. For loss networks, we have proven that even under very general assumptions, the optimal scheduling policy must be non-wasting, i.e., not rejecting any unblocked calls. For hold networks, a similar conjecture that no unblocked call should be held in queue is shown to be not generally true. Using sample path comparisons, we have proven the optimality of non-invasive hunting, which means the use of unshared channel first, in any loss or hold networks. For all hold networks, we have proven that no call should be held in queue if it can be assigned to an unshared channel. For what we call the W-hold network, we have also proven the optimality of the select from longer queue rule. Some of these results should have implications in the area of controlled queueing theory.

For loss networks, we have formulated the scheduling problem as a Markov decision problem and Howard's policy iteration method is implemented to solve for the optimal policy for any given access structure. In general, optimal policies are found to be traffic dependent and few general statements can be made. We have derived, analytically or numerically, and compared the performances of some representative loss networks. Results indicate that larger fractional reduction in blocking probability can be obtained by good access structures when traffic is low. Interestingly, results indicate that optimal performance is achieved by access structures that are in some sense asymmetric.

By a simulation program we have also compared the optimal and suboptimal queueing delays in a W-hold network with that in an M/M/2 queue. The fractional reduction in queueing delay is relatively constant at all offered traffic and therefore more significant when traffic is high.

Thesis Supervisor: Dr. Robert S. Kennedy
Title: Professor of Electrical Engineering

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my thesis supervisor, Professor Robert Kennedy, whose guidance has been invaluable to my learning process. I would also like to thank Professor Pierre Humblet and Professor John Tsitsiklis for being my thesis readers and for their many comments that have greatly improved this document. Throughout my years at M.I.T., I have benefited greatly from the interaction with and support from Professor Jeffrey Shapiro, my master thesis supervisor and graduate counselor. To the above persons and to many other faculty members who have toiled to create this excellent education environment for me, I can only express my deepest and heartfelt appreciation. The experience at M.I.T. will always be cherished in my life.

Soung, Greg, Ondria, Evan, Walid - the fellow students in Rm 35-427, have made the past two or three years very enjoyable for me. For everyone other than Soung, may their dissertations be finished soon.

As a personal note, I would like to thank for the help of Ming Li towards the completion of this thesis, and the friendship of Eddie Chong and Philip Rounseville over the years.

Albert Kai-sun Wong
February 1988
Massachusetts

<u>To</u>:

my parents,
my sister Dormei,   and
my brothers Billy and Kai-chi

Contents

Chapter 1

Introduction

As a communication medium, singlemode fiber can offer a transmission bandwidth of $10^{12} Hz$ or more. How this enormous bandwidth can be utilized in future data networks is the subject of a large amount of research. As speed of electronics still lags substantially behind that of fiber bandwidth, some issues are becoming apparent. One of which is the fact that the operation region of individual users shall most likely remain within electronic speed which is limited to a few hundred megabits per second or at most a few gigabits per second. In a very wideband network, therefore, it is highly likely that some form of concurrency, frequency concurrency in particular, would be employed to extend network bandwidth above the electronic speed limit.

1.1 Frequency Concurrency in Optical Communication Networks

We believe that there are at least two basic reasons for employing frequency concurrency in future wideband optical communication networks. First, as we have already mentioned, it is the only way to extend the data rate on a piece of fiber beyond the electronic speed limit. Second, unlike time multipexing, users do not have to operate at increased bit rate which is higher than what is needed for their own data transmission. High speed interface electronics, even if technologically feasible, may not be cost effective.

In optical communication at least two forms of frequency concurrency are of interest. They are:

1. *Wavelength Division Multiplexing (WDM)* - WDM systems, which can be implemented with today's technology, provides immediate motivations for research on multiple channel networks. Employing light sources of fixed but different wavelengths, and relying on frequency selective filters or grating structures for channel discrimination, the frequency separation between channels are relatively large[26,37,40]. But as the transmission window of optical fiber is fairly wide, WDM systems with up to ten wavelength channels are already commercially available.

2. *Optical Frequency Division Multiplexing (OFDM)* - the feasibility of OFDM may eventually depend on the success of tunable, stable, low linewidth light

sources and coherent detection. One can also envision the use of OFDM together with WDM to allow thousands of concurrent frequency channels on a piece of optical fiber.

Another form of frequency concurrency, *Subcarrier Frequency Division Multiplexing (SFDM)*, has also been considered. SFDM multiplexes data on different subcarrier frequencies and is probably not useful in terms of extending network bandwidth.

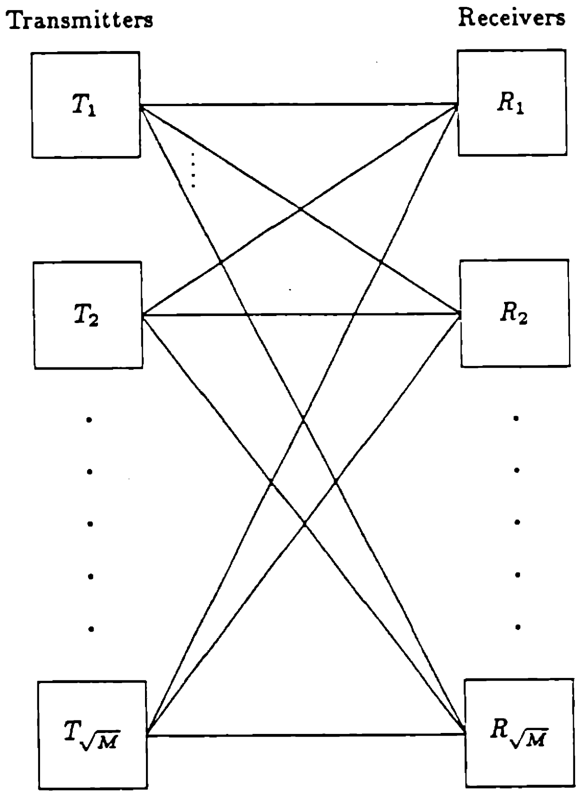
1.2 Limited User Access Capability

Let us consider how users are conventionally connected to a network. In a telephone network, for instance, each user (a telephone) is connected to a switching center through two dedicated wires (or two pairs of wires), one for transmission and one for reception. All switchings are done internally and not by the peripheral users. There is also no need for signal filtering or selection at the peripherals. In a local area network, however, all users are connected to a common transmission medium. Switching is normally not an issue as data are not "localized" and each data unit reaches all users in the network. On the other hand, filtering must be done by each user to extract only the data units in the network that are intended for himself. This filtering is mostly done dynamically, that is, each unit of data transmission is to be examined individually. Therefore, users are required to have processing capacity that can handle all the concurrent traffic in the network.

The focus of our research is on metropolitan area networks where the required bandwidth is presumed to be much higher than that of local area networks. In a future wideband optical network, while network bandwidth can be increased by having multiple frequency channels, the mismatch between the total network bandwidth and the network interface bandwidth will present constraints on network designs. The issue is that limited by the tunable range of lasers and frequency filters, or by cost and practicality, a network interface may no longer be able to tune in to all the channels in the network. Even if transmitters and receivers with wide tunable range become available, the processing bandwidth of network interfaces may still be limited to a small fraction of the total network bandwidth, so that data extraction may become impossible unless some control mechanisms, such as call setups, are employed to ensure that receivers, in particular, are always tuned to the appropriate channels for data reception.

With these considerations in mind, efforts have been made to come up with network designs that can make use of the large fiber bandwidth efficiently while allowing individual nodes to have only limited access capability. Several design approaches have been proposed:

1. Direct connectivity with no switch - We can guarantee logical connectivity between any two nodes by schemes such as one proposed by Marhic, Birk and Tobagi[23]. In their proposed network, as shown in Figure 1.1, each user can transmit on $\sqrt{M}$ channels and receive on a different set of $\sqrt{M}$ channels, where M is the total number of channels. Thus the bandwidth of user interfaces only has to be $\sqrt{M}$ while the total network bandwidth is M.



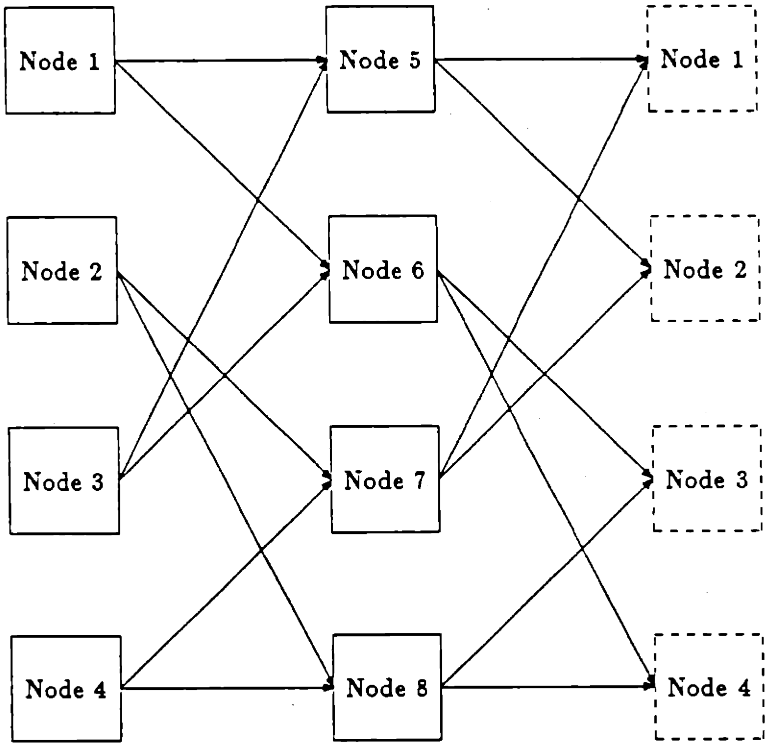
Number of Channels = M

Figure 1.1: Direct Connectivity Network

2. Distributed Switching - If we view each frequency channel as a subnetwork we can provide gateways to relay data from one channel to another, just as gateways do for inter-connecting local area subnetworks. Multiple relays may be necessary and the processing of data at each relay will add to the total delay. Acampora has proposed a so-called perfect shuffle network[2], which is shown in Figure 1.2 for the case where each node can transmit on two channels and receive on two channels. For the general case where each node can transmit on n channels and receive on n channels, the perfect shuffle network has kn^k nodes arranged in such a way that the maximum number of hops it will take from one node to reach another is k. If only one transmitter is allowed to transmit on each channel, the total number of channels is $M = nkn^k$.

3. Centralized switching - In this scheme, all transmitters transmit to the central node which switches data to channels on which respective receivers are listening. Thus for centralized switching it always takes two hops for data to reach the destination. For a very wideband network the central switch must be able to handle extremely high data rate. While efforts are being made to develop all photonic systems, a photonic frequency switch may remain impossible, although there should be no need of becoming overly obsessed with the idea of preserving the photonic nature of the signals throughout the network.

When individual nodes have limited channel access capability the *connectivity* between a transmitter and a receiver, which is defined as the number of common channels that both the transmitter and the receiver can access, will invariably be



k = 2, n = 2, M = 16

Figure 1.2: Perfect Shuffle Network

smaller than M, which is the total number of channels. For example, in the direct connectivity network shown in Figure 1.1, the connectivity between all transmitter receiver pairs is one. Another simple illustration of the limited connectivity problem is the subscriber loop scenario. In the subscriber loop scenario, we are focusing on a part of the network instead of the entire network. Addressing each part of a future metropolitan area network individually may very well be a justifiable approach. Here we are thinking of a local distribution network where one single piece of optical fiber with multiple frequency channels connects a central distribution headend with a large number of local users which are receivers only. We can also imagine the use of a similar network that is called the concentrator network[42] for local access where transmitters talk to a headend that listens on all channels. In the distribution network the receivers are limited in terms of the number of frequency channels that they can access. As shown in Figure 1.3, each receiver has access on, say, only two channels out of the large pool of frequency channels on the fiber. In a distribution network or a concentrator network where either the transmitter end or the receiver end has full channel access capability, the connectivity structure can be most easily seen. In the example shown in Figure 1.3, a particular "access structure" is drawn. That is, each receiver can access two adjacent frequency channels and each channel is shared by two receivers. The sharing of channels by different users that have different sets of accessible channels is the basis of this thesis research. One can have drawn the distribution network in such a way that the receivers listen on non-overlapping sets of channels. In which case the scheduling problem that we are to investigate will not exist.

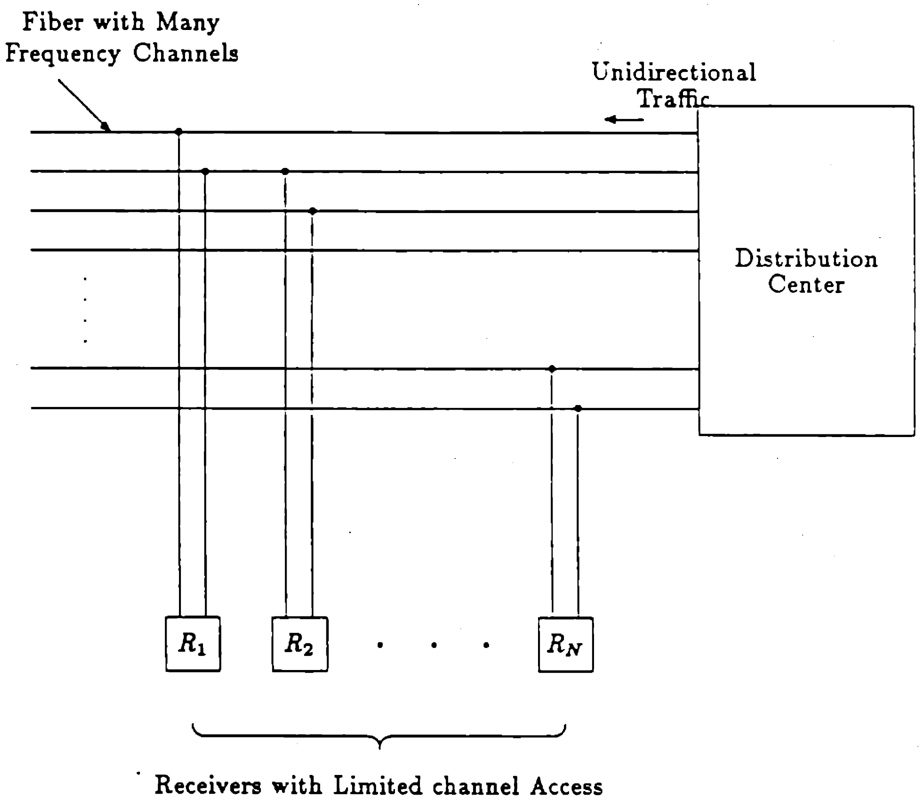


Figure 1.3: A Limited Access Subscriber Loop Scenario

1.3 The Channel Scheduling Problem

To address all the issues involved in a future optical communication network in one thesis is impossible. Recognizing the association of the cost of transmitters and receivers with channel access capability, Soung Liew has investigated the capacity assignment problem in non-switching multichannel networks[22]. His problem is to find the lowest cost solution which meets certain traffic requirements for each node pair. The cost that he is concerned with is assumed to be a function of the total number of channels, the total number of transmitters, and the total number of receivers (each of his transmitter and receiver can access only one channel). The basic relationship between the three parameters is that in a non-switching network, when receiver access capability is small, transmitter access capability has to be large and the number of channels should be kept small to the degree that it does not require splitting up traffic requirements of many transmitter-receiver pairs into smaller fractions and allocating them to multiple channels. His constraints are static in the sense that as long as the capacity allocated to each node pair is above the traffic requirement between the pair, and the total traffic in each channel is below unity, the solution is acceptable. The random nature of traffic is ignored and performance measures such as blocking probability or queueing delay are not considered.

In this thesis a different view is taken. We take into consideration the random nature of traffic and regard blocking probability or queueing delay as important measures of network performance. In addition to ensuring that users are logically connected, the connectivity between users, that is, the number of alternative channels that transmitter receiver pairs can communicate on, becomes an important parameter as it will greatly affect our performance measures. One may suggest that as bandwidth is plentiful, why should network congestion be taken as an important issue? We believe that low channel utilization to the extent that congestion issues can be completely ignored still remains quite unlikely. Although bandwidth on an optical fiber is plentiful, the switching and interface complexities involved in making use of the bandwidth is costly. Therefore, unless revolutionary design concepts are achieved, the bandwidth cannot be made use of for free.

The model we use, as explained in Chapter 2, is to treat all traffic as single hop calls which can be transmitted over different sets of channels. Thus calls are categorized into different classes according to the set of channels on which they can be transmitted. A class of call may represent the traffic between all transmitter and receiver pairs that have the same set of common channels. Or in the case of the distribution network, a class of call will represent all the traffic intended for a particular receiver. The focus of this thesis is to investigate how channel usage should be scheduled in such networks and how performance would vary with the choice of the scheduling method and with the connectivity between nodes.

Given an "access structure", the question that naturally arises is as follows: Assume that we always know the state of the network, which may have to be described by the set of busy channels and the number of waiting calls to be transmitted in each class, how should transmissions on channels be scheduled? That is, which channel should we use to transmit a particular call and which call should we transmit first?

We can see that the answer to the above question depends on the particular access structure we have. Therefore, the next question we naturally ask is how should the access structure be designed to start with? More importantly, we would like to address the issue of how much performance improvement can in fact be achieved and how much additional complexity will be introduced, as compared to the simple access structure where there is no sharing of channel among classes.

Both "loss networks" and "hold networks" are considered in this thesis. In loss networks there is no queueing and a call that is not assigned to a channel is rejected and will never return. In hold networks a call that is not assigned to a channel immediately will wait in a queue until it is assigned. The channel scheduling problem for loss networks is always simpler and is where more results are obtained. The channel scheduling problem for loss network is later found out to be identical to the limited availability problem addressed in telephony[32]. Most previous works in this area are very old and the focus is mostly on so-called "grading" structures with large number of channels. In this thesis, however, the focus is on simple access structures with small number of channels. It is conceivable that in the near future, the number of channels in a metropolitan area network will most likely be not very large. Furthermore, even when the number of channels is in fact very large, they can always be divided into smaller groups and the results obtained for networks with small number of channels may still be applicable. Some of the theoretical results obtained, though intuitively obvious in some cases, appear to be original and have implications in the controlled queueing area.

1.4 Organization of Thesis

The organization of this thesis is as follows: A more precise model and definition of terms will be given in Chapter 2. The first theoretical result is presented in Chapter 3. That is, the proof of the so-called non-wasting property of any optimal scheduling policy for any loss networks. For loss networks, the non-wasting property means that no unblocked call should ever be rejected. We shall also show that this result can be extended to the very general case when service times and arrivals are independent but not memoryless. For hold networks, where a non-wasting policy means one in which no unblocked call is ever held in queue, we shall show that such a policy is not necessarily optimal.

We then in Chapter 4 outline the formulation of the loss network channel scheduling problem as a Markov decision problem. Results in Chapter 3 imply that we only have to look for an optimal solution among the set of non-wasting policies. Some numerical results are presented and the optimal channel scheduling policy is shown to be not robust in the sense that it may change when the total traffic density changes, even when the access structure remains the same.

In Chapter 5 more theoretical results are presented. What we call non-invasive hunting in this thesis refers to the intuitively sensable rule of always using an unshared channel before using a shared channel. Indeed we shall prove that non-invasive hunting fully describes and is the optimal scheduling policy for all simple sharing loss networks. We shall also prove that non-invasive hunting must be obeyed in all optimal policies for loss networks. Similar results are obtained for non-invasive hunting in hold networks. In addition, we shall also prove that the optimal policy in a so-called W-hold network must obey the unshared channel non-wasting rule and the select from longer queue rule. These two rules respectively means that no call should ever be held in queue if it can be assigned to an unshared channel immediately and that when the shared channel becomes available, the next call to be transmitted on the shared channel should be from the longer queue, if one is to be transmitted immediately.

In Chapter 6 and 7 we focus on what we call "uniform-accessibility" networks, or networks in which the access capability, or accessibility, of all classes of calls is the same. The idea is to view the accessibility as a constraint, and try to find out what is the optimal access structure such that when optimal scheduling is used, the minimum blocking probability will be achieved. First, in Chapter 6 the ideal grading access structure is discussed. It is studied on the ground that it is analyzable. We have derived an upper and lower bound on its blocking probability, which is also compared with that of no-sharing networks. The intention is to give us some idea of at least how much reduction in blocking probability is possible. In Chapter 7, through the study of the simplest case when the number of channels is 3 and the accessibility is 2, we reach the interesting and somewhat counter-intuitive conclusion that optimal performance is achieved by some asymmetric access structures.

in Chapter 8, with a simulation program, we shall compare the queueing delay that can be achieved by the W-hold network with the use of optimal and sub-optimal non-wasting scheduling rules, with that of a no-sharing network.

Finally, results in this work will be summarized in Chapter 9 and their implications to network design will be briefly discussed.

Chapter 2

Problem Model

2.1 Basic Network Model

The basic model of our network is as follows. The communication network that we refer to in this thesis consists of M channels that are identical in all their characteristics. Data transmissions in the network are *calls* whose durations are independently and exponentially distributed with mean normalized to unity. As different transmitter and receiver pairs have different sets of common channels on which calls can be sent, we categorize calls into *classes* according to the set of channels on which they can be transmitted. The total number of classes will be denoted by K. The set of channels on which a class k call can be transmitted will be called the *accessible group* of class k calls, or AG_k. The number of channels in AG_k will be called the *accessibility* of class k and will be denoted by $|AG_k|$. In most of our work we shall assume that the accessibilities of all classes are the same and are equal to N_b, which can be viewed as the bandwidth of network interfaces or the connectivity between transmitter and receiver pairs. As the accessible groups of different classes of calls are distinct by definition, with the assumption that all accessible groups are of size N_b, the maximum possible number of classes would be $C(M, N_b)$, where $C(M, N_b)$ represents M choose N_b, or $\frac{M!}{(M-N_b)!N_b!}$. We assume that the arrival of each class of calls is independent of each others and is Poisson of rate λ_k for class k. Thus the total arrival rate, or the total traffic density of the network, will be

$$\Lambda = \sum_k \lambda_k$$

For a given Λ, the topology or *access structure* of the network will be fully described by specifying all the accessible groups (i.e., AG_k for $k = 1, 2, \ldots, K$) and the *traffic splitting coefficients*, ρ_k among all classes, where

$$\rho_k = \lambda_k / \Lambda$$

The network parameters are summarized as follows:

- Number of channels in network $= M$.
- Total number of classes of calls $= K$, each class k having a distinct accessible group of channels, represented by the set AG_k.
- Total offered traffic $= \Lambda$.
- Offered traffic to class $k = \lambda_k = \rho_k \Lambda$.
- Call arrivals are assumed to be Poisson and call durations are assumed to be

independently and exponentially distributed with mean equal to unity for all classes.

2.2 Loss Networks and Hold Networks

In a *loss network* we assume that incoming calls that are not assigned to some idle channel for transmission upon arrival are rejected from the network and will never return. In a *hold network* we assume that calls that are not assigned to channels immediately upon arrival are held in queues until they are assigned. In addition, we shall assume that service is nonpreemptive (i.e., established calls cannot be interrupted or relocated to other channels). With the above assumptions our system becomes a controlled Markov Process and the state of the network can be fully described by $n_k : k = 1,2,3,\ldots K$ and I, where n_k is the number of calls waiting in class k queue, or queue k, and where I is the set of idle channels. With infinite waiting rooms in queues the total number of states is infinite. The queueing model that we have described above is shown in Figure 2.1. An arrow connecting a class k queue with the mth channel means that channel m is within the accessible group of class k.

When a loss network is considered there will be no queue and the state can be fully described by the set of idle channels I. With M channels, the total number of states will be 2^M.

For a loss network where there is no queueing our objective will be to minimize

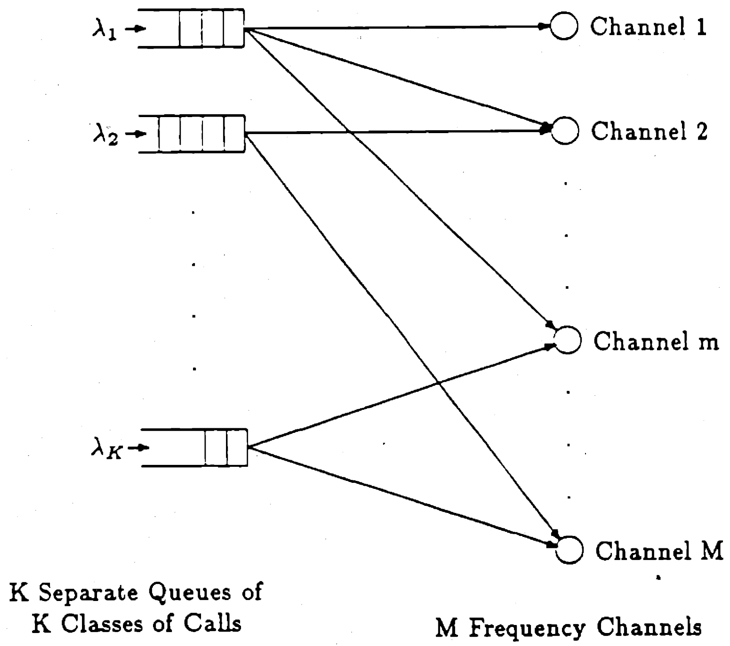


Figure 2.1: A Hold Network with Arbitrary Access Structure

overall probability of loss or blocking probability, P_b, which is

$$P_b = \lim_{T \to \infty} \frac{N_r(T)}{N_a(T)} \tag{2.1}$$

where $N_r(T)$ is the total number of rejected calls up to time T and $N_a(T)$ the total number of arrivals up to time T.

When the network is in state S such that I is the set of idle channels and there is a class k call arrival, we must either select a channel from $AG_k \cap I$ to transmit the call or have the call rejected. This is called the *channel selection problem.* $AG_k \cap I$ is simply the set of channels that are in the accessible group of class k and are idle.

For a hold network the objective will be to minimize expected queueing delay. Besides the channel selection problem, for the hold network there is also the *call selection problem*, which is to choose among queues of different classes of waiting calls the next to be transmitted when a channel accessible to these classes becomes idle when a call previously in service departs.

2.3 Examples: The W-loss Network and the W-hold Network

We shall further illustrate the channel scheduling problem by the W-loss network and the W-hold network examples shown in Figure 2.2 and Figure 2.3 respectively.

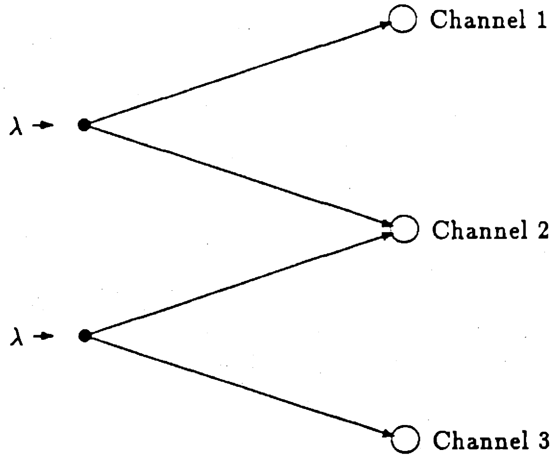


Figure 2.2: The W-loss Network

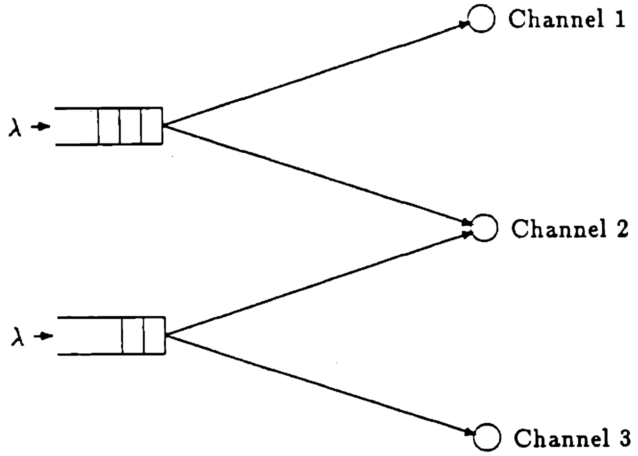
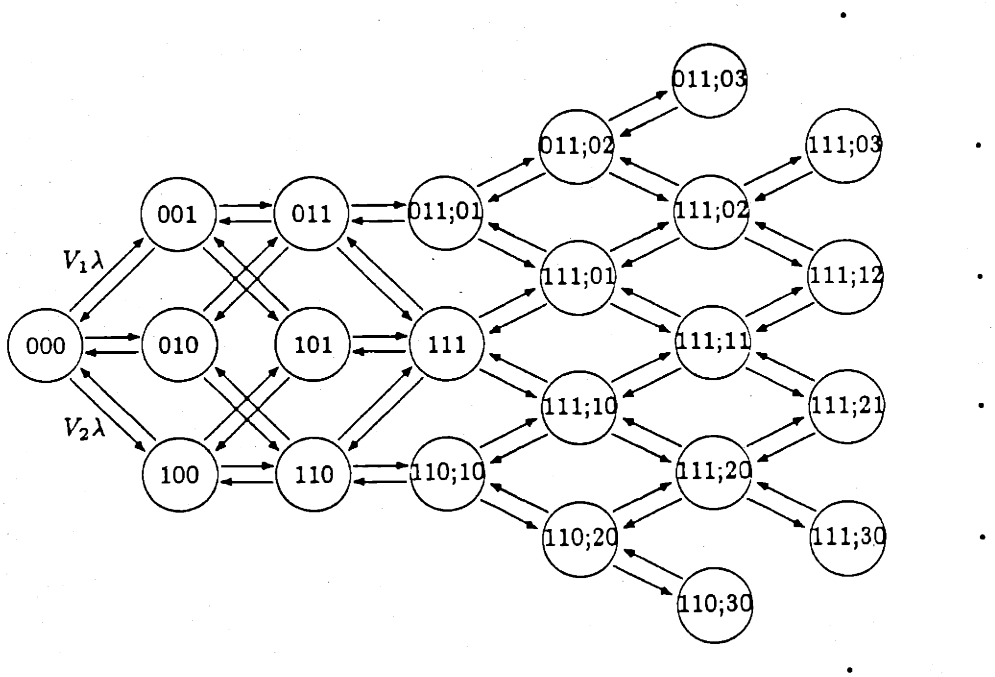


Figure 2.3: The W-hold Network



All Right-to-Left Transition Rates are 1
Left-to-Right Transition Rates are Policy Dependent

Figure 2.4: Corresponding Markov Chain for the W-hold Network

In this simplest non-trivial network that one can think of, there are three channels and two classes. Accessible group 1 consists of channels 1 and 2, and accessible group 2 consists of channels 2 and 3. Therefore channel 1 and 3 are unshared channels and channel 2 is a shared channel. Arrival rates of both classes are assumed to be equal. That is:

- $M = 3$
- $K = 2$
- $AG_1 = (1,2)$
- $AG_2 = (2,3)$
- $\lambda_1 = \lambda_2 = \lambda = \frac{1}{2}\Lambda$

With the assumption that $\lambda_1 = \lambda_2 = \lambda$, the corresponding Markov chain of the W-hold network is shown in Figure 2.4. With infinite waiting room, the state space is infinite, and a state $S = (b_1 b_2 b_3; n_1 n_2)$ represents the state where there are n_1 waiting calls in queue 1 and n_2 waiting calls in queue 2, while b_m is 0 if the mth channel is idle and is 1 if the mth channel is busy. For the W-loss network, there is no queue and the Markov chain is truncated along the dashed line. The number of states is given by $N = 2^3 = 8$ and a state $S = (b_1 b_2 b_3)$ will represent the state such that channel m is busy if $b_m = 1$.

For example, when both channel 1 and channel 2 are idle and there is an arrival from class 1, decision has to made whether the assignment should be made to channel 1 or to channel 2. This is the channel selection problem. When there is a call departure from channel 2 while there are waiting calls in both class 1 queue and class 2 queue, decision has to be made whether the next call assigned to channel 2 should be from queue 1 or from queue 2, or if any should be assigned immediately at all. This is the call selection problem. It is therefore apparent that with this Markovian formulation, decisions have to be made only when *events* occur. We will introduce the notation that an event ξ is represented by:

- A_k if it is an arrival from class k,
- D_m if it is a departure from channel m.

2.4 Representation of a Scheduling Policy as a Decision Table

Any channel scheduling policy R can thus be viewed as a decision table such that

$$R = [r(i, \xi)]$$

where the entry $r(i, \xi) = j$, associated with state i and event ξ, specifies the policy by specifying the next state j that the network should end up in when it is initially at state i and an event ξ occurs. Alternatively, we can specify a scheduling policy by a ***channel selection table*** and a ***call selection table.*** The channel selection table is represented by the matrix F such that

$$F = [f(i, k)] \quad k = 1, 2, \ldots, K$$

with the (i, k)-th element $f(i, k) = m$ specifying the channel m to which assignment should be made when the initial state is i and there is an arrival from class k. The channel selection table alone is sufficient for loss networks. The call selection table is represented by the matrix G, such that

$$G = [g(i, m)] \quad m = 1, 2, \ldots, M$$

with the (i, m)-th element $g(i, m) = k$ specifying the class of the waiting call to be transmitted next when the initial state is i and there is a departure from channel m. In addition, $f(i, k) = 0$ and $g(i, m) = 0$ will respectively mean that no assignment is made when network is in state i and there is a class k arrival or there is a departure from channel m.

The decision table R, or F and G combined will uniquely specify the rate transition matrix of the Markov chain associated with the network. For example, in figure 2.4, the decision variable V_1 is 1 if $f((000; 00), 1) = 1$ and is 0 if $f((000; 00), 1)$ $= 2$, etc. For the loss network as the state space is finite, the number of decision variables is also finite. Each decision variable has a finite number of choices and therefore the set of deterministic channel selection policies is finite. In Chapter 4 we shall show how the channel scheduling problem for the loss system can be formulated as a Markov decision problem. But first in Chapter 3 we shall show that for any optimal scheduling policy, no unblocked calls should be rejected. This implies that $r(i, A_k)$ is limited to at most $|AG_k|$ alternatives for each state i and each class of call k, instead of at most $|AG_k| + 1$ if rejection of unblocked call is allowed. For a M-channels network with constant accessible group size (i.e., $|AG_k| = N_b$, $k =$ 1,2,... K), the total number of alternative policies to be considered is thus at most

$2^M K N_b$.

Chapter 3

Non-Wasting Policy

In this section we shall give the proof of an important property of any optimal channel scheduling policy for a loss network, that is, no call should be rejected if some idle channel can accomodate the call. A class k call is said to be a *blocked* call if and only if it arrives when no channel in its accessible group is idle. It is otherwise said to be an *unblocked* call.

Definition 3.1 For loss networks, a channel scheduling policy is said to be *non-wasting* if and only if no unblocked call is rejected. That is, if network is in state i with I_i being the set of idle channels and there is an arrival from class k such that $AG_k \cap I_i \neq \emptyset$, the arriving call should never be rejected. For hold networks, a channel scheduling policy is non-wasting if and only if no unblocked call is held in queue. That is, n_k must be equal to zero if $AG_k \cap I_i \neq \emptyset$.

3.1 No Rejection of Unblocked Calls in Loss Networks

Theorem 3.1 *In a loss network, optimal channel scheduling policy must be non-wasting.*

3.1.1 Deferred Rejection Rule

Proof of Theorem 3.1:

Assume that there is a policy R under which there exists a state such that an arriving call c is rejected while some channel m that is in the accessible group of c is idle. We will show that improvement can always be made by constructing an alternative policy R', called the *deferred rejection* policy, such that instead of rejecting call c, we assign it to channel m immediately. For all subsequent assignments made under R that do not involve channel m, R' will follow policy R exactly. For the first assignment that involves channel m in R, there are two possibilities:

- (1) Call c may have already departed from channel m. In this case we shall construct R' such that R' will follow policy R and make assignment to channel m.
- (2) Call c may still be occupying channel m. In this case R' will reject the new arrival.

For case (1) it is obvious that one more call would have been served using policy R'. For case (2), the new incoming call, finding channel m busy, will be rejected regardless of whether it can be assigned to some other channels which are idle. Thus this is what we call deferred rejection. Since the service times of calls are independent, exponentially and identically distributed for all classes of calls, the state of the network is fully described by the set of busy channels regardless of the ages (the age of a call is the amount of time the call has already been occupying a channel) of the calls in the network. Therefore the network will always end up in the same state no matter whether case (1) or case (2) is true. Case (1) will always be true with some non-zero probability and thus with some non-zero probability, one more call will have been served using policy R', while there is nothing to lose in case (2) when deferred rejection has to be applied. Therefore improvement is always possible on any policy that rejects unblocked calls. In other words, such a policy cannot be optimal and any optimal policy must obey the non-wasting rule.

QED

3.2 Extension to General Service Time Distribution

In this section we shall demonstrate that for loss networks, the optimality of non-wasting rule can be shown to be true under very general conditions. First let us consider the case when call durations have general but independent and identical distribution. When call durations are not exponentially distributed the network is not memoryless. However, the state of the network can still be represented by

$$S = (b_1 b_2 \cdots b_M; t_1 t_2 \cdots t_M)$$

where t_m represents the age of the call on channel m if there is one and is 0 if channel m is idle. Consider a policy R under which a call c_1 that can be assigned to channel m is rejected at time 0. Assume that we attempt to make a similar proof as in Section 3.1 by constructing a deferred rejection policy R' such that instead of rejecting the unblocked call c_1, assignment is made to channel m immediately. For all subsequent assignments under policy R that do not involve channel m, assume that R' will follow these assignments exactly. Let c_2 be the first call assigned to channel m under R, which occurs at time τ, and let S be the resulting state after c_2 is assigned, such that,

$$R : S = (b_1 b_2 \cdots b_M; t_1 t_2 \cdots, t_m = 0^+, \cdots t_M)$$

If c_1 has already departed from channel m, R' will be able to follow R and assign c_2 to channel m. If this is true both policies will end up in identical states with R' having one extra call assigned and departed and thus better.

If channel m is still busy and if we apply deferred rejection, the resulting state under R' will be:

$$R' : S' = (b_1 b_2 \cdots 1 \cdots b_M; t_1 t_2 \cdots, t'_m > 0, \cdots t_M)$$

S and S' will differ only in that the call c_1 occupying channel m in S' has an older age than the call c_2 that is just assigned to channel m in S, i.e., $t'_m > t_m$.

At first glance, it is not clear that S' is always better than S. It may be worse

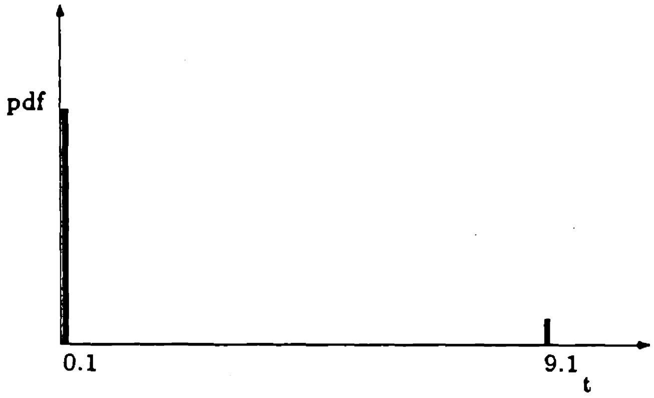


Figure 3.1: Example of an Call Duration Probability Density Function

off to have a call with an older age because the probability density function of call durations may be such that a call having an older age may have a longer expected residual life than a call having a younger age. The simplest example is that when the probability distribution of call durations is such that it is, say, 0.1 with probability 0.9 and 9.1 with probability 0.1, as shown in Figure 3.1. The expected duration of a new call will thus be 1 but the expected residual life of a call with age τ such that $9.1 > \tau > 0.1$ will be 9.1 - τ and will be larger than 1 if τ is less than 8.1.

The above problem disappears if we compare the two policies probabilistically, as to be explained in Section 3.2.1. The idea is that for the duration of c_1 to be t with some probability (or probability density), the duration of c_2 will be t with the same probability also. In a sense, the comparison of R and R' in the proof of

Theorem 3.1 is also a probabilistic one. When deferred rejection is applied the calls occupying channel m in R and R' are different. The problem does not surface as the residual lives of the two different calls have identical distributions. The idea of comparing two sample paths probabilistically is used again in Chapter 5, when we make use of the idea of call duration randomization.

3.2.1 Consistent Deferred Rejection and the V-objective

Theorem 3.2 *In a loss network with general but identical call duration distribution and general arrival process which may or may not be memoryless, optimal channel scheduling rule must be non-wasting as long as call durations are independent of each other and of the arrival process, in addition to the assumption that call durations and arrivals are independent of the scheduling policy.*

Proof:

We shall prove Theorem 3.2 by proving that if there is a policy R that rejects an unblocked call c_1 at time 0 when it can be assigned to some channel m that is idle, we can always construct another policy R' that can always increase the expected number of assignments made up to time T for any $T > 0$.

When the arrival process is general, in addition to the age of each call, the past history of the arrival process must also be included to specify the state of the network. Let us assume that this past history up to time τ can be summarized by $\Phi(\tau)$ so that the state of the network can still be specified. Furthermore, let

$V_l^T(R,S)$ be the probability that l or more assignments are made up to time T given that the initial state is S and policy R is used.

Definition 3.2 The *V-objective* is the probability that l or more assignments are made up to time T given some initial state S and some policy R.

In Chapter 5 we shall make use of a similar quantity that is called the *U-objective.*

Now assume that there is a policy R that rejects an unblocked call c_1 at time 0 when it can be assigned to an idle channel m, we construct a policy R' such that R' assigns c_1 to channel m immediately and follows all subsequent assignments made under R as long as they do not involve channel m. Assume that policy R first makes $n-1$ assignments to channels other than channel m and make first assignment to channel m at time τ and let c_2 be the call assigned. We shall construct R' such that it simply rejects c_2 regardless of whether channel m has become idle. Thus this is what we call *consistent deferred rejection.*

Definition 3.3 The *consistent deferred rejection rule* means that if a policy R rejects a call c_1 that can be assigned to some idle channel m, instead we assign c_1 to channel m immediately but in the future when c_2 is the first call that is assigned to channel m under R, we consistently reject c_2, regardless of whether channel m has become idle.

Let d_1, d_2 be the durations of c_1 and c_2 respectively and let $p_d(t)$ be the probability density function of all call durations.

Let $S_\tau(d_2 = \sigma)$ denote the state at time τ given we know specifically that the duration of c_2 is σ.

Furthermore, let $p_e^\sigma(n)$ be the probability that there are n events up to time σ. $p_e^\sigma(n)$ is identical under both R and R'.

Under R, using iterated expectation, we have for $T > \tau$, $l > n$,

$$V_l^T(R, S_0) = \sum_n p_e^\sigma(n) \int V_{l-n}^{T-\tau}(R, S_\tau(d_2 = \sigma)) p_d(\sigma) d\sigma \tag{3.1}$$

Under R', the resulting state at time τ, S'_τ, is almost the same as S except for channel m. We have,

$$V_l^T(R', S_0) = \sum_n p_e^\sigma(n) \int V_{l-n}^{T-\tau}(R, S'_\tau(d_1 = \sigma)) p_d(\sigma) d\sigma \tag{3.2}$$

where $S'_\tau(d_1 = \sigma)$ is the state at time τ under R' given we know specifically that the duration of c_1 is σ.

For any σ, the state $S'_\tau(d_1 = \sigma)$ is always better or at least as good as the state $S_\tau(d_2 = \sigma)$. The reason is that any subsequent assignments made for $S_\tau(d_2 = \sigma)$ can always be made for $S_\tau(d_1 = \sigma)$, as the two states are the same except that we know channel m will always become idle first for the latter.

Let $N_s(T)$ be the number of assignments made up to time T. Given a policy R, the expected number of assignments up to time T is

$$E\{N_s(T)\} = \sum_{l=1}^{\infty} V_l^T(R, S_0) \tag{3.3}$$

We have,

$$V_1^T(R', S_0) = 1 \tag{3.4}$$

considering c_1 that is already assigned to channel m.

For $l = 1$, there is always some positive probability that no assignment is made in R up to time T, say, when there is no arrival up to time T. Therefore, $V_l^T(R', S_0)$ is obviously maximized for $l = 1$. For $l > 1$, R' is at least as good. Therefore, for any policy R that rejects unblocked calls, we can always construct an alternative policy that can increase the expected number of assignments made up to any time T. In other words, such a policy cannot be optimal. QED

The essence of the proof for Theorem 3.2 is as follows. For the constructed alternative policy R', up to any time T there is always some probability that consistent deferred rejection is not applied and one more assignment is made. While consistent deferred rejection is applied, for each state reached in R, with the same probability a state that is at least as good will be reached in R'.

3.3 Holding of Unblocked Calls in Hold Networks

One may be tempted to believe in similar statement as Theorem 3.1 when hold networks are considered and when performance criterion is the expected queueing time. That is, no call should be held in queue when it can be assigned to some idle channels immediately. It turns out that this conjecture is not true in general, although it may be frequently true in many practical cases. A convincing counter example is given in Figure 3.1.

Let us define the busy period of a class be the period of time during which the queue of the class is non-empty. For class 1 shown in Figure 3.1, which has only one channel in its accessible channel group, such that,

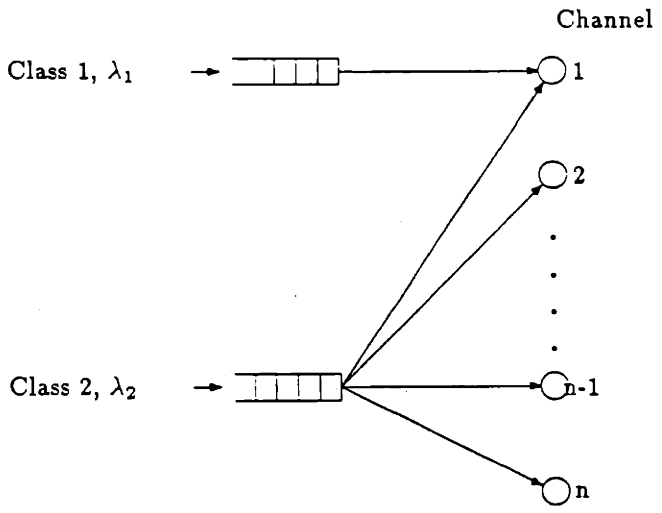
$$AG_1 = (1)$$

assume that arrival rate λ_1 is large (almost 1). Therefore for group 1 the busy periods are long and the idle periods are short. For class 2 which has many channels in its accessible channel group, such that,

$$AG_2 = (1, 2, \ldots, n) \quad \text{where } n \gg 1$$

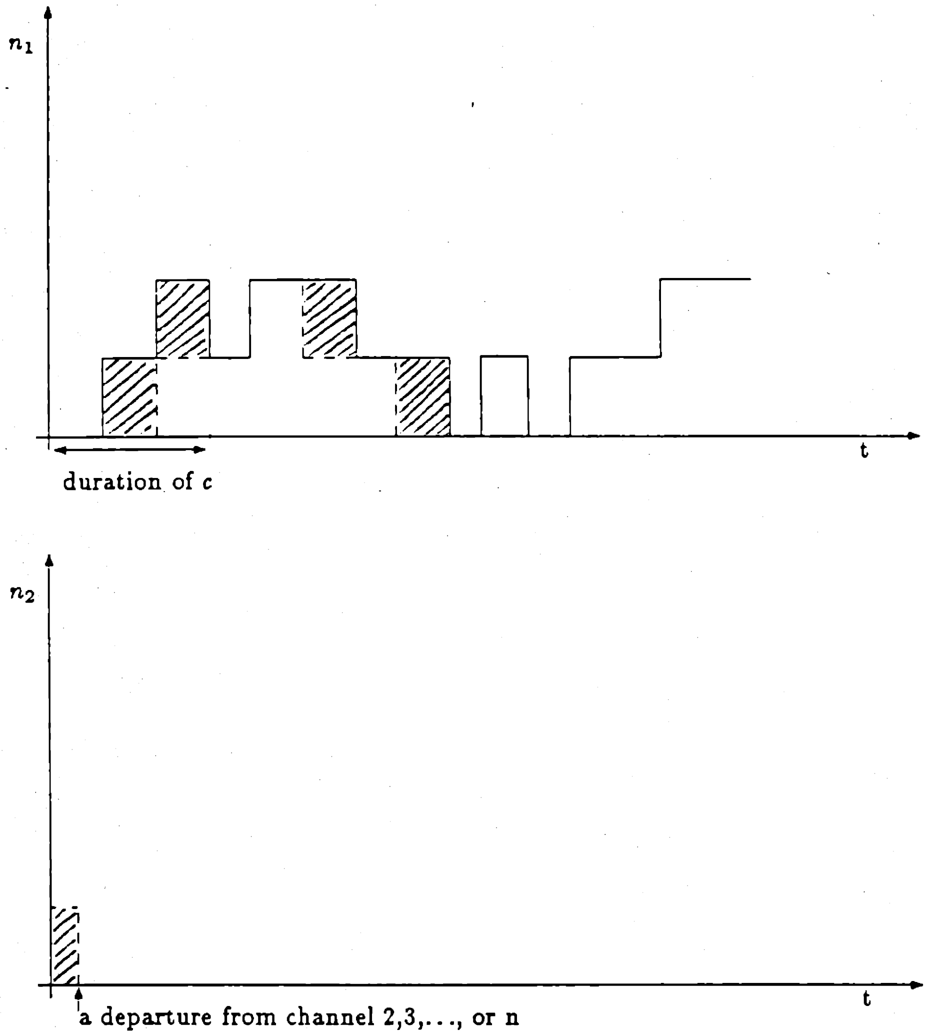
assume that arrival rate λ_2 is small, i.e., $\lambda_2 \rightarrow 0^+$. Therefore, for group 2 the busy periods are short and the idle periods are long.

Now assume we are in the state where all channels are busy except channel 1, which is shared by both group 1 and group 2, and a call c of class 2 arrives. If we obey the rule that no unblocked call should be kept on hold, the arriving call should be served by the shared channel immediately. We can see that this is not optimal since busy period for group 2 is short, the arriving class 2 call will have to wait only a short while, approximately $\frac{1}{n-1}$ units of time on the average, to find a freed unshared channel in group 2, and most likely it will not be blocking any other class 2 calls as idle periods for class 2 are long. Therefore the added delay penalty is only about $\frac{1}{n-1}$. If c is assigned to the channel 1 immediately, the delay penalty on class 1 calls will be large since busy period of group 1 is long and the number of calls which will have to wait additional amount of time because of the assigned call



Assuming $\lambda_1 \gg \lambda_2$, and all channels busy except channel 1

Figure 3.2: Counter Example for No Holding of Unblocked Calls



Shaded regions are added delay penalty for each case

Figure 3.3: Number of Waiting Calls in Queues

on the shared channel will be large.

To make the above argument more quantitative, let us consider the following. Let

$$\lambda_1 = 1 - \epsilon \quad \text{for } \epsilon \text{ some very small positive number}$$

The probability that the first call from class 1 will arrive before c departs will approximately be $\frac{1}{2}$. The expected delay incurred on this class 1 call alone will be 1, which is the expected residual life of c if it is still occupying channel 1. Therefore the total expected delay incurred on class 1 calls is at least almost 0.5. When n is large, the delay incurred by holding the call c can be made arbitrarily small, and the probability that it will be delaying other class 2 calls will be very small. A typical plot of the number of holding calls in class 1 and class 2 is shown in figure 3.2. The dashed lines represent the resulting number of holding calls of each class if call c is held until another channel in group 2 becomes free. The solid lines are when call c is assigned to channel 1 immediately. The shaded region in each case represents the additional penalty of each choice and it is clear that it is better to hold the class 2 call in queue than using up the shared channel.

The contrast between the result in Section 3.1 and the result obtained here illustrates the fact that for the same network with the loss model and the hold model, the optimal channel selection policies are in general not the same, even when only the states with all queues empty are considered in the hold system.

Since we have proven the fact that for any loss network the optimal channel scheduling policy must be non-wasting, in the following chapter when the channel scheduling problem is formulated as a Markov decision problem, there is no need to consider alternatives that reject unblocked calls. Unnecessary complications are thus avoided.

Chapter 4

Markov Decision Formulation of Channel Selection Problem

In this chapter we shall outline the formulation of the channel selection problem for the loss network as a Markov decision problem.

4.1 Imbedded Markov Chain

While the Markov chain describing the state of the network is continuous time in nature, we shall first seek to set up our problem as a discrete time problem. We shall do so by considering the imbedded chain at the occurrence of an event, where again an event is either a call arrival or a call departure. A call arrival may either result in assignment to an idle channel, or result in rejection (which we have proven in the previous chapter that should only occur when all channels in the corresponding accessible group are busy).

4.1.1 Transition Probability Matrix for the Imbedded Chain

If we look at the imbedded Markov chain, the transition probabilities between states are policy dependent and are given as follows:

$$\text{For } j = D(i,m), \qquad p_{ij} = \frac{1}{nbusy(i) + \Lambda} \tag{4.1}$$

$$\text{For } j = r(i,A_k), \qquad p_{ij} = \frac{\sum\limits_{k:r(i,A_k)=j} \lambda_k}{nbusy(i) + \Lambda} \tag{4.2}$$

$$\text{For } j = i, \qquad p_{ii} = \frac{\sum\limits_{k:f(i,k)=\emptyset} \lambda_k}{nbusy(i) + \Lambda} \tag{4.3}$$

where,

- $D(i,m)$ represents the state j which results when initial state is i and there is a departure from channel m.
- $nbusy(i)$ represents the occupancy, or the number of busy channels in state i.

and again,

- A_k represents the event that there is a class k call arrival.

- $r(i, \xi)$ represents the state that the network will end up in given that it is initially in state i and there is an event ξ.
- $f(i, k)$ represents the channel that a class k call will be assigned to when the network is initially in state i. $f(i, k)$ is 0 if class k calls are rejected in state i.
- Λ is the total arrival rate and λ_k the arrival rate of class k calls.

p_{ij} is the probability that given the network is initially in state i, it will next end up in state j. It is also equal to the probability that the occurrence of the event, say event l, that leads to the corresponding transition in the imbedded chain preceeds occurrences of all other events possible in state i, i.e.,

$$t_l < t_1, t_2, \ldots, t_{l-1}, t_{l+1}, \ldots, t_L$$

where t_n is the occurrence time of the nth possible event in state i. As the occurrence time of all events are exponentially distributed and independent,

the probability that event l occurs first equals to:

$$\int_0^\infty \lambda_l e^{-\lambda_l t} Pr(t < t_1, \ldots, t_{l-1}, t_{l+1}, \ldots, t_L) dt$$
$$= \int_0^\infty \lambda_l e^{-\lambda_l t} e^{-\lambda_1 t} \cdots e^{-\lambda_{l-1} t} e^{-\lambda_{l+1} t} \cdots e^{-\lambda_L t} dt$$
$$= \lambda_l / \sum_{n=1}^{L} \lambda_n$$

We can see that only the transition probabilities corresponding to call arrivals are policy dependent. Given a policy, $r(i, \xi)$ and $f(i, k)$ are given and thus p_{ij}, the transition probability matrix

$$P = [p_{ij}]$$

is uniquely specified for the imbedded chain. It is possible, however, that two or more different policies will have the same transition probability matrix. In which case these policies are identical in terms of blocking probability performance. It is also possible that two transition probability matrices look different, but in fact correspond to policies that are identical but for a different labelling of states or channels.

4.1.2 The g-Objective

Let w_{ij} denote the reward associated with the $i-j$ state transition. For each state i, we shall assign a unit reward to each transition that corresponds to a departure, such that

$$w_{ij} = \begin{cases} 1 & \text{if } j = D(i,m) \text{ for some } m \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

For each state i, the expected return is:

$$z_i = \sum_j p_{ij} w_{ij} = \frac{nbusy(i)}{nbusy(i) + \Lambda} \tag{4.5}$$

and is independent of the particular policy used.

For the imbedded chain, the expected return per transition will be:

$$g = \sum_i z_i \pi_i \tag{4.6}$$

where π_i is the steady state probability of state i of the imbedded chain, i.e., the probability of finding the system in state i immediately after an event.

Corollary 4.1 Maximizing g is equivalent to minimizing P_b, the blocking probability.

It is simple to show that maximizing g is equivalent to minimizing P_b. Let $N_a(T)$ be the number of arrivals in the time interval $(0,T)$, $N_d(T)$ be the number of departures in $(0,T)$, and $N_r(T)$ the number of rejected calls. $N_e(T)$, number of events in $(0,T)$ is

$$N_e(T) = N_a(T) + N_d(T) \tag{4.7}$$

and

$$\begin{aligned} g &= \lim_{T\to\infty} \frac{N_d(T)}{N_e(T)} & (4.8)\\ &= \lim_{T\to\infty} \frac{N_d(T)}{N_a(T)+N_d(T)} & (4.9)\end{aligned}$$

By definition, the blocking probability is,

$$\begin{aligned} P_b &= \lim_{T\to\infty} \frac{N_r(T)}{N_a(T)} & (4.10)\\ &= \lim_{T\to\infty} \frac{N_a(T)-N_d(T)}{N_a(T)} & (4.11)\\ &= \lim_{T\to\infty} \frac{(\frac{N_a(T)-N_d(T)}{N_a(T)+N_d(T)})}{(\frac{N_a(T)}{N_a(T)+N_d(T)})} & (4.12)\\ &= \frac{1-2g}{1-g} & (4.13)\\ &= 1-\frac{g}{1-g} & (4.14)\end{aligned}$$

which is monotonically decreasing in g since its derivative is equal to $-\frac{1}{(1-g)^2}$ and is always negative as g must be less than 0.5 (at most half of the events can be departures).

Therefore, maximizing g is equivalent to minimizing P_b, and P_b can be easily calculated after g is found. The advantage of optimizing g instead lies in the fact that the expected immediate return for all states with the same number of busy channels is identical, independent of the scheduling policy, and can be easily calculated as z_i by equation 4.5.

4.2 Linear Programming Formulation of Channel Scheduling Problem

In this section we shall first formulate the channel scheduling problem for loss networks as a primal linear programming problem. Then we shall give the dual linear programming problem view of Howard's policy iteration algorithm.

4.2.1 Primal Problem Formulation

The idea of formulating the channel scheduling problem as a linear programming problem is to randomize the choice of policies[1]. Let x_j^r be the joint probability of being in state j and choosing policy r for state j, and let $R_j = (r_{j1}, r_{j2}, \cdots, r_{jn_j})$ be the set of alternative policies for state j. For the channel selection problem R_j can be considered as a decision vector which specifies channel assignments for all classes of arrivals at state j. We can formulate the channel scheduling problem as a linear programming problem as follows:

[1] see [20]

Linear Programming Formulation

Primal Problem:

$$\min \sum_{j\in S}\sum_{r\in R_j} -z_j^r x_j^r \qquad (\text{or } \max \sum_{j\in S}\sum_{r\in R_j} z_j^r x_j^r)$$

$$\text{subject to}$$

$$\begin{aligned}\sum_{r\in R_j} x_j^r - \sum_{i\in S}\sum_{r\in R_i} p_{ij}^r x_i^r &= 0 \qquad \text{for } j = 1,2,\ldots N-1 \\ \sum_{j\in S}\sum_{r\in R_j} x_j^r &= 1 \\ &\mathit{and} \\ x_j^k &\geq 0 \quad \text{for } j \in S, r \in R_j\end{aligned}$$

The constraints equations are simply the conservation of probability flows and the summation of steady state probabilities to unity. The objective is to maximize $g = \sum_{j\in S}\sum_{r\in R_j} z_j^r x_j^r$. For the channel scheduling problem, we have mentioned earlier that the rewards z_i^r's are independent of the policy. Nevertheless, here we have first written down the general objective where the rewards are policy dependent.

Alternatively in matrix form, we can express the primal problem as:

to minimize

$$[-z_1^{r_{11}} - z_2^{r_{12}} \ldots - z_1^{r_{1n_1}} | - z_2^{r_{21}} \ldots - z_2^{r_{2n_2}} | \ldots | - z_N^{r_{N1}} \ldots - z_N^{r_{Nn_N}}] \begin{bmatrix} x_1^{r_{11}} \\ x_{12} \\ \vdots \\ x_1^{r_{1n_1}} \\ x_2^{r_{21}} \\ \vdots \\ x_2^{r_{2n_2}} \\ \vdots \\ x_N^{r_{N1}} \\ \vdots \\ x_N^{r_{Nn_N}} \end{bmatrix}$$

subject to:

$$\left[\begin{array}{cccc|ccc|c|ccc} 1 - p_{11}^{r_{11}} & 1 - p_{11}^{r_{12}} & \cdots & 1 - p_{11}^{r_{1n_1}} & -p_{21}^{r_{21}} & \cdots & -p_{21}^{r_{2n_2}} & \cdots & -p_{N1}^{r_{N1}} & \cdots & -p_{N1}^{r_{Nn_N}} \\ -p_{12}^{r_{11}} & -p_{12}^{r_{12}} & \cdots & -p_{12}^{r_{1n_1}} & 1 - p_{22}^{r_{21}} & \cdots & 1 - p_{22}^{r_{2n_2}} & \cdots & -p_{N2}^{r_{N1}} & \cdots & -p_{N2}^{r_{Nn_N}} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \end{array}\right] \begin{bmatrix} x_1^{r_{11}} \\ x_1^{r_{12}} \\ \vdots \\ x_1^{r_{1n_1}} \\ x_2^{r_{21}} \\ \vdots \\ x_2^{r_{2n_2}} \\ \vdots \\ x_N^{r_{N1}} \\ \vdots \\ x_N^{r_{Nn_N}} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

and

$$x_j^r \geq 0 \text{ for } j \in S, r \in R_j$$

This problem is a standard linear programming problem and can be solved using the Simplex method but obviously the number of columns will be very large and keeping track of these columns will be difficult.

4.2.2 The Policy Iteration Algorithm

For the primal problem outlined in section 4.2.1, there is a corresponding dual problem:

$$\max \ f$$

$$\text{subject to}$$
$$u_i - \sum_{j=1}^{N-1} p_{ij}^r u_j + g \leq -z_i^r \text{ for } i \in S, r \in R_i$$

Given a deterministic policy R, a transition probability matrix P^R is uniquely specified for the network and a corresponding basic feasible solution can be obtained for the primal problem.

$$\left[\begin{array}{c} I_{N-1} - (P^R)^T_{N-1} \\ \hline 1 \;\; 1 \cdots 1 \end{array}\right] \begin{bmatrix} x_1^{r_1} \\ x_2^{r_2} \\ \vdots \\ x_N^{r_N} \end{bmatrix} = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ \hline 1 \end{array}\right]$$

where the I_{N-1} matrix is the identity matrix with the last row omitted and the $(P^R)^T_{N-1}$ matrix is the transpose of the P^R matrix, and with the last row omitted. Now for ease of notation the r's are single subscripted and r_i represents a particular decision vector for state i.

If the solution is non-optimal, the corresponding basic solution to the dual variables will not be feasible. If the basic solution is also feasible, we know an optimal solution has been reached. That is, if u_i's and f are the solution to

$$\left[\begin{array}{cccc|c} u_1 & u_2 & \ldots & u_{N-1} & f \end{array}\right] \left[\begin{array}{c} I_{N-1} - P^R_{N-1} \\ \hline 1 \;\; 1 \ldots 1 \end{array}\right] = \begin{bmatrix} -z_1^{r_1} & -z_2^{r_2} & \cdots & -z_N^{r_N} \end{bmatrix}$$

and if u_i's satisfy

$$u_i - \sum_{j=1}^{N-1} p_{ij}^r u_j + f \;\; \leq \;\; -z_i^r \;\; for \; i \in S, r \in R_i$$

then R is optimal.

Equivalently, if we solve for $v_i = -u_i$ and $g = -f$, that is,

$$\left[\begin{array}{cccc|c} v_1 & v_2 & \dots & v_{N-1} & g \end{array}\right] \left[\begin{array}{c} I_{N-1} - P^R_{N-1} \\ \hline 1 \;\; 1 \dots 1 \end{array}\right] = \left[\begin{array}{cccc} z_1^{r_1} & z_2^{r_2} & \cdots & z_N^{r_N} \end{array}\right]$$

while obviously now $g = \sum_{i=1}^{N} \pi_i^R z_i^R$.

And if the v_i's satisfy

$$v_i - \sum_{j=1}^{N-1} p_{ij}^r v_j + g \;\; \geq \;\; z_i^r \;\; for \; i \in S, r \in R_i$$

then R is optimal and g is maximized.

The dual problem formulation of the Markov decision problem thus explains the existence of the so-called policy iteration algorithm according to Howard [2], which operates as follows:

1. Value Determination - For a given policy R, solve
$$v_i - \sum_{j=1}^{N-1} p_{ij}^k v_j + g = z_i^k \quad \text{for } i = 1,2,3,\dots,N \;\; v_N = 0$$
2. Policy Improvement - For each state i, find the policy r' that maximizes
$$z_i^{r'} + \sum_{j=1}^{N-1} p_{ij}^{r'} v_j$$
3. Iteration - Go back to (1) until no further improvement is possible.

[2] see [12]

The policy iteration algorithm is equivalent to performing N pivotings at the same time. Howard showed that step (2) will always result in improvement as follows:

Proof:

Since r' maximizes $z_i^{r'} + \sum_{j=1}^{N-1} p_{ij}^{r'} v_j$ for all i,

$$v_i + g \leq z_j^{r'} + \sum_{j=1}^{N-1} p_{ij}^{r'} v_j \quad \forall i$$

Now let v_i', $i = 1, 2, 3, N-1$ be the new values and g' be the new expected reward, such that:

$$v_i' + g' = z_j^{r'} + \sum_{j=1}^{N-1} p_{ij}^{r'} v_j'$$

Therefore, we have:

$$\begin{aligned} (v_i' - v_i) + (g' - g) &\geq \sum_{j=1}^{N-1} p_{ij}^{r'} (v_j' - v_j) \\ \Rightarrow (v_j' - v_i) + (g' - g) &= C_i + \sum_{j=1}^{N-1} p_{ij}^{r'} (v_j' - v_j) \end{aligned}$$

where

$$C_i \geq 0 \qquad \forall i$$

therefore,

$$(g' - g) = \sum_{i=1}^{N} \pi_i' C_i \geq 0$$

where π_i''s are the steady state probabilities associated with the Markov chain with transition probabilities $p_{ij}^{r'}$ and must therefore be non-negative. As $(g' - g) \geq 0$, improvement is made. **QED**

From linear programming we know that it is impossible for an optimal policy to remain undiscovered. Since z_i^r's are independent of the policy, finding policy improvements corresponds to assigning channels to each class of calls to reach states with largest present values of v_i. This can be done by considering alternative channel assignments and reconstructing the probability transition matrix at each iteration without storing all alternative transition probabilities for all alternative policies. The policy iteration algorithm is implemented to solve the channel selection problem with the purpose of trying to reach some conclusions about the characteristics of the optimal channel selection policy. Some results are presented in the next section.

4.3 Results from the Policy Iteration Algorithm

OPTM.F77 is the FORTRAN program written to find the optimal solution to the channel selection problem for loss networks using Howard's policy iteration algorithm. The program can handle up to seven channels and twenty classes of calls. Input to the program is contained in the file "INDATA" and output is generated in the file "OUTDATA". In this section we shall present results from some sample cases.

4.3.1 The W-Loss Network

For the W-loss network shown in Figure 2.2, the optimal channel selection table $f(i,k)$ is found for different α's, the offered load per channel. A sample output file is shown in Figure A.1 in the appendix. The output file contains both the channel selection table F and the policy table R. The value of each state i, v_i, is also listed. The last entry in the v_i's column, the 8th one in this case, is in fact the value of g. Note that all the v_i's are negative. This is not surprising as z_8 is the largest for state 8 and v_8 is assumed to be zero.

It is found that the channel selection table for the W-loss network is independent of α and is shown in Table 4.1.

S = $(b_1 b_2 b_3)$	f(i,1)	f(i,2)
(000)	1	3
(100)	2	3
(010)	1	3
(001)	1	2
(110)	0	3
(101)	2	2
(011)	1	0
(111)	0	0

Table 4.1: Optimal Channel Selection Table for W-Loss Network

From the channel selection table for the W-loss network, we can summarize the optimal channel selection policy by what we call the non-invasive hunting rule. That is, when a call arrives, assignment is always made to an unshared channel first if possible. As we know implicitly that only non-wasting policies are of interest, and that there is only one shared and one unshared channel for each class, the specification that the unshared channel should always be chosen first completely describes the optimal policy. For more general access structures in which shared channels are shared among different classes, non-invasive hunting will not be sufficient to describe the optimal policy. The optimality of non-invasive hunting for the W-loss network is intuitively quite obvious. In Chapter 5, a rigorous proof is given.

4.3.2 Variability of Optimal Policy with Respect to Total Traffic Density

Given the result in section 4.3.1, one may wonder if the optimal policy is determined by the access structure alone. Our numerical results have shown that it is not true. When we maintain the same access structure (i.e., with all the accessible groups and traffic splitting coefficients remaining the same) but vary the total traffic density Λ, the decision table is found to be changing in general. We illustrate this result by the following example:

Let

- $AG_1 = (1, 2)$
- $AG_2 = (2, 3)$
- $AG_3 = (3, 4)$
- $AG_4 = (4, 5)$

- $AG_5 = (5, 6)$
- $AG_6 = (6, 1)$
- $AG_7 = (1, 3)$
- $\rho_k = \frac{1}{7}$, for $k = 1, 2, \ldots, 7$

For this network of six channels and seven classes with equal traffic splitting, we obtain the optimal channel selection tables for different values of Λ, the total traffic density. Shown in Figure A.2 and Figure A.3 are the results for the cases when $\Lambda = 0.1$ and $\Lambda = 0.9$. To save space the output files are only shown in part. The two channel selection tables are different in quite a few places and we can also see that the difference is not an aberration due to the existence of symmetical policies.

For instance, for states 5 and 6 such that

$$S_5 = (000100) \quad \text{and} \quad S_6 = (000010)$$

we have for $\Lambda = 0.1$,

$$v_5 = -2.34685$$

and

$$v_6 = -2.34697$$

While for $\Lambda = 0.9$, we have

$$v'_5 = -1.47463$$

and

$$v'_6 = -1.47302$$

While for $\Lambda = 0.1$, $v_5 > v_6$, for $\Lambda = 0.9$, $v_6' > v_5$. Therefore if initially we are in state 1 such that $S_1 = (000000)$ and there is a class 4 arrival, the better state to go to will be state 5 when $\Lambda = 0.1$ and will be state 6 when $\Lambda = 0.9$.

The result here indicates that in general, the optimal channel scheduling policy cannot be deduced from the access structure alone. Therefore there are very few general characterizations that we can really make for the optimal channel scheduling policy.

In the next chapter, however, the optimality of a policy called non-invasive hunting that is independent of traffic density is proven first for the W-loss network. Extensions are then made for other types of networks.

Chapter 5

Non-invasive Hunting

In this chapter we will first prove the optimality of *non-invasive hunting* for the W-loss network and then for all loss networks. For all *simple sharing networks*, non-invasive hunting is sufficient to completely specify the optimal policy. Second, we will prove the optimality of certain scheduling rules for the W-hold Network.

We hereby state the following definition:

Definition 5.1 *Non-invasive hunting* means that if possible, assignment will always be made to an unshared channel first. In other words, assignment to shared channel will only be made when an arriving call finds all the unshared channels in its accessible group busy.

5.1 The W-loss Network

We have already proven in Chapter 3 that for any loss network the optimal channel scheduling policy must be non-wasting. Here we will show that the optimal policy can indeed be completely specified for simple network structure such as that of the W-loss network.

Theorem 5.1 *Non-invasive hunting is optimal for the W-loss network.*

5.1.1 The U-objective and Blocking Probability

We shall prove Theorem 5.1 by inductively showing that for all l, T, and S_0, non-invasive hunting maximizes what we call the *U-objective*, which is:

Definition 5.2

$$U_l^T(R, S_0) = \begin{array}{l} Pr\{N_e(T) \geq l \\ \qquad \text{given that the initial state is } S_0 \text{ and policy } R \text{ is employed}\} \end{array}$$

where $N_e(T)$ is the total number of events up to time T while an event ξ is defined as either a call arrival or a call departure at service completion.

For the W-loss network the channel selection policy is completely specified by the non-invasive hunting rule as all choices involves only two channels, one shared and one unshared. What we are going to show is that non-invasive hunting maximizes the U-objective over the set of stationary and non-stationary policies.

Assume that there is a T, l dependent policy

$$R^{Tl} = [r^{Tl}(S, \xi)]$$

and let Γ be the complete specification of R^{Tl} for all T and l. We have,

$$U_l^T(\Gamma, S) \quad = \quad \int_\Omega p_{\omega|S}(\xi, \tau|S) U_{l-1}^{T-\tau}(\Gamma, r^{(T-\tau)(l-1)}(S, \xi)) d\omega \tag{5.1}$$

In general, maximizing $U_l^T(R, S_0)$ is a continuous time finite-horizon Markov decision problem and the optimal decision table

$$R_{opt}^{\tau L} = \left[r_{opt}^{\tau L}(S, k)\right]$$

will depend on τ as well as L, and let us assume that it can be completely specified for all $0 \leq \tau \leq T$ and $1 \leq L \leq l$ as Γ^*. The (S, k)-th entry of $R_{opt}^{\tau L}$, $r_{opt}^{\tau L}(S, k)$, will specify the state S_1 that the network will end up in if the network is initially in state S and there is a class k arrival, while the objective is to maximize the probability that the total number of events up to time τ is greater than or equal to L.

Lemma 5.1 *R_{opt}^{Tl} must satisfy the following optimality condition:*

For

$$S^* = r_{opt}^{\tau L}(S_0, k)$$

and any state resulting from any other policy

$$S = r(S_0, k)$$

for the same initial state S_0,

$$\begin{aligned} U^T_{l-1}(\Gamma^*, S^*) &= \int_\Omega p_{\omega|S^*}(\xi, \tau|S^*) U^{T-\tau}_{l-2}(\Gamma^*, r_{opt}^{(T-\tau)(l-1)}(S^*, \xi)) d\omega \\ &\geq U^T_{l-1}(\Gamma^*, S) \end{aligned}$$

where $\Omega = \{\omega : (\xi, \tau)\}$ is the joint event space of the first event and its occurrence time.

Lemma 5.1 follows immediately from dynamic programming.

Definition 5.3 A state S_1 is said to be *Tl- superior* to state S_2, or

$$S_1 \overset{Tl}{\geq} S_2$$

if and only if for some particular time T and integer l,

$$U_l^T(\Gamma, S_1) \geq U_l^T(\Gamma, S_2)$$

Therefore, Lemma 5.1 simply states that a Tl-optimal policy R^{Tl}_{opt} must always make assignments so that the state reached is $T(l-1)$-superior to all states that are reached by other policies.

Definition 5.4 A state S_1 is said to be *T-stationary l-superior* to state S_2, or

$$S_1 \overset{l}{\geq} S_2$$

if and only if for a particular integer l, S_1 is Tl-superior to S_2 for all T, and is said to be *Tl-stationary superior*, or

$$S_1 \overset{stat}{\geq} S_2$$

if and only if S_1 is Tl-superior to S_2 for all T and l.

Definition 5.5 R_{opt}^{Tl} is said to be T-stationary l-optimal if it is independent of T and is equal to $R^{(l)}$ for some particular l. If R_{opt}^{Tl} is independent of both T and l and is equal to R^*, it is said to be Tl-stationary optimal

Lemma 5.2 *If there is a Tl-stationary policy R^* that maximizes $U_l^T(R, S_0)$ for all l, T, and S_0, R^* is optimal in terms of minimizing overall blocking probability P_b.*

Lemma 5.2 is obvious as given any initial state, R^* also maximizes the expected total number of events up to any time T, which is

$$E\{N_e(T)\} = \sum_{l=1}^{\infty} U_l^T(R, S_0) \tag{5.2}$$

And the expected number of arrivals up to time T is,

$$E\{N_a(T)\} = \Lambda T \tag{5.3}$$

independent of initial state or scheduling policy.

Therefore, the expected number of departures, which is the expected number of events minus the expected number of arrivals, is

$$E\{N_d(T)\} = \sum_{l=1}^{\infty} U_l^T(R, S_0) - \Lambda T \tag{5.4}$$

will also be maximized.

According to Corollary 4.1, P_b is minimized.

5.1.2 Optimality of Non-invasive Hunting

Proof of Theorem 5.1:

We shall now proceed to prove Theorem 5.1. Assume that the inital state at time 0^- is $S_{0^-} = (00b_3)$ and there is a class 1 call arrival at time 0. With non-invasive hunting the arriving call will be assigned to channel 1 and thus

$$S_{0^+} = (10b_3)$$

The probability that there are one or more events up to time T will be,

$$\begin{aligned} & U_1^T(\Gamma, S_{0^+}) \\ & = 1 - Pr\{\text{No arrival from either sources and no departure up to time } T\} \\ & = 1 - e^{-(1+b_3+2\lambda)T} \end{aligned}$$

which is obviously maximized as long as the policy is non-wasting. Therefore, the state S_{0^+} reached by non-invasive hunting is T-stationary 1-superior to any other states reached by other policies. Therefore, non-invasive hunting must be T-stationary l-optimal for $l = 2$. It does not really matter whether we begin our induction from $l = 1$ or from $l = 2$. For $l = 1$, all policies must be the same as which policy we are under will not affect the occurrence of the first event, i.e.,

$$U_1^T(R_1, S) = U_1^T(R_2, S) \text{ for all } R_1,\ R_2$$

Next we shall show that if non-invasive hunting is T-stationary l-optimal for $l =$ L - 1, it must be T-stationary l-optimal for $l = L$. The assumption that non-invasive hunting is T-stationary l-optimal for $l = L$ - 1 ensures the following:

$$S' = (10b_3) \overset{L-2}{\geq} S = (01b_3) \qquad (5.5)$$

for $b_3 = 0$ and $b_3 = 1$.

Now assume there is a policy R that assigns to channel 2 when channel 1 is idle, i.e,

$$R : S_{0^-} = (00b_3) \xrightarrow{R} S_{0^+} = (01b_3)$$

Consider the alternative policy R' that assigns to channel 1 instead, i.e,

$$R' : S_{0^-} = (00b_3) \xrightarrow{R'} S_{0^+} = (10b_3)$$

Now let τ be the occurrence time of the first event that occurs next and let S_1 and S_1' be the state reached by R and R' respectively after the first next event.

If next event in R is D_2, or departure from channel 2, next event in R' will be D_1, and same state will be reached by R and R'. If next event is departure from channel 3, which would occur only when b_3 is 1, then the resulting states are $S_1 = (010)$ and $S_1' = (100)$. We know that S_1' is $(T-\tau)(L-2)$-superior to S_1, according to equation 5.5, since if the initial state is (000) and there is an arrival from source 1, assignment will be made to channel 1 due to the assumed T-stationary $(L-1)$-optimality of non-invasive hunting.

If next event is A_1, or arrival from source 1, R will make assignment to channel 1 and R' will make assignment to channel 2 and same state will be reached by both policies. If next event is A_2 and channel 3 is idle, $S_1 = (011)$ and $S_1' = (101)$, which is $(T-\tau)(L-2)$-superior to S_1 again if initial state is (001) and if there is an arrival from source 1, assignment would be made to channel 1 due to the assumed T-stationary $(L-1)$-optimality of non-invasive hunting. Finally, if next event is A_2

and channel 3 is busy, the arriving call will be blocked in R but not in R'. But we can construct R' such that the call is rejected as well. The resulting state in R', $S_1' = (101)$, is again superior to $S_1 = (011)$, the resulting state in R, according to equation 5.5. If deferred rejection is applied, R' can be even better.

Thus assuming that non-invasive hunting is T-stationary $(L-1)$-optimal, non-invasive hunting must also be T-stationary L-optimal as for any policy R that does not obey non-invasive hunting, an improved policy R' can always be found.

Thus we have completed the inductive proof for the Tl-stationary optimality of non-invasive hunting and therefore Theorem 5.1. QED

The above proof has been tedious and wordy but as it serves as the example to several subsequent proofs, we shall summarize the sample path proof as follows:

$\xi_1; R$	S_1	$\xi_1; R'$	S_1'	Relation
D_2	$(00b_3)$	D_1	$(00b_3)$	same
$D_3{}^\dagger$	(010)	D_3	(100)	$S_1' \overset{L-2}{\geq} S_1$
A_1	$(11b_3)$	A_1	$(11b_3)$	same
A_2	(011)	A_2	(101)	$S_1' \overset{L-2}{\geq} S_1$

Note: † For $b_3 = 1$ only

The interpretation of the sample path table is that the corresponding first next event, ξ_1, for R and R' in the same row of the table will occur with equal probability for all occurrence time.

5.2 Generalizations

5.2.1 To Simple-Sharing Loss Networks

It became immediately obvious that the proof of Theorem 5.1 in the previous section is independent of the arrival rates to the two different classes and can be easily generalized to any loss networks that have multiple classes, each having different arrival rate and multiple number of unshared channels and multiple number of shared channels that are shared by all classes.

Definition 5.6 A loss network or a hold network is called a *simple-sharing* network if and only if all shared channels in the network are shared by all the classes.

Theorem 5.2 *Non-invasive hunting is optimal for any simple-sharing loss network.*

Proof:

As in the previous section, it is obvious that any non-wasting policy will have the largest $U_2^T(\Gamma, S)$ for all T and S. Therefore non-invasive hunting is T-stationary l-optimal for $l = 1$ and $l = 2$. Now assuming that non-invasive hunting is T-stationary

l-optimal for $l = L-1$, we will show that non-invasive hunting must be l-optimal for $l = L$ by showing that improvement can always be made on a policy R that does not obey non-invasive hunting. Assume that there is a policy R that makes assignment to a shared channel m_s when it can be assigned to some unshared channel m_u. We can construct the alternative policy R' such that assignment is made to channel m_u instead. If next event is an arrival, R' will follow assignment that would have been made by R if possible or R' will reject the next arrival if it would have been blocked in R. For both cases and the case when the next event is a departure of other than the last assigned call, the resulting state under R', S_1', will otherwise be the same as S_1, the resulting state under R, except that m_u is busy in S_1' instead of channel m_s. Therefore in these cases S_1' is $(L-2)$-superior to S_1 due to the assumed T-stationary $(L-1)$-optimality of non-invasive hunting.

If the next event is the departure of the last assigned call, R' and R will end up in the same state. If the next event is an arrival and R makes assignment to channel m_u, R' will not be able to follow R but R' can always make assignment to channel m_s instead and both policies will again end up in the same state. This is what we call *deferred shared channel assignment.* Therefore improvement is always possible on R and non-invasive hunting must be L-optimal assuming that it is $(L-1)$-optimal. Thus our inductive proof is completed and non-invasive hunting also minimizes P_b according to Lemma 5.2.

Therefore the idea is that at a later time R' can always make assignment to the shared channel m_s previously left idle. Using up the shared channel m_s may cause some future arrivals to be blocked in R but not in R'. We have completed our proof assuming that R' will also reject these calls. Further improvement can be achieved if deferred rejection is applied. Therefore non-invasive hunting is strictly better than any other policies. **QED**

5.2.2 To General Access Structures

For a simple-sharing loss network, optimality of non-invasive hunting rule completely specifies the optimal policy (clearly it does not matter which channel within a set of unshared channels or shared channels should be used first). For a general access structure non-invasive hunting is not sufficient to specify the optimal policy as there are different types of shared channels. However, the following still holds:

Theorem 5.3 *For a loss network with some unshared channels, in the optimal policy assignments should never be made to any shared channel if they can be made to some unshared channels.*

Proof:

We can simply follow the proof of Theorem 5.2 exactly. **QED**

Therefore, what we have shown is the simple fact that in any loss network, assignments should always be made to unshared channel before they are made to shared channels.

5.3 W-hold Network

In this section we shall demonstrate that the following can be shown for the W-hold network:

- I - An optimal policy must obey non-invasive hunting
- II - There should be no holding of calls if the corresponding unshared channel is idle.
- III - Given that non-wasting policy is used, for channel 2, call selection from longer queue is better than from shorter queue.

In our model for the W-hold network, the arrival rates form the two classes are assumed to be equal. As we shall see from the proofs later on, I and II is valid even when the arrival rates are not equal.

Before we offer the proofs, we first present the following lemma:

5.3.1 The U-objective and Queueing Delay

Lemma 5.3 *If a channel scheduling policy R^* maximizes $U_l^T(R,S)$ for all T, l and initial state S, R^* will also minimize the average queueing delay in the network.*

As in Lemma 5.2, R^* maximizes the expected number of events and thus the expected number of departures up to time T for any T. Therefore R^* minimizes the expected number calls in system for all time T, which is the expected number of arrivals plus the number of calls in the initial state minus the expected number of departures. By Little's Theorem R^* minimizes the average system sojourn time and thus the average queueing delay.

5.3.2 Optimality of Non-invasive Hunting

Theorem 5.4 *An optimal channel scheduling policy for the W-hold network must obey non-invasive hunting.*

Proof:

Here we shall prove with greater generality that it is always better to first made assignment to an unshared channel than to a shared channel. The reason is that we have not established the optimality of non-wasting rule for hold networks and cannot eliminate states such as $S = (00b_3; n_1 n_2)$, $n > 0$, from consideration.

As in the proof of Theorem 5.1, non-invasive hunting must be T-stationary l-optimal for $l = 2$.

Let $(00b_3; n_1 n_2)$ be the initial state and assume that $n_1 \geq 1$. Let us compare R and R' such that

$$S_0\text{-} \xrightarrow{R} (01b_3; n_1\text{-}1\ n_2)$$

and

$$S_0\text{-} \xrightarrow{R'} (10b_3; n_1\text{-}1\ n_2)$$

For the more general initial condition S_{0^-} defined as above, we have to consider the possibility that there may be some immediate assignments made by R before the first event occurs. If the immediate assignment is to channel 3, R' can always follow R and there is no loss of generality for the subsequent sample path comparison. If the immediate assignment is from queue 1 to channel 1, then R' can assign to channel 2 and ends up in the same state. If there is no immediate assignments, sample path comparisons can be summarized in the following table:

$\xi_1; R$	S_1	$\xi_1; R'$	S'_1	Relation
D_2	$(00b_3; n_1n_2)$	D_1	$(00b_3; n_1n_2)$	same
D_3	$(010; n_1n_2)$	D_3	$(100; n_1n_2)$	$S'_1 \overset{l-2}{\geq} S_1$
A_1	$(01b_3; n_1n_2)$	A_1	$(10b_3; n_1n_2)$	same
A_2	$(011; n_1n'_2)$	A_2	$(101; n_1n_2\prime)$	$S'_1 \overset{l-2}{\geq} S_1$

Therefore, for any R that does not obey non-invasive hunting, an alternative policy R' that always ends up in a better or equivalent state can be constructed. Therefore, R cannot be optimal. **QED**

5.3.3 Optimality of Unshared Channel Non-wasting Rule

Definition 5.7 The *unshared channel non-wasting* rule means that no call should be held in queue if it can be assigned to an unshared channel immediately.

Theorem 5.5 *Optimal channel scheduling policy for the W-hold network must be unshared channel non-wasting.*

Proof:

In other words, we are to prove that the optimality of the unshared channel non-wasting rule for the W-hold network.

First, for $l = 2$, l-optimal policy must be unshared channel non-wasting, since

$$\begin{aligned} & Pr\{1 \text{ or more events up to time } T\} \\ & \quad = 1 - Pr\{\text{no arrival or departure up to time } T\} \\ & \quad = 1 - e^{-(\Lambda + nbusy(S_{0^+}))T} \end{aligned}$$

and the number of busy channel immediately after assignment, $nbusy(S_{0^+})$, is maximized only for a non-wasting policy.

Now assumed that an unshared channel non-wasting policy is T-stationary l-optimal for $l = L - 1$. Let the initial state be:

$$S_{0^-} = (0b_2b_3; n_1n_2)$$

$$n_1 > 0$$

Assume that there is a policy R that does not obey the unshared channel non-wasting rule, i.e.,

$$S_{0^-} \xrightarrow{R} S_{0^+} = (0b_2b_3; n_1n_2)$$

We will show that improvement is always possible by using some alternative policy R', which makes assignment from queue 1 to channel 1 immediately, such that,

$$S_{0^-} \xrightarrow{R'} S_{0^+} = (1b_2b_3; n_1 - 1\ n_2)$$

When we make the sample path comparison here there is one subtle difficulty. That is, the call which is assigned to channel 1 in R' is the same call that is left behind in queue in policy R. If we treat the two calls as having the same duration, sample path comparison will become tricky. The reason is that conditioned on what the first next event is, the duration of the call left behind in queue will have a different probability distribution. This complication can be easily resolved through a conceptual operation that we refer to as *call duration randomization.*

Definition 5.8 *Call duration randomization* refers to the conceptual operation that before we compare the sample paths of two initial states that result from two alternative policies, we randomize the durations of all the calls and give each call an independent and exponentially distributed duration with unit mean.

It is obvious that randomizing the call durations will not change the two initial states. The idea is simply to make our sample path comparison easier.

If channel 1 becomes idle in R' before the first event occurs in R, we shall devise a fictitious call to occupy channel 1 again. At the occurrence of the next event R' can always follow any assignments made by R if it does not involve a call from queue 1. In these cases, S_1', the resulting state in R', will have one more call assigned to channel 1 (which is occupied either by the original call or a fictitious call) as compared to S_1, the resulting state in R. S_1' must be T-stationary $(L-2)$-superior to S_1 due to the assumed T-stationary $(L-1)$-optimality of the unshared channel non-wasting rule. If the next assignment in R is from queue 1, it must be to channel 1 as we have proven the T-stationary $(L-1)$-optimality of non-invasive hunting in Section 5.3.2. As under R' channel 1 is either busy or occupied by a fictitious call, no assignment is made. Even so R' will end up in the same state as R and will have nothing to lose. Since we have the freedom of not occupying channel 1 with fictitious calls, there must exists another policy that is better or at least as good as R'.

Therefore, with the assumed T-stationary $(L-1)$-optimality of the unshared channel non-wasting rule, improvement can always be made on a policy that does not obey the unshared channel non-wasting rule for $l = L$. This completes our inductive proof. **QED**

5.3.4 Optimality of Select from Longer Queue Rule

Theorem 5.6 *For the W-hold network, given that a non-wasting policy must be used, call selection from the longer queue is better than from the shorter queue.*

Proof:

Again, for $l = 2$, all non-wasting policies are T-stationary l-optimality and the optimality of select from longer queue rule is obvious.

Now assume T-stationary l-optimality of select from longer queue rule for $l = L-1$ and let the initial state be

$$S_{0^-} = (101; n_1 n_2)$$

Therefore decision has to be made whether the next call to be assigned to channel 2 should be from queue 1 or from queue 2.

With no loss of generality, assume that

$$n_1 > n_2 > 0$$

Because for $n_2 = 0$, there can be no choice. And for $n_1 = n_2$, selection from either queue is equivalent from symmetry. We can also focus on initial states with all three channels busy due to the non-wasting assumption and the presence of calls in the two queues.

Assume that there is a policy R that selects from shorter queue. We compare it to a policy R' that selects from longer queue, i.e,

$$\begin{aligned} S_{0^-} &\xrightarrow{R} S_{0^+} = (111; n_1\, n_2\text{-}1) \\ S_{0^-} &\xrightarrow{R'} S_{0^+} = (111; n_1\text{-}1\, n_2) \end{aligned}$$

For the case that $n_2 - 1 \neq 0$, R' can always follow R if there is any assignment at the occurrence of the next event. Again, applying call duration randomization, sample path comparison can be summarized in the table below:

$\xi_1;R$	S_1	$\xi_1;R'$	S_1'	Relation
A_1	$(111;n_1+1\,n_2-1)$	A_1	$(111;n_1n_2)$	$S_1' \overset{L-2}{\geq} S_1$
A_2	$(111;n_1n_2)$	A_2	$(111;n_1-1\,n_2+1)$	$S_1' \overset{L-2}{\geq} S_1$, same if $n_1 = n_2+1$
D_1, D_2	$(111;n_1-1\,n_2-2)$	D_1, D_2	$(111;n_1-2\,n_2)$	$S_1' \overset{L-2}{\geq} S_1$, same if $n_1 = n_2+1$
D_3	$(111;n_1\,n_2-2)$	D_3	$(111;n_1-1\,n_2-1)$	$S_1' \overset{L-2}{\geq} S_1$

Thus R' is always better or at least as good.

For the case if $n_2 - 1 = 0$ and if the first event is departure from channel 3, then

$$\begin{aligned} S_1' &= (111;n_1\text{-}1\ 0) \\ S_1 &= (110;n_1\ 0) \end{aligned}$$

Therefore it is left to show that

$$S_1' = (111;n_1\text{-}1\ 0) \overset{L-2}{\geq} S_1 = (110;n_1\ 0)$$

Indeed we can show that S_1' is Tl-stationary superior to S_1 for any $n_1 - 1$. We can use the inductive proof again. For $l = 1$, obviously $S_1' \overset{l}{\geq} S_1$ is true, as there is one more assigned call for S_1' and thus it has a higher total departure rate.

Now assume $S_1' \overset{l}{\geq} S_1$ for $l = L - 1$, it must be true also for $l = L$ as at the next event, whatever assignment made for S_1, we can always make the same assignment for S_1', except when $n_1 - 1 = 0$ and next event is departure from channel 1. But if that happens the resulting state S_2 for S_1 will be $(110;00)$ and the resulting state S_2' for S_1' will be $(011;00)$, which is identical to S_2 due to symmetry. **QED**

5.4 Conjecture on the Optimality of the Shared Channel Non-wasting Rule for the W-hold Network

One conjecture that we have not been able to prove, however, is that the optimal scheduling policy for the W-hold network must also use the shared channel non-wastingly. In Chapter 3 we have shown that the optimality of non-wasting rule is not generally true for hold networks. However, due to the symmetry for the W-hold network, there is a strong suspicion that it should be true. Using the same type of inductive proof that we have developed in this chapter, we can see that a shared channel non-wasting policy must be T-stationary l-optimal for $l = 1$ and 2 for all initial states as assignment is always made if possible and thus overall departure rate is maximized. Here we shall illustrate why similar inductive proof cannot be carried through.

Let us consider one particular initial state:

$$S_{0^-} = (101; 11)$$

For S_{0^-} the decision to be made is which of the two waiting calls should be assigned to channel 2, or if the network should hold both calls in queue and make no assignment at all. Let R be a policy that hold the calls in this case and let R' be a non-wasting policy that select a call to be assigned to channel 2 immediately. For this case it should not matter which call to select due to the symmetry. Therefore let us assumed that call from queue 1 is selected, i.e,

$$S_{0^-} = (101; 11) \xrightarrow{R} S_{0^+} = (101; 11)$$

$$S_{0^-} = (101;11) \xrightarrow{R'} S_{0^+} = (111;01)$$

In this case we will apply call duration randomization on all calls and ignore first event in R' if it is departure from channel 1. But if departure from channel 1 in R' preceeds the first event in R, we shall mark the corresponding sample path with an asterisk sign $*$.

Assume that unshared channel non-wasting rule is T-stationary $(L-1)$-optimal. We can summarize sample path comparision in the following table:

$\xi;R$	S_1	$\xi;R'$	S'_1	Relation	$Pr\{\cdot\}$
D_1	$(111;00)$	D_2^*	$(011;00)$	$S'_1 \overset{L-2}{\geq} S_1$	$\frac{1}{(3+2\lambda)(2+2\lambda)}$
		D_2	$(011;00)$	same	$\frac{1}{3+2\lambda}$
D_3	$(111;00)$	D_3^*	$(011;00)$	$S'_1 \overset{L-2}{\geq} S_1$	$\frac{1}{(3+2\lambda)(2+2\lambda)}$
		D_3	$(111;00)$	same	$\frac{1}{3+2\lambda}$
A_1	$(111;11)$	A_1^*	$(111;01)$	$S'_1 \overset{L-2}{\geq} S_1$	$\frac{\lambda}{(3+2\lambda)(2+2\lambda)}$
		A_1	$(111;11)$	same	$\frac{\lambda}{3+2\lambda}$
A_2	$(111;11)$	A_2^*	$(011;02)$	unknown	$\frac{\lambda}{(3+2\lambda)(2+2\lambda)}$
		A_2	$(111;02)$	$S'_1 \overset{L-2}{\leq} S_1$	$\frac{\lambda}{3+2\lambda}$

Therefore, for the last sample path at least, R will end up in a better state than R' and thus we cannot guarantee the optimality of R'. The problem is in the delayed commitment made by R that allows it to be better under some situations. The sample path comparision fails even if we look further forwards into subsequent events.

One conclusion we can obtain here is that since the difference between $U_l^T(\Gamma, S_1)$ and $U_l^T(\Gamma, S_1')$ is bounded by 1 and thus finite, a non-wasting policy must at least be almost as good as any policy that holds unblocked calls when λ approaches 0.

Another point worth to be mentioned is that attempts to show the optimality of non-wasting policy for hold network by showing the Tl-stationary optimality of such policy for all initial states is bounded to fail for many other types of hold networks. Consider the simple shared network as shown in Figure 5.1, where there are two classes of equal arrival rates with one shared channel and n unshared channels for each class.

For this example only let us change the objective to be maximizing the expected total number of assignments made up to time T. In other words we are using the V-objective instead for this example. From Lemma 5.3 it is also equivalent to minimizing the expected number of calls left in queue. Assume that the initial state is that all channels are busy except the shared channel, and that there is one call in each of the two queues. Consider the limit as λ, the arrival rate to each class, approaches zero. A non-wasting policy R' maximizes the probability that there is at least one assignment for any T, as the probability will uniformly be one. After the first assignment is made, as probability of having new arrival is small, the probability of having two or more assignments up to time T will roughly be

$$1 - e^{-(n+1)T}$$

for as soon as one of the $n + 1$ channels that the remaining call can access to

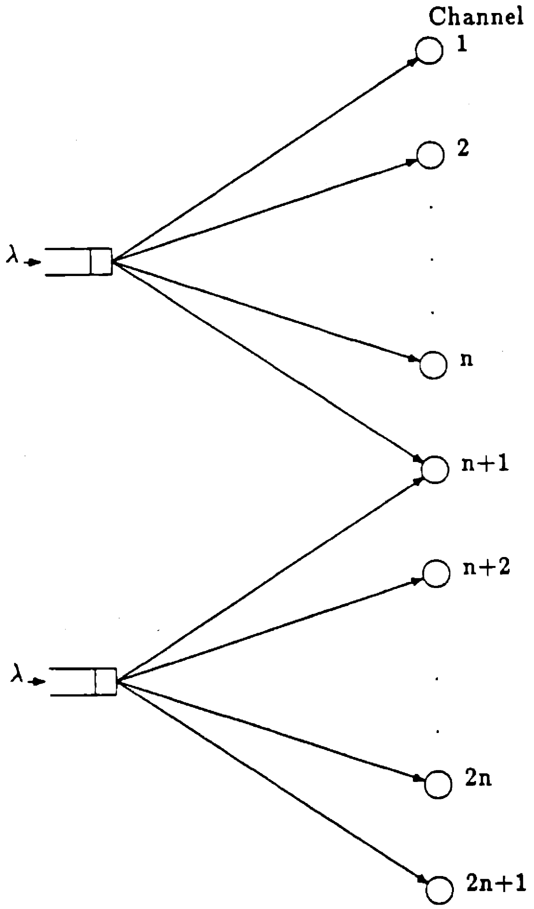


Figure 5.1: A Simple-sharing Hold Network

becomes idle, the second assignment can be made.

But now consider a policy R that holds both calls until one of the $2n$ busy channels becomes idle, and then assigns one of the calls to the channel that has become free, and the other call to the shared channel. For R the probability of making at least one assignment up to time T will be the same as the probability of making at least two assignments and is

$$1 - e^{-2nT}$$

Therefore, while R' maximizes the probability of having at least one assignment up to time T, R maximizes the probability of having at least two assignments up to time T. The difference is due to the delayed commitment made by R.

5.5 Extension to General Hold Networks

We can see that Theorem 5.4 and 5.5 can be extended to any hold network. That is, the optimal scheduling policy for any hold network must obey the non-invasive hunting rule and the unshared channel non-wasting rule. We proof the optimality of these two scheduling rules first for the W-hold network due to the fact that attempts to give rigorous proofs for the general case immediately will be too confusing.

5.5.1 Optimality of Non-invasive hunting

Theorem 5.7 *An optimal channel scheduling policy for any hold network must obey non-invasive hunting.*

Proof:

As in the Proof of Theorem 5.4, non-invasive hunting must be T-stationary l-optimal for $l = 2$.

Assume that non-invasive hunting is obeyed in any T-stationary $(L-1)$-optimal policy. Assume that there is a policy R that makes assignment to a shared channel m_s while assignment can be made to an unshared channel m_u. We can then construct an alternative policy R' such that assignment is made to channel m_u instead. As in the proof of Theorem 5.4, if there is any immediate assignment made by R before the first event occurs, R' can always follow R if the immediate assignment is not made to m_u, and there will be no loss of generality for the initial condition. If the immediate assignment is made to m_u, R' can always made assignment to m_s instead and with call duration randomization, R' and R will end up in the same state. This is what we called deferred shared channel assignment in the Proof of Theorem 5.2. So we have taken care of the possibility that there may be immediate assignments made by R.

Now for the first event in R, if it is departure from channel m_s, it will be departure from channel m_u in R', and the two policies result in the same state. If a call is blocked in R but not in R', we will hold the call in R' also. Or if the next assignment made by R is one that can be followed by R', for either case R' must result in a $(L-2)$-superior state due to the assumed $(L-1)$-optimality of non-invasive hunting and as the only difference between the resulting state in R' and the resulting state in R is that in the former, the unshared channel m_u is busy and the shared channel m_s is idle, while the converse is true in the latter. If the next assignment in R is one that cannot be followed by R', the assignment made by R must be to channel m_u, in which case R' can always make a deferred shared channel assignment and make assignment to channel m_s. In this case both policies end up in the same state.

Therefore, our inductive proof is completed.

QED

5.5.2 Optimality of Unshared Channel Non-wasting Rule

Theorem 5.8 *An optimal scheduling rule for any hold network must obey the unshared channel non-wasting rule.*

Proof:

As in the proof of Theorem 5.5, for $l = 2$, an l-optimal policy must be non-wasting and thus unshared channel non-wasting.

Assume that a $(L-1)$-optimal policy is unshared channel non-wasting and assume that there is a policy R that does not obey the unshared channel non-wasting rule such that a call c that can be assigned to an unshared channel m_u immediately is held in a queue, say queue k. We will construct the alternative policy R' which assigns c immediately to channel m_u. If channel m_u becomes idle before the occurrence of the first next event in R, we shall devise a fictitious call to occupy channel 1 again. If the next event is a departure, R' will result in a superior state due to the assumed

$(L-1)$-optimality of the unshared channel non-wasting rule, as the resulting state in R' differs from the resulting state in R only in that one more call is assigned to an unshared channel. If the next event is an arrival and if either no assignment is made in R or the assignment made in R does not involve a call from queue k, it can always be followed by R' and R' will result in a superior state. If there is an assignment made in R that cannot be followed by R', it must be an assignment made from queue k. If the assignment is made to channel m_u or another unshared channel, the two policies will result in the same state. If the assignment is made to a shared channel, R' will result in a superior state.

Therefore, R' is always better or at least as good and our inductive proof is completed. QED

In summary, in this chapter we have proven that non-invasive hunting must be obeyed in all optimal scheduling policies for all loss and hold networks. We have also proven that the unshared channel non-wasting rule must also be obeyed in all optimal scheduling policies for all hold networks. For the W-hold network, we have proven that selection from the longer queue is better than selection from the shorter queue. However, the optimality of using the shared channel non-wastingly for the W-hold network remains as a conjecture.

Chapter 6

Ideal Grading and Random Scheduling

In this chapter and the next we shall focus on what we call *uniform-accessibility* networks.

Definition 6.1 A *uniform-accessibility* network is one in which all classes have the same accessibility. In other words, the number of channels in all accessible channel groups is the same.

The accessibility parameter N_b can be viewed as the bandwidth of individual users and thus a measure of the cost of network interfaces. The fundamental question discussed in this chapter and Chapter 7 is as follows: Given that the total offered traffic to an M-channel uniform-accessibility network is Λ and that the accessibility of all classes is restricted to N_b, what is the minimum blocking probability that can be obtained? Furthermore, how should the access structure of the network be designed to achieve this minimum blocking probability?

6.1 Ideal Grading

In section 2.1 we have explained what we mean by "access structure". It is stated here again as a definition.

Definition 6.2 For a network with M channels and total offered traffic Λ, *access structure* refers to the network topology which can be fully described by the following parameters:

1. The total number of classes, K,
2. The accessible channel group AG_k for each class k, and
3. The fraction of total traffic assigned to each class, ρ_k.

For a uniform-accessibility network, the maximum number of classes is $C(M, N_b)$, as it is the maximum number of distinct subsets of N_b channels out of M channels. Therefore the optimal access structure problem can simply be viewed as a *traffic splitting* problem, for which we want to decide what fraction of the total traffic should be assigned to each class so that minimum overall blocking probability can be achieved. That means we want to find the optimal ρ_k for $k = 1, 2, \cdots, C(M, N_b)$. ρ_k will be zero if the particular class do not exist in the access structure.

The simplest access structure is the case where K, the number of classes, is equal to $\frac{M}{N_b}$, so that each channel belongs to one accessible group only. As there is no shared channel at all the network will decompose into K disjoint networks such that each can be modeled as an $M/M/N_b/N_b$ chain, where $M/M/N_b/N_b$ is the standard notation of a Markovian queue with N_b servers and no waiting room (blocked call lost). With general access structures the number of classes will be greater than K and there will be shared channels and each shared channel may be shared by a different number of classes. This is what we refer to as *grading*[1], a word from very early works in congestion theory in telephony. For a network with general access structure and shared channels, we say that grading is use. Networks with no shared channel will be referred to as *no-grading* networks.

Definition 6.3 For a uniform-accessibility network, the *ideal grading* [2], as named by Erlang, is when the total traffic is evenly distributed among the maximum possible number of classes, which is $C(M, N_b)$. In other words, $K = C(M, N_b)$ and $\rho_k = \frac{1}{K}$ for $k = 1, 2, \cdots, K$. A loss network with ideal grading is called an *ideal-loss-network.*

For $N_b = 2$, the ideal gradings for $M = 3$ and for $M = 4$ are shown in Figure 6.1 and Figure 6.2 respectively.

6.1.1 Random Scheduling

Definition 6.4 *Random scheduling* refers to the non-deterministic scheduling rule such that when the network is in state i with I_i being the set of idle channels, upon

[1] see [32] Chapter 7

[2] Again, see [32] Chapter 7

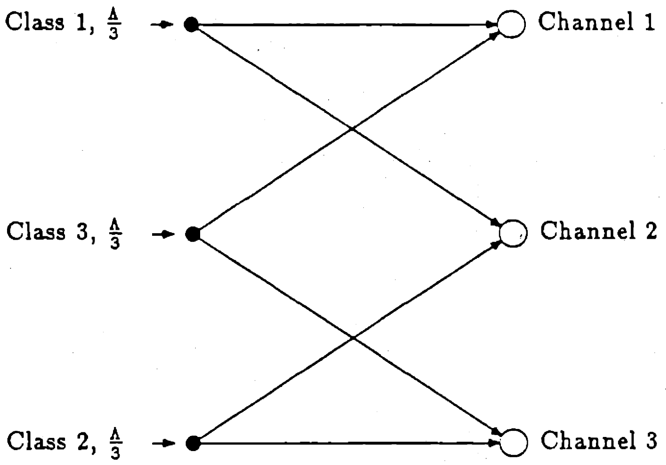


Figure 6.1: Ideal-Loss-Network for $M = 3$ and $N_b = 2$

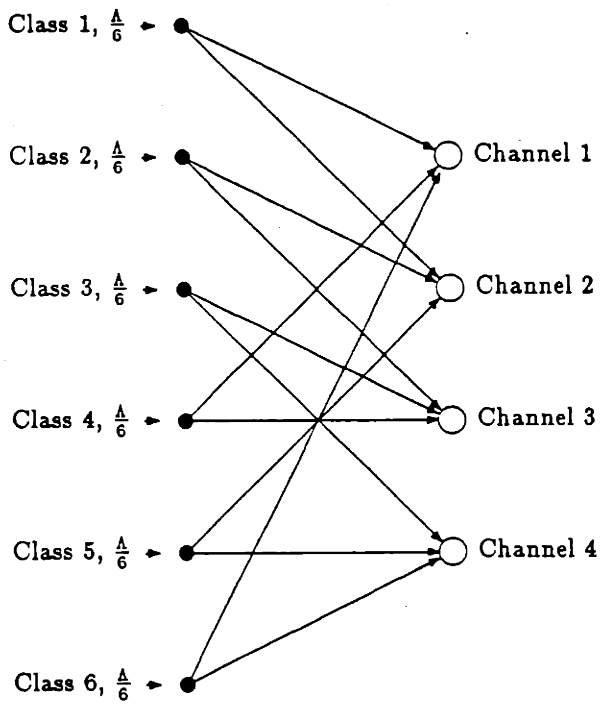


Figure 6.2: Ideal-Loss-Network for $M = 4$ and $N_b = 2$

a class k arrival, assignment will be made to any one of the channels in $AG_k \cap I_i$ with equal probability.

In this section we shall first demonstrate the complete symmetry in an ideal-loss-network and then show the optimality of random scheduling in such a network. Although we shall show later that ideal-grading is not in general the optimal access structure, the fact that closed form solution of its blocking probability can be obtained allows us to have some ideas about how much blocking probability can be reduced by grading.

Lemma 6.1 *Both the continuous time Markov chain and the imbedded chain (as developed in Chapter 4) describing the ideal-loss-network with random scheduling are reversible*[3].

<u>Proof</u>

Let the matrix $Q = [q_{ij}]$ be the transition rate matrix describing the ideal-loss-network with random scheduling. Let M be the total number of channels. The total number of states in the network will be $N = 2^M$. The total number of classes is $C(M, N_b)$ and let λ be the arrival rate of each class, such that

$$\lambda C(M, N_b) = \Lambda$$

The total number of states with n busy channels will be equal to $C(M, n)$. Now let us examine the transition rate between any two states i and j. Let us define the

[3] For definition of reversible Markov chain, see [15]

occupancy of a state as follows:

Definition 6.5 For a loss network, the occupancy of a state i is the number of busy channels in state i. Alternatively, we say that state i is an n-*occupancy* state if and only if $nbusy(i) = n$.

As state transition is due to either one single departure or one single arrival, q_{ij} is non-zero only if

$$| \, nbusy(i) - nbusy(j) \, | = 1$$

In other words, for the continuous time chain transitions occur only between states with occupancies that differ by 1. For the imbedded chain, transitions can be from a state back to itself when blockings occur.

We have:

1. For $j = D(i, m)$ for some busy channel m,
$$q_{ij} = 1. \tag{6.1}$$

2. For j such that $i = D(j, m)$,
$$\begin{aligned} q_{ij} &= \lambda\{\sum_{n=1}^{N_b} \frac{1}{n} C(M - nbusy(i) - 1, n-1) C(nbusy(i), N_b - n)\} \\ &= \lambda\{\sum_{n=1}^{N_b} \frac{1}{n} \frac{(M - nbusy(i) - 1)!}{(n-1)!(M - nbusy(i) - n)!)} C(nbusy(i), N_b - n)\} \\ &= \frac{\lambda}{M - nbusy(i)} \sum_{n=1}^{N_b} C(M - nbusy(i), n) C(nbusy(i), N_b - n) \\ &= \frac{\lambda}{M - nbusy(i)} [C(M, N_b) - C(nbusy(i), N_b)] \qquad (6.2) \end{aligned}$$

$$= \frac{\Lambda}{M - nbusy(i)}[1 - \frac{C(nbusy(i), N_b)}{C(M, N_b)}] \tag{6.3}$$

3. and $q_{ij} = 0$ otherwise.

We can also see that whenever $q_{ji} = 1$, $q_{ij} = \frac{\Lambda}{M-nbusy(i)}[1 - \frac{C(nbusy(i),N_b)}{C(M,N_b)}]$. The interpretation of equation 6.3 is that $\Lambda[1 - \frac{C(nbusy(i),N_b)}{C(M,N_b)}]$ is the amount of traffic that is not blocked in state i, and that all $M - nbusy(i)$ idle channels are equally likely to become busy next. Also, for each transition that corresponds to an assignment, there is a transition in the opposite direction with rate equal to 1 that corresponds to a departure.

The Markov chain for the ideal-loss-network with $M = 3$ and $N_b = 2$ is shown in Figure 6.3. What we can see is that for each n-occupancy state i, there are $M - n$ non-zero and equal transition rates between state i and $M - n$ states with $n + 1$ occupancy, and there are n unity transition rates between state i and n states with $n - 1$ occupancy.

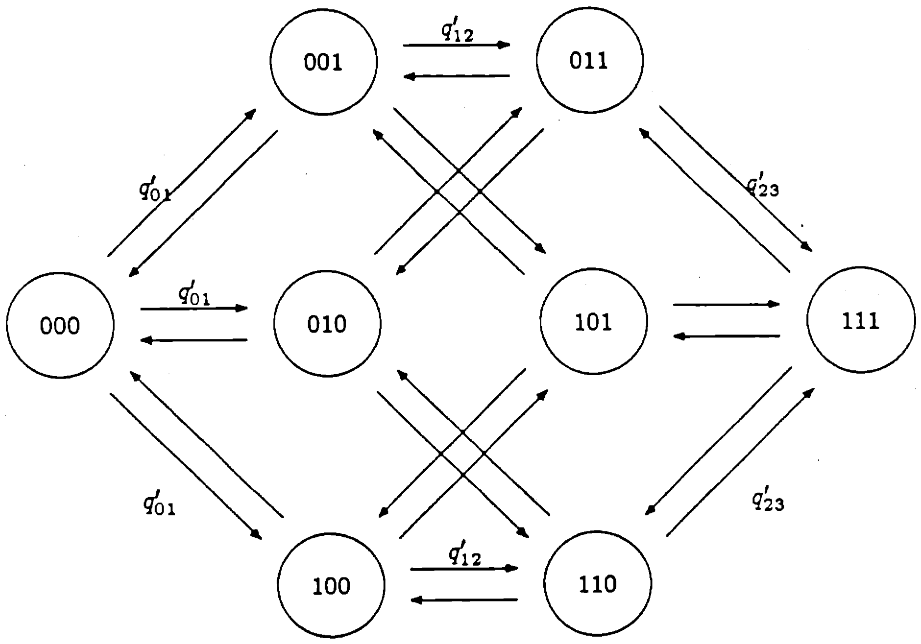
For the imbedded chain described in Chapter 4, the transition probabilities are given by:

$$p_{ij} = \frac{q_{ij}}{nbusy(i) + \Lambda} \tag{6.4}$$

and

$$p_{ii} = 1 - \sum_{j} p_{ij} \tag{6.5}$$

What we have arrived at is that for the ideal grading with random scheduling, in the continuous time chain the transition rates are the same as long as they corre-



$q'_{n(n+1)} = q_{ij}$ s.t. $nbusy(i) = n$ and $nbusy(j) = n+1$

All transition rates corresponding to departures equal 1

Figure 6.3: Continuous Time Markov Chain for an Ideal-Loss-Network with $M = 3$ and $N_b = 2$

spond to assignments made in states with the same occupancy or to departures from states with the same occupancy. The same is true for the transition probabilities of the imbedded chain.

It can be easily seen that Kolmogorov condition[4] is satisfied for any cycles in the continuous time Markov chain or the imbedded chain. For any cycle in either Markov chains, if there is a transition from an n-occupancy state i to an $(n+1)$-occupancy state j, there must be an associated transition from an $(n+1)$-occupancy state j' to an n-occupancy state i'. This is due to the fact that as a cycle must begins and ends in the same state, it must go through as many transitions that correspond to assignments from some n-occupancy states as there are that correspond to departures from $(n+1)$-occupancy states. For the reverse cycle, there are corresponding transitions from state j to state i and from state i' to state j'. And we have:

$$q_{ji} = q_{j'i'}$$

$$p_{ji} = p_{j'i'}$$

and

$$q_{i'j'} = q_{ij}$$

$$p_{i'j'} = p_{ij}$$

Therefore the product of transition rates in any cycle must be equal to the product of transition rates in the reverse cycle. Kolmogorov condition is satisfied and both the continuous time chain and the imbedded chain must hence be reversible. QED

[4] see [15]

The reversibility property is useful later on when we derive the upper and lower bounds of the blocking probability for ideal-loss-networks with random scheduling.

Lemma 6.2 *For the ideal-loss-network with random scheduling, in both the continuous time chain and the imbedded chain, steady state probabilities for states with the same occupancy are the same.*

Proof:

This is obvious from the reversibility of the Markov chains. Let θ_i be the steady state probability of state i for the continuous time chain (while π_i is the steady state probability of state i for the imbedded chain). There is only one state with 0-occupancy, and let this state be state $i = 1$, with steady state probability θ_1. Due to detailed balance for a reversible chain, for any 1-occupancy state i,

$$\theta_i = \theta_1 q_{1i}$$

q_{1i}'s are the same for all i's that have occupancy equal to 1. Therefore θ_i's must be the same for all 1-occupancy states.

If θ_i's are the same for all n-occupancy states, they must be the same for all $(n+1)$-occupancy states. The reason is that for any $(n+1)$-occupancy state i, it must be related to some n-occupancy state j by

$$\theta_i = \theta_j q_{ji}$$

and all θ_j's are equal and q_{ji}'s are equal. By induction we know that Lemma 6.2 must be true for the continuous time chain. Making the same argument for the imbedded chain, we know that Lemma 6.2 must also be true for the imbedded chain. **QED**

Furthermore, θ_i for an n-occupancy state is given by:

$$\theta_i = B^{-1} \prod_{l=0}^{n-1} \frac{\Lambda}{M-l}(1 - \frac{C(l,N_b)}{C(M,N_b)}) \tag{6.6}$$

where B is a normalization constant and is given by

$$B = \sum_{n=0}^{M} C(M,n) \prod_{l=0}^{n-1} \frac{\Lambda}{M-1}(1 - \frac{C(l,N_b)}{C(M,N_b)}) \tag{6.7}$$

Lemma 6.3 *Random scheduling is an optimal policy for the ideal-loss-network.*

<u>**Proof:**</u>

Originally, for the value determination step, there are $N = 2^M$ value determination equations:

$$v_i - \sum_{j=1}^{N-1} p_{ij}^r v_j + g = z_i^r \quad \text{for } i = 1,2,3,\ldots,N \quad v_N = 0$$

If we can show that the values v_i's as determined by the value determination step of the policy iteration algorithm are the same for all states with the same occupancy, i.e.,

$$v_i = v_j = u_n$$

for all i, j such that

$$nbusy(i) = nbusy(j) = n$$

for random scheduling, than random scheduling must be optimal as no further improvement can be made.

Now let us assume that the above is true. Then each value determination equation for an n-occupancy state will be reduced to the following:

$$u_n - p'_{n(n+1)}u_{n+1} - p'_{n(n-1)}u_{n-1} + g = z'_n \qquad \forall i \text{ s.t. } nbusy(i) = n, \qquad (6.8)$$
$$n = 0, 1, \ldots, M \;\; u_M = 0$$

where now,

$$p'_{n(n+1)} = \frac{\Lambda(1 - \frac{C(n,N_b)}{C(M,N_b)})}{n + \Lambda} \qquad (6.9)$$
$$p'_{n(n-1)} = \frac{n}{n + \Lambda} \qquad (6.10)$$
$$p'_{nn} = \frac{\Lambda\frac{C(n,N_b)}{C(M,N_b)}}{n + \Lambda} \qquad (6.11)$$
$$z'_n = p'_{n(n-1)} \qquad (6.12)$$

What has happened here is that the 2^M equations we originally have for the value determination step are reduced to $M + 1$ equations. Now there must be a unique solution (assuming $u_M = v_N = 0$) for u_0, u_1, u_2, $\cdots$, u_{M-1} and g as

$$P' = [p'_{n_1 n_2}]$$

is obviously a $(M+1) \times (M+1)$ stochastic matrix. The solutions obtained will also satisfy the 2^M equations we originally have. Therefore we can indeed find solutions for v_i's such that $v_i = v_j$ for all i and j such that $nbusy(i) = nbusy(j)$, and the solution found must be the unique solution. No further improvement can be made and random scheduling must be an optimal policy for an ideal-loss-network. QED

6.1.2 Blocking Probability in an Ideal-Loss-Network

Let ϕ_n be the probability that an ideal-loss-network is in an n-occupancy state. We have,

$$\begin{aligned} \phi_n &= \sum_{i:nbusy(i)=n} \theta_i && (6.13) \\ &= C(M,n) \prod_{l=0}^{n-1} \frac{\Lambda}{M-l}(1 - \frac{C(l,N_b)}{C(M,N_b)}) && (6.14) \end{aligned}$$

Therefore,

$$\phi_n = \phi_{n-1} \frac{\Lambda}{n}(1 - \frac{C(n-1,N_b)}{C(M,N_b)}) \qquad (6.15)$$

Since $\sum_n \phi_n = 1$, we also have,

$$\phi_n = \frac{\prod_{l=1}^{n} \frac{\Lambda}{n}(1 - \frac{C(n-1,N_b)}{C(M,N_b)})}{\sum_{n=0}^{M} \prod_{l=1}^{n} \frac{\Lambda}{n}(1 - \frac{C(n-1,N_b)}{C(M,N_b)})} \qquad (6.16)$$

The blocking probability for an ideal-loss-network with random scheduling is first given by Erlang[5] as:

$$P_b = \sum_{n=N_b}^{M} \phi_n \frac{C(n, N_b)}{C(M, N_b)} \tag{6.17}$$

This result is clear intuitively. Due to the complete symmetry in an ideal-loss-network, the process describing the number of busy channels in the network is also a Markov process and can be represented by an occupancy chain. The steady state probability of finding n busy channels is given by ϕ_n. With n busy channels, a fraction of $\frac{C(n,N_b)}{C(M,N_b)}$ of traffic arrivals will be blocked.

6.1.3 Channel-Reduced-Ideal-Network

Definition 6.6 For an ideal-loss-network with AG_k's, $k = 1, 2, \ldots, C(M, N_b)$, being the accessible groups, a *channel-reduced-ideal-network* is a network such that one or more channels are removed from all accessible groups that contain them, while the arrival rates to all classes remain the same.

A channel-reduced-ideal-network will no longer be a uniform-accessibility network as the accessible groups containing some of these removed channels will be reduced in size while those that do not contain these channels will not be affected.

We have the following theorem for channel-reduced-ideal-network:

[5] see [32] Chapter 7

Theorem 6.1 *Random scheduling is an optimal channel selection policy for any channel-reduced-ideal-network.*

Proof:

We can view the Markov chain corresponding to a channel-reduced-ideal-network as a subspace or a truncated version of the Markov chain corresponding to the original ideal-loss-network. This truncated chain will contain all the states with the removed channels always being busy and transitions into and out of the truncated chain are ignored. Let r be the number of channels removed. An n-occupancy state in the channel-reduced-ideal-network will correspond to an $(n+r)$-occupancy state in the untruncated Markov chain. If we examine any n-occupancy state i there will be $M-r-n$ non-zero and equal transition rates between i and $M-r-n$ $(n+1)$-occupancy states, and n unit transition rates between i and n $(n-1)$-occupancy states. Following the proof of Lemma 6.3, we can show that with random scheduling the values as determined by the value determination step must be the same for all states that have the same occupancy. Therefore no improvement can be made and random scheduling must be optimal for a channel-reduced-ideal-network. QED

6.1.4 Bounds on Blocking Probability in Ideal-Loss-Network

Theorem 6.2 *If P_b is the blocking probability in an M-channel ideal-loss-network* *with total traffic Λ and accessibility group size N_b, P_b must satisfies:*

$$P_b \geq (\frac{\Lambda}{M}(1-P_b))^{N_b} \tag{6.18}$$

$$\textit{and} \quad P_b \leq \frac{(\Lambda(1-P_b))^{N_b}}{M(M-1)\cdots(M-N_b+1)} \tag{6.19}$$

Proof:

From previous developments in this chapter, it is apparent that for the ideal-loss-network with random scheduling, the overall blocking probability is the same as the probability that a random arrival from any particular class is blocked. Let m_1, m_2, $\cdots$, m_{N_b} be the channels in the accessible channel group AG_k, i.e,

$$AG_k = (m_1, m_2, \cdots, m_{N_b})$$

Then the probability that a class k call is blocked will be equal to the probability that all channels in AG_k are busy, i.e.,

$$\begin{aligned} P_b &= Pr\{m_1\text{busy}\}Pr\{m_2\text{busy} \mid m_1\text{busy}\} \\ &\quad \cdots Pr\{m_{N_b}\text{busy} \mid m_1 m_2 \cdots m_{N_b-1}\text{busy}\} \end{aligned} \tag{6.20}$$

Firstly, we know

$$Pr\{m_1\text{busy}\} = \frac{\Lambda}{M}(1-P_b) \tag{6.21}$$

as it is the throughput per channel as the probability that any channel is busy must be the same due to symmetry. Secondly, given m_1 is busy, the probability that any other channel is busy is also the same due to symmetry, i.e.,

$$Pr\{m_2\text{busy} \mid m_1\text{busy}\} = Pr\{m_i\text{busy} \mid m_1\text{busy}\} \ \forall i \neq 1 \tag{6.22}$$

Since the Markov chain is reversible, we can solve for the probability that m_2 is busy given m_1 is busy from the truncated chain, by which we means the Markov chain with all the states such that m_1 is idle truncated. The truncated chain is identical to the Markov chain representing a channel-reduced-ideal-network, resulting from the removal of channel 1. Since random scheduling is optimal for a channel-reduced-ideal-network, the probability that any other channel is busy given m_1 busy is maximized as blocking probability is minimized and throughput is maximized. Now let us imagine that there is a sub-optimal scheduling rule for the channel-reduced-ideal-network which acts as if channel m_1 exists, assigns calls to m_1 while in fact these calls are rejected, and keeps track of whether m_1 should be supposely busy or idle by assuming that the non-existing channel m_1 will be held up for an exponentially distributed duration for each call assigned. For the sub-optimal policy thus the probability that any other channel is busy will be $\frac{A}{M}(1 - P_b)$. For the original chain which is optimal, the probability that any other channel is busy given m_1 is busy must therefore be larger. Now imagine that we are given m_1, m_2 busy, by considering the sub-optimal scheduling policy that acts as if they are available for assignment we can easily see that the probability any other channel is busy given two channels are busy must also be larger than $\frac{A}{M}(1 - P_b)$. Thus we have proven the lower bound.

For the upper bound, imagine that with one channel removed, the total throughput cannot be larger than the original Markov chain. Thus the throughput per channel for the truncated chain, which is the same as the probability that any other channel is busy given m_1 is busy, must be smaller than $\frac{A}{M-1}(1 - P_b)$. With similar consideration when two or more channels are given busy, the upper bound is established. **QED**

The nice feature for these two bounds is that when M, the total number of channels becomes very large, the bound will become very tight and we can treat the solution to

$$P_b = (\frac{\Lambda}{M}(1 - P_b))^{N_b} \tag{6.23}$$

as the blocking probability of the corresponding ideal-loss-network. The lower bound that we have established will be referred to as the *ideal lower bound.* The idea behind the ideal lower bound is that channel usages are positively correlated. That is, given that a channel is busy in the network, the probability that any other channel is busy is generally increased. Our numerical results have shown that this ideal lower bound is not in general observed for more complicated access structures. This implies that in some access structures with optimal scheduling, the usages of some channels must be negatively correlated.

6.2 Comparison To $M/M/N_b/N_b$ No-Grading Network

In this section the subject is to compare the performance of the ideal-loss-network with a network with no grading but has the same accessible group size and the same offered traffic per channel, α. It is well known that for a loss network with N_b channels, the blocking probability is given by the Erlang B-Formula:

$$P_B = \frac{(N_b\alpha)^{N_b}}{N_b!}(\sum_{n=0}^{N_b}\frac{(N_b\alpha)^n}{n!})^{-1} \tag{6.24}$$

We shall denote as P_I the ideal lower bound blocking probability that satisfies equation 6.19 with equality. The blocking probability for the $M/M/N_b/N_b$ system, P_B, is plotted along with P_I for various N_b's against different value of α, the offered traffic per channel. The results are shown in Figure 6.4. For the same N_b the lower curve always corresponds to P_I. Indeed we can see that the reduction in blocking probability can be quite substantial, particularly when N_b is large.

We can also consider the asymptotic behaviour when N_b is large. For the $M/M/N_b/N_b$ case, with $\alpha < 1$,

$$\begin{aligned} P_B &= \lim_{n\to\infty} \frac{(N_b\alpha)^{N_b}}{N_b!}(\sum_{n=0}^{N_b} \frac{(N_b\alpha)^n}{n!})^{-1} && (6.25) \\ &\approx \frac{1}{N_b!}(\frac{N_b\alpha}{e^\alpha})^{N_b} && (6.26) \\ &\approx \frac{(\alpha e^{1-\alpha})^{N_b}}{\sqrt{2\pi N_b}} \quad \text{using Stirling's approximation} && (6.27) \end{aligned}$$

Therefore, when N_b is large, blocking probability P_B will approximately be multiplied by a factor of $\alpha e^{1-\alpha}$ with each unit increase in N_b for the $M/M/N_b/N_b$ network. For the ideal-loss-network, the multiplication factor would approximately be α when P_b is small. Since $e^{1-\alpha} > 1$ for $\alpha < 1$, blocking probability is reduced faster in the case of the ideal-loss-network. For $\alpha > 1$, the approximation for P_B will not be valid.

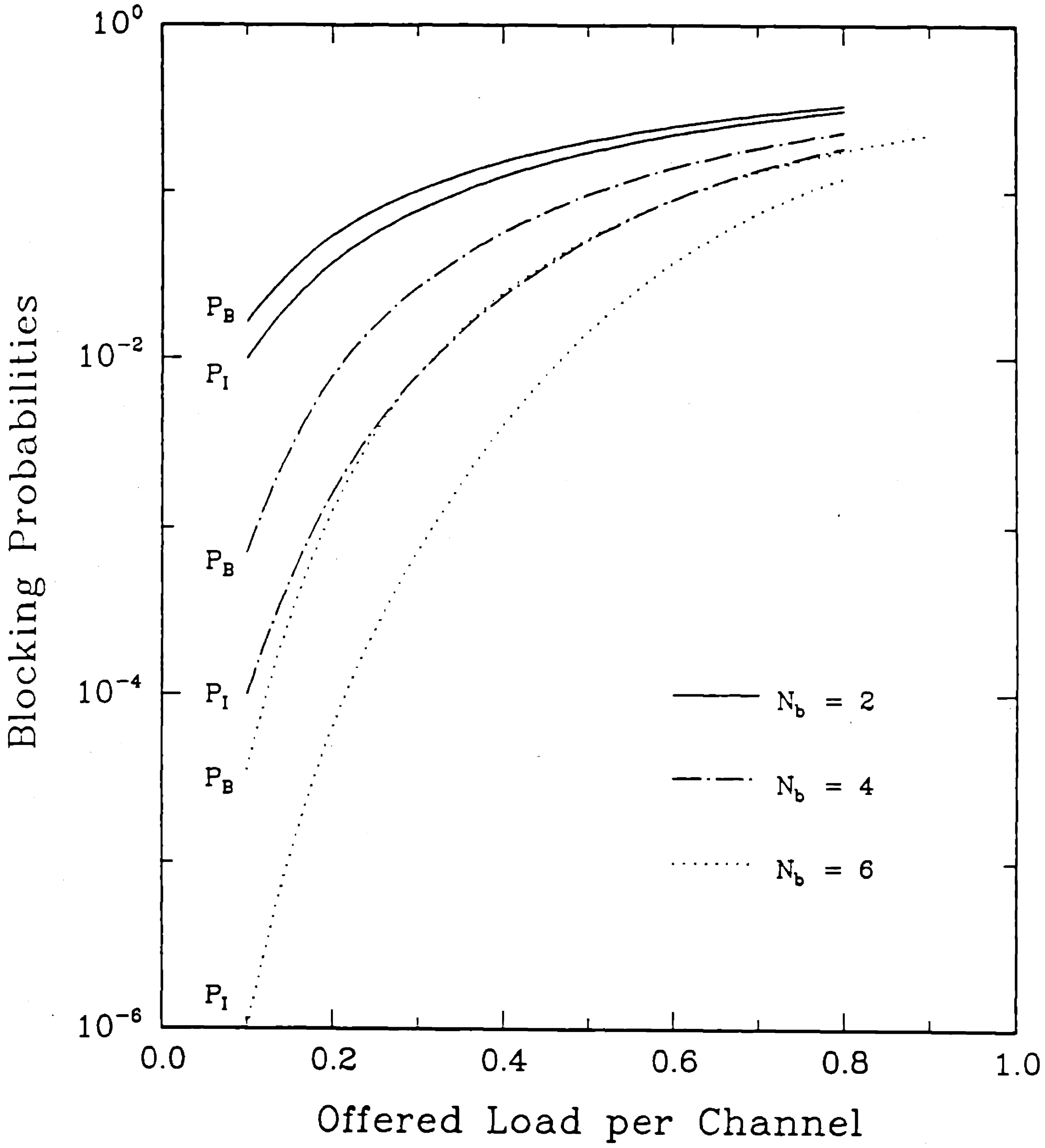


Figure 6.4: Blocking Probability For Ideal-Loss-Network and For $M/M/N_b/N_b$

Chapter 7

Optimal Access Structure and Traffic Splitting

In Chapter 6 we have given a detailed analysis of the ideal grading access structure. We have derived what we call the ideal lower bound on the blocking probability of ideal-loss-networks. In this chapter we shall continue to consider the optimal access structure or traffic splitting problem for uniform-accessibility networks. We shall first derive the analytical solution to the blocking probability for the W-loss network. Then two other major results will be presented. The first result is that we have demonstrated from the simple example with $M = 3$ and $N_b = 2$ that optimal traffic splitting is to split traffic somewhat "unevenly". The second result is that we have found out numerically that the ideal lower bound is not generally observed by other access structures.

7.1 Blocking Probability of W-Loss Network

7.1.1 Analytical Approach

In Chapter 5 we have proven that non-invasive hunting is the optimal scheduling rule for the W-loss network. From the policy iteration algorithm that we have implemented, blocking probability for any loss network can be obtained numerically. In this section we shall demonstrate that for the W-loss network, the blocking probability can actually be found through analytical approach.

Let us refer to Figure 2.2 once again and let us consider the class 1 arrivals that find channel 1 busy the overflow traffic from channel 1 and notate the overflow process as OT_1. Similarly, we consider the class 2 arrivals that find channel 3 busy the overflow traffic from channel 3 and notate the overflow process as OT_2. Overflow traffic cannot affect subsequent state of channel 1 or channel 3 or subsequent arrivals to either classes. Therefore channel 1 and channel 3 are independent and the two traffic overflow processes are also independent. Thus the total arrival process to channel 2 can be treated as the superposition of the two independent traffic overflow processes.

Corollary 7.1 *Each individual overflow traffic in a W-loss network is a renewal process.*

Given we have an overflow call from class 1, channel 1 must be busy and therefore subsequent overflows are independent of previous statistics. Therefore each individual overflow traffic is obviously a renewal process. Let $P_o(t)$ be the probability that the interval time between two arrivals from the same overflow stream is larger than or equal to t. In other words, $P_o(t)$ is one minus the interarrival time cumulative distribution function of each overflow process. The reason for using $P_o(t)$ is simply out of convenience for later calculations.

We can obtain the $P_o(t)$ characterization of each overflow process as follows:

$$P_o(t) \quad = \quad e^{-\lambda t} + \int_0^t (1 - e^{\tau})\lambda e^{-\lambda\tau} P_o(t-\tau)d\tau \tag{7.1}$$

The first term on the right-hand-side is the probability that there is no arrival from class 1 while the second term is the probability that there has been one class 1 arrival but channel 1 has already become idle so that it is not seen as an overflow.

Recognizing that the second term on the right-hand-side of equation 6.27 is a convolution integral and taking Laplace Transforms of both sides, we have,

$$P_o^*(s) \quad = \quad \frac{1}{s+\lambda} + P_o^*(s)\lambda(\frac{1}{s+\lambda} - \frac{1}{s+\lambda+1}) \tag{7.2}$$

where $P_o^*(s) = \mathcal{L}\{P_o(t)\}$, the Laplace transform of $P_0(t)$.

After some manipulations, we have,

$$P_o^*(s) \quad = \quad \frac{s+\lambda+1}{s^2+(s\lambda+1)+\lambda^2} \tag{7.3}$$

Since OT_2 is independent of OT_1, an OT_1 arrival can be regarded as a random incident of OT_2. The residual time before the next OT_2 arrival can be characterized by[1]:

$$P^*_{res}(s) = \frac{1}{s}(1 - \frac{P^*_o(s)}{\eta}) \tag{7.4}$$

where $P^*_{res}(s)$ is the Laplace transform of one minus the cumumlative distribution function of the residual time, and η is the expected interarrival time of each overflow process.

As $\frac{\lambda}{1+\lambda}$ is the probability that any call arrival will find the unshared channel busy, the overflow arrival rate of each overflow process is

$$\frac{\lambda^2}{1+\lambda}$$

Therefore, the expected interarrival time of each overflow process is,

$$\eta = \frac{1+\lambda}{\lambda^2} \tag{7.5}$$

Now given the characterizations $P^*_o(s)$ and $P^*_{res}(s)$, how can the blocking probability be found? Let us first consider what we call the *knockout* assumption:

Definition 7.1 The *knockout* assumption means that services can be pre-emptive and a subsequent arriving call can displace an established call on a channel and thus eject the call from the network.

[1] see [31] p172

Let us consider the following:

Corollary 7.2 *For a loss network with knockout, the knockout probability is the same as the blocking probability without knockout.*

This corollary is obvious since call durations are exponentially distributed the resulting state is the same no matter whether the new arrival is rejected or if the call in progress is knockout and rejected. The total number of calls rejected with or without knockouts will always be identical. Therefore, blocking probability as defined by the fraction of calls rejected(or ejected) is the same. However, knockout does favor the completion of shorter calls.

Therefore, for the W-loss network, the probability that an overflow call is not blocked by channel 2 is the same as the probability that the arrival will not be subsequently knocked out when knockout is allowed. This probability is the probability that the new arrival will depart before there is another overflow arrival from either class. This probability is given by:

$$\begin{aligned} P_x &= Pr\{\text{a call on channel 2 is not knocked out}\} && (7.6) \\ &= \int_0^\infty e^{-t} P_{res}(t) P_o(t) dt && (7.7) \end{aligned}$$

where e^{-t} is the probability density that service is completed at time t.

Therefore, blocking probability P_b is the probability that an arrival finds the unshared channel busy and also subsequently knocked out from channel 2. It is given by:

$$P_b = \frac{\lambda}{1+\lambda}(1 - P_x) \tag{7.8}$$

The right-hand-side of equation 7.7 can readily be recognized as the Laplace Transform of the product of the two time functions evaluated at $s = 1$. Multiplication in time domain corresponds to convolution in the frequency domain along an appropriate vertical contour. We have,

$$P_x \; = \; \frac{1}{i2\pi} \int_{\gamma+i(-\infty)}^{\gamma+i(+\infty)} P^*_{res}(1-s) P^*_o(s) ds \tag{7.9}$$

with the vertical contour in the domain of convergence of $P_o(t)$.

From equations 7.3 and 7.4, after some manipulations, we obtain:

$$P^*_{res}(s) = \frac{\eta[s^2 + (2\lambda+1)s + \lambda^2] - (s + \lambda + 1)}{\eta s[s - \frac{-(2\lambda+1)+\sqrt{4\lambda+1}}{2}][s - \frac{-(2\lambda+1)-\sqrt{4\lambda+1}}{2}]} \tag{7.10}$$

and

$$P^*_o(1-s) = \frac{2 + \lambda - s}{[s - \frac{(2\lambda+3)+\sqrt{4\lambda+1}}{2}][s - \frac{(2\lambda+3)-\sqrt{4\lambda+1}}{2}]} \tag{7.11}$$

A program has been written to compute the pole locations and the integral given by equation 7.9 using the residue method. Results are tabulated for different offered load per channel and are shown in table 7.1.

λ	α		1	2	3	4	5	P_z	P_b
1.35	0.90	Poles	.0000	-.5851	-3.1149	4.1149	1.5851	.3711	.3613
		residues[2]	.0000	.3568	.0143	-.0626	-.3084		
.90	.60	Poles	.0000	-.3276	-2.4724	3.4724	1.3276	.5005	.2366
		residues	.0000	.4895	.0109	-.0691	-.4314		
.75	.50	Poles	.0000	-.2500	-2.2500	3.2500	1.2500	.5603	.1884
		residues	.0000	.5510	.0093	-.0705	-.4898		
.60	.40	Poles	.0000	-.1780	-2.0220	3.0220	1.1780	.6312	.1383
		residues	.0000	.6239	.0073	-.0708	-.5603		
.45	.30	Poles	.0000	-.1133	-1.7867	2.7867	1.1133	.7143	.8868
		residues	.0000	.7092	.0050	-.0690	-.6453		$\times 10^{-1}$
.15	.10	Poles	.0000	-.0175	-1.2825	2.2825	1.0175	.9097	.1178
		residues	.0000	.9090	.0007	-.0455	-.8642		$\times 10^{-1}$
.015	.01	Poles	.0000	-.0002	-1.0298	2.0298	1.0002	.9923	.1140
		residues	.0000	.9923	.0000	-.0071	-.9852		$\times 10^{-3}$

Table 7.1: P_b for the W-Loss Network: Analytical Solution

[2] The values of the residues in Table 7.1 should actually be all multiplied by 2π. We have omitted doing so as the factor of 2π will be cancelled in the computation of the integral

7.1.2 Comparison with the Ideal-Loss-Network

The blocking probabilities for the W-loss network and the ideal-loss-network with $M = 3$ and $N_b = 2$ are shown in table 7.2 as a function of α, the offered load per channel. For the W-loss network non-invasive hunting scheduling policy is employed while for the ideal-loss-network, random selection is used.

α	W-Loss Network	Ideal Grading
0.9	0.36130	0.356901
0.6	0.23662	0.234412
0.5	0.18844	0.187500
0.4	0.13831	0.138817
0.3	0.088677	0.0905281
0.1	0.11781×10^{-1}	0.13353×10^{-1}
0.01	0.11404×10^{-3}	0.14848×10^{-3}

Table 7.2: Blocking Probabilities for the W-Loss Network and the Ideal-Loss-Network

What we can see from the above table is that at high offered traffic, ideal grading gives better performance while when offered traffic is small, W-loss network is better. The intuitive explanation for this behaviour is as follows: When offered traffic is small, the probability that all three channels are busy will be small, so most of the blocking occurs when the network has just two busy channels. With a W-loss network an optimal scheduling policy can minimize the probability that two channels in the same accessible group are busy given that two channels are busy so that no call will be blocked. For the ideal grading case, given that two channels are busy, one third of the calls will be blocked regardless of which two channels are busy. Thus the W-loss network can perform better at low traffic.

7.1.3 Optimal Traffic Splitting for $M = 3$, $N_b = 2$

For a uniform-accessibility M-channel network with $M = 3$ and $N_b = 2$, as the total number of classes is small, we can write a brute force search program to estimate the optimal traffic splitting among the three classes. Let:

$$AG_1 = (1,2)$$

$$AG_2 = (2,3)$$

$$AG_3 = (1,3)$$

What we have found, interestingly, is that the optimal splitting is to split traffic almost equally to class 1 and class 2, and then split some smaller fraction of traffic to class 3, as shown in Figure 7.1. The optimal splitting coefficients are dependent of the total traffic density and become almost equal when total offered traffic is very large. When offered traffic is small ρ_3 becomes almost zero. The optimal splitting coefficients as a function of total traffic density is shown in Figure 7.1. We shall notate as P_{opt} the blocking probability that can be obtained with optimal traffic splitting and optimal scheduling policy. P_{opt} is shown shown against P_B the Erlang B blocking probability for an $M/M/2/2$ network with same offered load per channel in Figure 7.2.

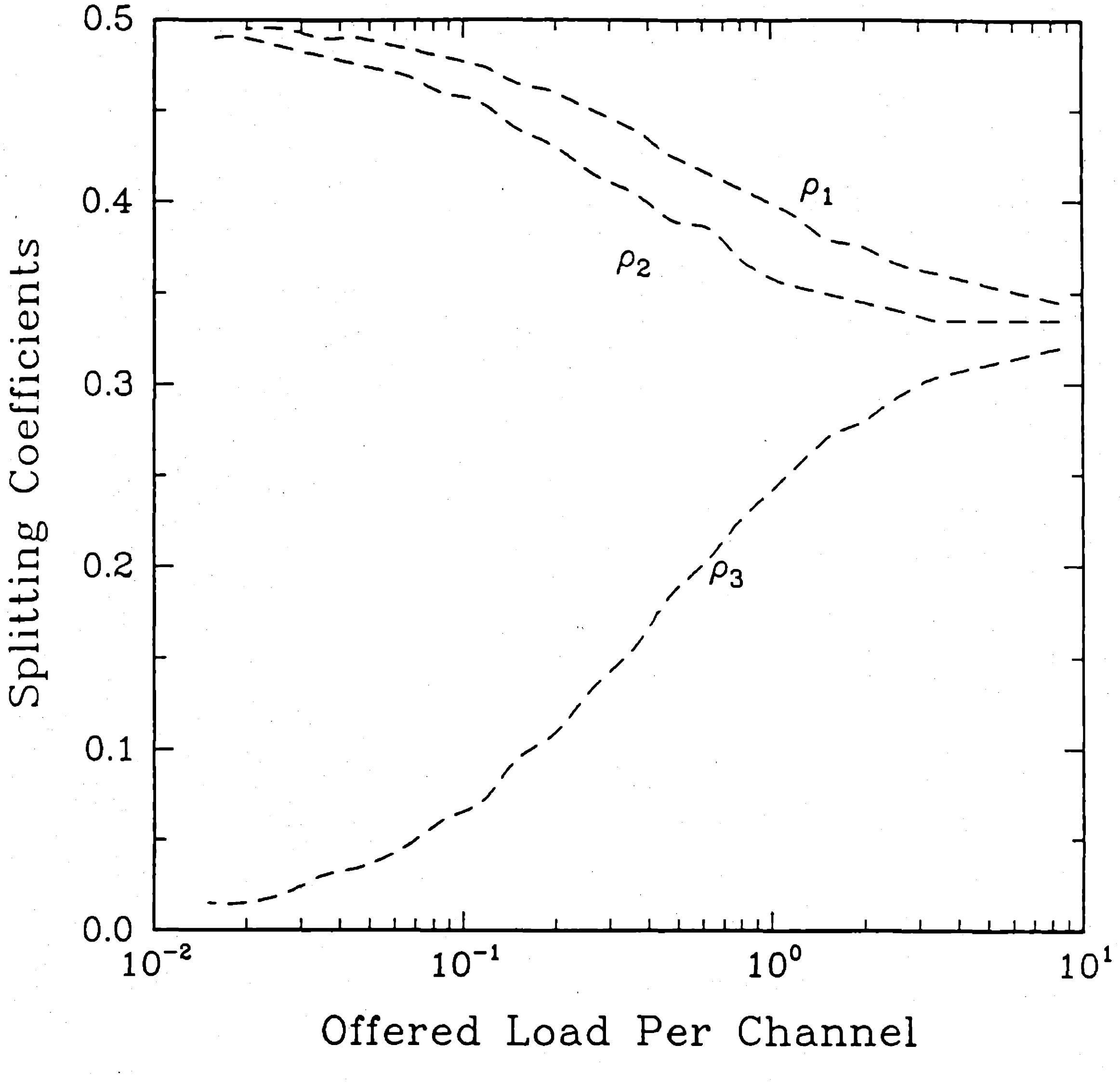


Figure 7.1: Optimal Traffic Splitting, $M = 3$, $N_b = 2$

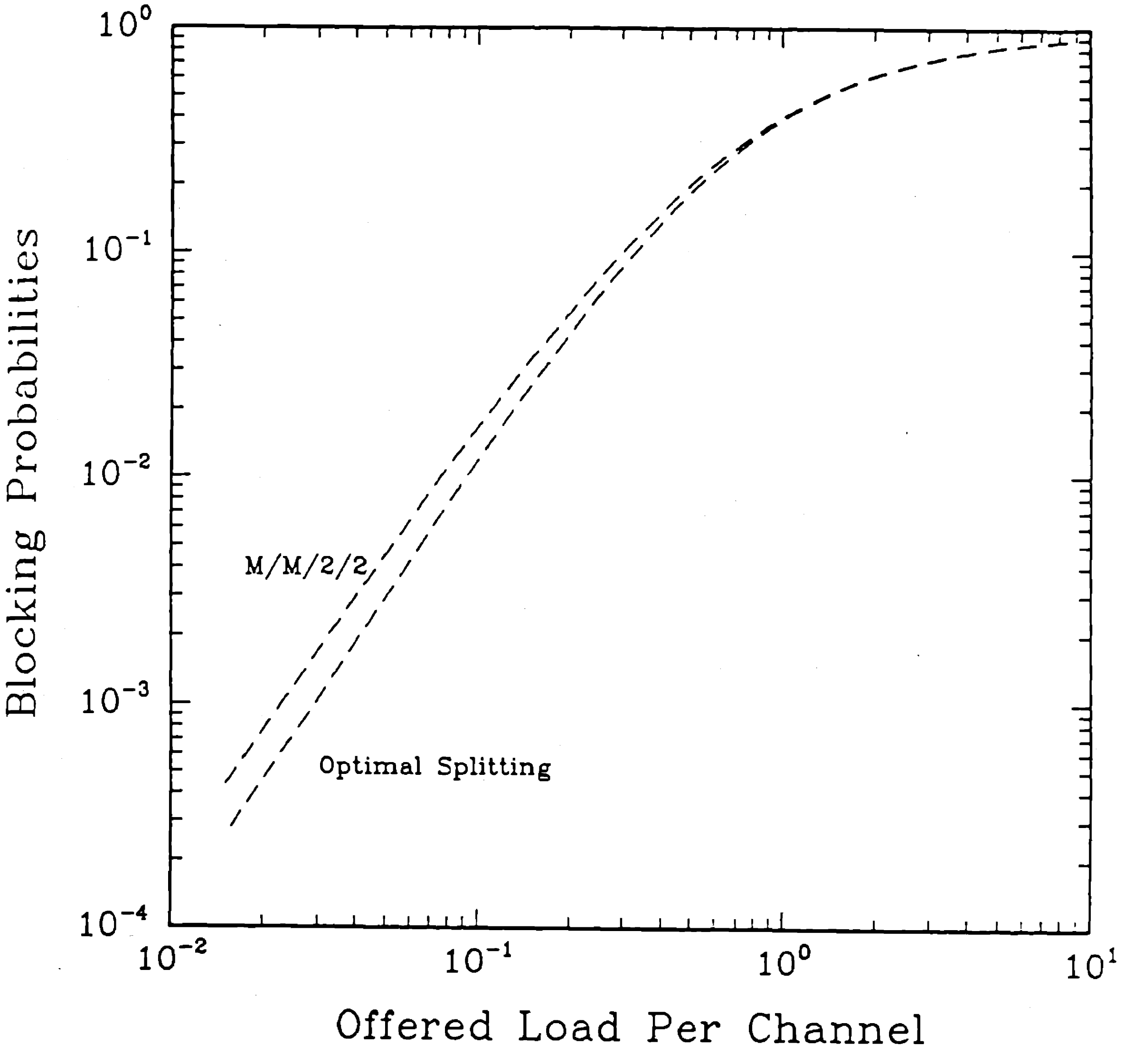


Figure 7.2: Corresponding Blocking Probability

In addition, we have found that the optimal scheduling policy is independent of the offered traffic. With $\rho_1 > \rho_2 > \rho_3$, the decision table $F = [f(i,k)]$ is given in table 7.3.

	f(i,k)		
S=$(b_1 b_2 b_3)$	k=1	k=2	k=3
(000)	1	3	3
(100)	2	3	3
(010)	1	3	3
(001)	1	2	1
(110)	0	3	3
(101)	2	2	0
(011)	1	0	1
(111)	0	0	0

Table 7.3: Channel Selection Table, M=3, N_b=2, Optimal Splitting

The above decision table indicates one interesting feature, which is the fact that for each class, a channel is still hunted for in a fixed order. Channel assignment is always made in the specific order of channel 3, 1, and 2. The intuitive reasoning is that channel 3 is shared by the two classes with the smallest traffic and thus should be used first, while channel 2 is shared by the two classes with the largest traffic and thus should be used last.

The result we obtained here is somewhat counter-intuitive in the sense that one

one normally expects optimality to be found at some symmetrical points. One reasonable explanation of the optimality of splitting traffic unevenly is as follows: when traffic to each class is uneven, given a choice, channel assignments can be made optimally, While if traffic is splitted evenly, certain states will become identical by symmetry and nothing can be gained through optimal scheduling. To illustrate this, say if ρ_1 and ρ_2 are equal and there is an arrival from class 3 while both channel 1 and channel 3 are idle, then assigning to either channel is no better than assigning to the other. On the other hand, if ρ_1 and ρ_2 are not equal, than making an optimal assignment may lead to some improvement. Therefore, the result we have obtained is not too surprising.

7.2 The Ideal Lower Bound

Let us first look at the loss network with $M = 3$ and $N_b = 2$. Again, let P_{opt} be the optimal blocking probability that is achieved by optimal traffic splitting, P_B the blocking probability of an $M/M/2/2$ queue, and P_I the ideal lower bound blocking probability. They are compared under the same offered load per channel, α, in Table 7.4.

α	P_{opt}	P_B	P_I
0.9	0.35560	0.3665	0.34791
0.6	0.23201	0.2466	0.21938
0.5	0.18462	0.2000	0.17157
0.4	0.13549	0.1509	0.12305
0.3	.086981	0.1011	.078720
0.1	$.11690 \times 10^{-1}$	$.1639 \times 10^{-1}$	$.98049 \times 10^{-2}$
0.01	$.11402 \times 10^{-3}$	$.1960 \times 10^{-3}$	$.99980 \times 10^{-4}$

Table 7.4: P_{opt}, P_B and P_I for M=3, N_b=2

The results in Table 7.4 show that for $M = 3$ and $N_b = 2$, while P_{opt} represents reduction of blocking probability by almost a half as compared to P_B at low traffic, it is always larger than P_I, the ideal lower bound blocking probability. Therefore the ideal lower bound is observed by any loss network with three channels and uniform accessibility of two.

However, when M increases and the access structure becomes more complicated, the ideal lower bound is not generally observed. For the case of $M = 6$, let us consider the following access structure with seven classes and even traffic splitting,

$$AG_1 = (1,2), AG_2 = (2,3), AG_3 = (3,4), AG_4 = (4,5),$$

$$AG_5 = (5,6), AG_6 = (6,1), AG_7 = (1,4)$$

and

$$\rho_k = \frac{1}{7} \quad \text{for } k = 1, 2, \ldots, 7$$

Let P_b be the blocking probability obtained for this network with optimal channel selection policy. At different α's, P_b and P_I are shown as follow:

α	P_b	P_I
0.9	.35328	.34791
0.5	.17841	.17151
0.3	$.80157\times10^{-1}$	$.76720\times10^{-1}$
0.2	$.38098\times10^{-1}$	$.37088\times10^{-1}$
0.1	$.94367\times10^{-2}$	$.98049\times10^{-2}$
0.01	$.81210\times10^{-4}$	$.99980\times10^{-4}$

Again, we can see that for small value of α's, the ideal lower bound is not satisfied. If the ideal lower bound is not observed for some access structures with a small value of $M = M_1$, we know that it will also not be satisfied for some access structures with larger value of $M = M_2$, or at least for the case when M_2 is a multiple of M as we can always divide a larger network with many channels into many smaller networks with fewer channels.

The results in this chapter shows that even for the simplest case, little can be said about the optimal access structure in terms of meaningful lower bound on the blocking probability.

The conclusion that can be drawn from this Chapter is that at low traffic, the performance of the W-loss network with non-invasive hunting is very close to the optimal. As can be seen observed from Table 7.2 and Table 7.4, at α's of 0.1 and 0.01, the W-loss network represents about 30% to 45% reduction in blocking probability as compared to the $M/M/2/2$ queue under the same load. This reduction should be of interest, particularly when we can see that the structure of the W-loss network is quite simple, and that the implementation of non-invasive hunting should not be too complicated. Individual users only have to attempt to transmit on unshared channels first. Complete knowledge of network state is not necessary. When traffic is high, both the W-loss network and the ideal-loss-network with random scheduling are fairly close to the optimal, although the latter is slightly better. The fractional reduction in blocking probability achieved by the optimal structure as compared to the $M/M/2/2$ queue, however, is rather small at high traffic.

Chapter 8

Simulation Study of the W-Hold Network

For hold networks, as the state space is infinite, the Markov decision formulation as developed in Chapter 4 cannot be applied. One approach is to use approximations by assuming some maximum queue sizes so that a finite-state-space system is obtained. Let us consider the W-hold network alone. If we assume that the maximum queue sizes allowed for the two queues are x_1 and x_2, the total number of possible states is:

$$N = 2^3(x_1 + 1)(x_2 + 1) \tag{8.1}$$

Knowing that the optimal scheduling policy must be unshared channel non-wasting, as proven in Chapter 5, we can eliminate from consideration all states that have a non-empty queue with the corresponding unshared channel idle. Then the remaining permissible states are:

- $(1b_21; n_1n_2)$, $n_1 = 0, 1, \cdots, x_1$, and $n_2 = 0, 1, \cdots, x_2$
- $(1b_20; n_10)$, $n_1 = 0, 1, \cdots, x_1$
- $(0b_21; 0n_2)$, $n_2 = 0, 1, \cdots, x_2$
- $(000; 00)$

The number of states one then have to consider will be:

$$N = 2[(x_1 + 1)(x_2 + 1) + (x_1 + 1) + (x_2 + 1) + 1] \tag{8.2}$$

We have not been able to prove that the optimal scheduling policy for the W-hold network is completely non-wasting in that it also does not hold calls in queue when the shared channel is idle. But assume that we restrict ourselves to non-wasting policies, the permissible states are:

- $(111; n_1n_2)$, $n_1 = 0, 1, \cdots, x_1$, and $n_2 = 0, 1, \cdots, x_2$
- $(110; n_10)$, $n_1 = 0, 1, \cdots, x_1$
- $(011; 0n_2)$, $n_2 = 0, 1, \cdots, x_2$
- $(000; 00)$, $(100; 00)$, $(001; 00)$, $(010; 00)$, and $(101; 00)$

Therefore, the number of states to be considered is:

$$N = (x_1 + 1)(x_2 + 1) + (x_1 + 1) + (x_2 + 1) + 5 \tag{8.3}$$

We can see that for the general case with no state reduction, even with small values of x_1 and x_2, the dimension of the problem will easily become too large to be handled.

Therefore, instead of the approximation approach, a simulation program is written to simulate the performance of the W-hold network under different scheduling rules. Instead of simulating the continuous time chain, we again make use of the idea of the imbedded chain. In each state $S = (b_1 b_2 b_3; n_1 n_2)$, the probabilities of various events to be the first one to occur are:

$$Pr\{\xi_1 = A_i\} = \frac{\lambda_i}{b_1 + b_2 + b_3 + \lambda_1 + \lambda_2} \tag{8.4}$$

$$Pr\{\xi_1 = D_i\} = \frac{b_i}{b_1 + b_2 + b_3 + \lambda_1 + \lambda_2} \tag{8.5}$$

where λ_1 and λ_2 are equal for the W-hold network.

SIMUL.f77 is the FORTRAN simulation program written In the simulation program, a random variable is generated in each state to decide which event is to occur next according to the probabilities given in equation 8.4 and 8.5. To compile statistics of the average queueing delay, we keep track of the number of waiting calls in queues each time a new call arrives. Starting from the initial state $S_o = (000; 00)$, the program simulates the occurrences of 800,000 events. Approximately half of these events will be arrivals as only a finite number of calls will be found in the final state, and thus the number of departures must be almost equal to the number of arrivals. Let $\overline{n}$ be the ensemble average of the total number of calls in both queues seen by an arriving customer, over the 800,000 events. Neglecting the effect of the particular choice of the initial state, we obtain from Little's theorem the average queueing delay experienced by a call as:

$$\overline{t_q} = \frac{\overline{n}}{\lambda_1 + \lambda_2} \tag{8.6}$$

First, by setting λ_2 to zero, we effectively have an $M/M/2$ queue. The validity of the simulation result is tested by comparing it against the calculated queueing delay of an $M/M/2$ queue. Excellent agreement is found. Then we apply the simulation program under two different non-wasting scheduling policies. Policy one, R_1, is to use both non-invasive hunting and select from longer queue rules. Policy two, R_2, is to use non-invasive hunting only and make random call selections whenever channel 2 becomes available while both queue 1 and queue 2 are non-empty. The reason of considering the two policies is as follows: The actual implementation of the non-invasive hunting rule appears to be relatively simple - transmitters only have to always attempt to transmit on the unshared channel first. The select from the longer queue rule, on the other hand, requires knowledge of the number of calls in both queues and may not be as easy to be implemented. Therefore we are interested in seeing how much degradation will be introduced if the select from longer queue rule is not implemented. The average queueing delay under the two different policies, is plotted along with the average queueing delay of an $M/M/2$ queue under the same offered load per channel. The result is shown in Figure 8.1.

What we have seen from the result is as follows. First, the average queueing delay of an $M/M/2$ queue ranges from approximately 20 percent higher in low load to approximately 30 percent higher in high load of 0.9 than that of a W-hold network under the optimal non-wasting policy R_1. Second, the average queueing delay under

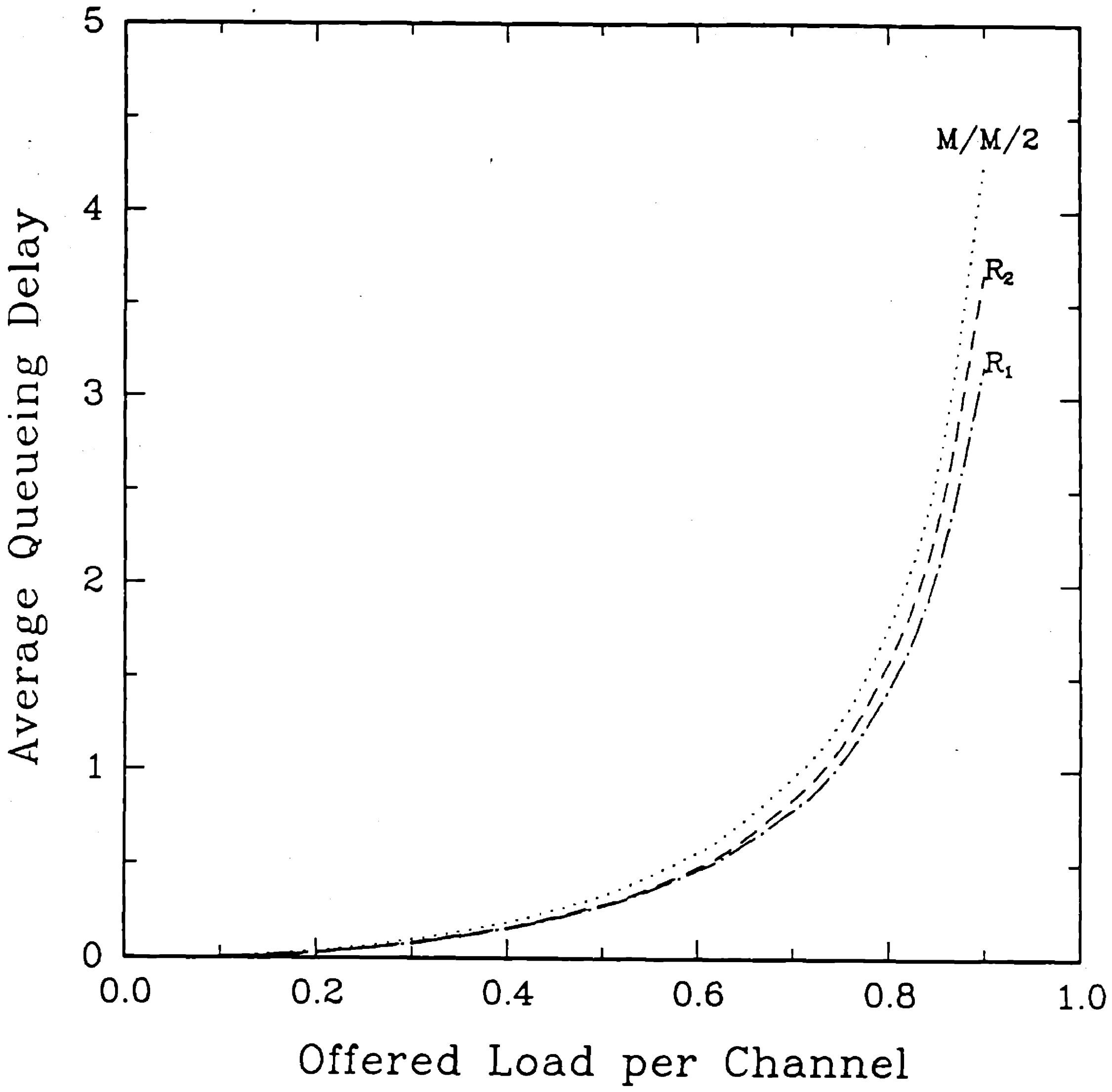


Figure 8.1: Average Queueing Delay in W-hold Network Under Optimal and Suboptimal Non-wasting Policies

R_2 is almost identical to that under R_2 in low load while is about 15 percent higher in a high load of 0.9. The intuitive explanation is that when offered load is small, the probability of finding both queues non-empty is small, and the probability of finding two non-empty queues with different number of waiting calls is even smaller. Therefore the select from longer queue rule is rarely actually applied even under R_1 and consequently there is no significant difference under the two policies.

In conclusion, by using the W-hold network with non-invasive hunting alone, queueing delay is reduced by roughly speaking 10 to 15 percent as compared to the $M/M/2$ queue. This should be of interest from a network design point of view in light of the ease with which this reduction can be achieved.

Chapter 9

Discussions and Summary

9.1 Comments on our Problem Model

9.1.1 The Loss Network Model and The Hold Network Model

One may question whether the loss network model or the hold network model is of more interest in a future metropolitan area network. If the emphasis in future networks is on datagram or packet switching, one then have to admit that the hold model should be more appropriate. The delay analysis in a real network, however, will almost certainly involve many more factors other than the queueing delay that we have considered in this thesis. The form of access control, for instance, will affect the queueing model that we should use, besides the control delay that should be included. In a network with switches, the delay encountered within a switch will most likely be very significant also. As a matter of fact, the switch is regarded by many as where the bottleneck is in a future broadband metropolitan area network.

One cannot, however, eliminate the possibility of finding application of circuit switching in future networks. For instance, if the data traffic in the network are, say, high bandwidth video sessions of relatively long durations with little room for statistical multiplexing, then the loss network model with blocking probability as performance measure may very well be the appropriate choice.

9.1.2 Access Control and Channel Scheduling

In this thesis, we have concentrated only on finding the optimal channel scheduling policy. In general, applying a channel scheduling policy requires the knowledge of the state of the network. In a network with a centralized access controller, it is easier for us to assume that the scheduling function is also performed by the centralized controller that has complete knowledge of the state of the network. In the subscriber loop scenario discussed in Chapter 1, access control is not an issue as the distribution center is the only node that transmits data. However, in networks with decentralized access control, say when users acquire use of channels through contention or possession of tokens, it becomes doubtful whether it will be realistic to have any sophisticated scheduling policy implemented. Such a consideration is another reason why we have concentrated on simple access structures such as the W-loss and W-hold networks. We have pointed out in Chapter 8 that the implementation of the non-invasive hunting rule should be quite simple, even without centralized control. It is not necessary for an individual node to have complete knowledge of the state of the network. All that an individual node has to do to effect non-invasive hunting is to attempt to transmit on the unshared channel in its accessible channel group first.

We have mentioned in the last section that the particular access control scheme employed is crucial in determining the delay analysis model that should be used. For instance, for various collision access schemes, the analysis of contention resolution delay may deviate substantially from that of the Markovian queueing delay model that we have adopted. Propagation delay, detection delay for carrier sensing or collision detection, retransmission strategy, etc., will all become important parameters to be considered. For other access control schemes, different sets of parameters will be of relevance. It is also worth pointing out that there exists a class of implicit-token demand assignment multiple access schemes which make use of the unidirectional nature of optical signal propagation. The idea is for nodes to sense and attach their own transmissions to the end of transmissions from upstream nodes. The physical ordering of nodes on the fiber will determine the order of transmission as well. A thorough discussion can be found in [9].

It is impossible for us to pin point a particular access control scheme for our works. Roughly speaking, the Markovian model we adopt should be accurate to the degree that access control delay is small when compared to call durations.

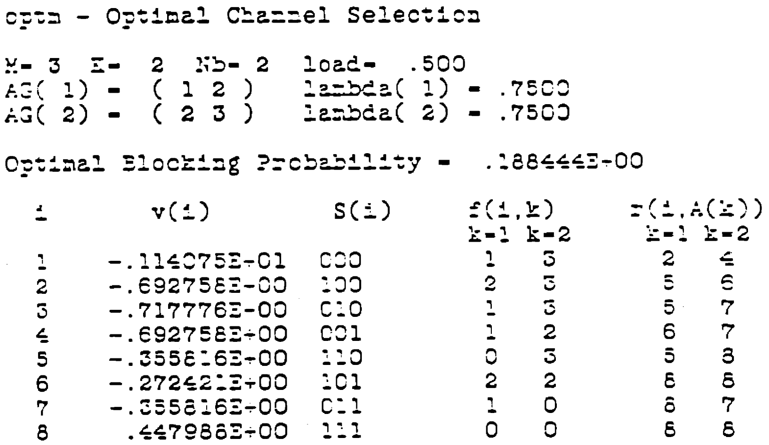
9.2 Summary of Results

In this thesis a number of ideas are developed. In Chapter 3 the ideas of deferred rejection and consistent deferred rejection are used to proof the optimality of the non-wasting rule for loss networks. In Chapter 4, we have used the ideas of the imbedded chain and the maximization of the g-objective, which is the ratio between the total number of departures and the total number of events. In Chapter 5, we have used the ideas of Tl-optimality and the maximization of the expected number of events up to any time T to prove the optimality of the non-invasive hunting rule for all loss and hold networks, the unshared channel non-wasting rule for all hold networks, and the selection from longer queue rule for the W-hold network. The conceptual operation of call duration randomization was also introduced to facilitate some sample path comparisons. In Chapter 6 we have used the concept of reversibility in Markov chains to prove the optimality of random scheduling for the ideal-loss-network and channel-reduced-ideal-network. From these two results the upper and lower bounds on the optimal blocking probability of an ideal-loss-network was obtained. In Chapter 7, the concept of knockout is used to deduce the blocking probability for the W-loss network through overflow traffic analysis.

From the numerical results, we have found out that relatively few general characterizations can be made about the optimal channel scheduling policy for arbitrary access structures. It does not appear that much can be deduced from the mathematical structure of the problem. However, focusing on the case when N_b, the accessibility of users, is small and is equal to 2, we have found out numerically and by simulation that indeed some performance improvement is possible through the use of some simple access structures such as the W-loss or W-hold networks. Here an important observation is that apparently little complexities have to be added for the use of these simple access structures. Finally, in our attempt to find the optimal access structure for networks with a particular number of channels and N_b, we have reached an interesting conclusion that minimium blocking probability is often achieved by some "uneven" or asymmetic access structures. This is somewhat counter-intuitive in that one may first expect optimality to occur at some symmetrical point. But then it is not so surprising, as the presence of asymmetries provides freedom for the optimal scheduling policy to operate on.

Appendix A

Sample Results From Optimization Program

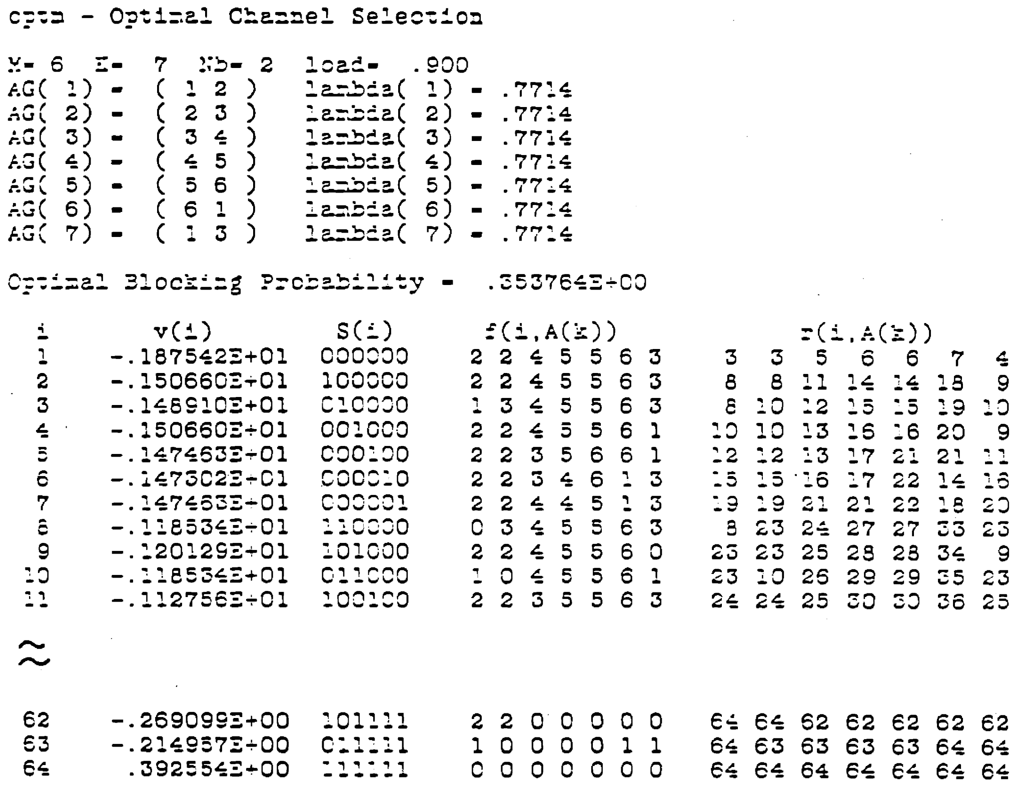


```
optm - Optimal Channel Selection

M= 3  K=  2  Nb= 2  load=  .500
AG( 1) =  ( 1 2 )   lambda( 1) = .7500
AG( 2) =  ( 2 3 )   lambda( 2) = .7500

Optimal Blocking Probability =  .188444E+00

  i        v(i)        S(i)     f(i,k)      r(i,A(k))
                               k=1 k=2      k=1 k=2
  1     -.114075E+01  000       1   3        2   4
  2     -.692758E-00  100       2   3        5   6
  3     -.717776E-00  010       1   3        5   7
  4     -.692758E+00  001       1   2        6   7
  5     -.355816E+00  110       0   3        5   8
  6     -.272421E+00  101       2   2        8   8
  7     -.355816E+00  011       1   0        8   7
  8      .447988E+00  111       0   0        8   8
```

Figure A.1: OUTDATA for the W-loss Network

```
optm - Optimal Channel Selection

M= 6  K=  7  Nb= 2  load=  .100
AG( 1) =  ( 1 2 )  lambda( 1) = .0857
AG( 2) =  ( 2 3 )  lambda( 2) = .0857
AG( 3) =  ( 3 4 )  lambda( 3) = .0857
AG( 4) =  ( 4 5 )  lambda( 4) = .0857
AG( 5) =  ( 5 6 )  lambda( 5) = .0857
AG( 6) =  ( 6 1 )  lambda( 6) = .0857
AG( 7) =  ( 1 3 )  lambda( 7) = .0857

Optimal Blocking Probability =  .101638E-01

 i       v(i)        S(i)       f(i,A(k))              r(i,A(k))
 1    -.284508E+01  000000   2 2 4 4 6 6 3     3  3  5  5  7  7  4
 2    -.234964E+01  100000   2 2 4 5 5 6 3     8  8 11 14 14 18  9
 3    -.234820E+01  010000   1 3 4 5 5 6 3     8 10 12 15 15 19 10
 4    -.234964E+01  001000   2 2 4 5 5 6 1    10 10 13 16 16 20  9
 5    -.234685E+01  000100   2 2 3 5 6 6 1    12 12 13 17 21 21 11
 6    -.234697E+01  000010   2 2 3 4 6 1 3    15 15 16 17 22 14 16
 7    -.234685E+01  000001   2 2 4 4 5 1 3    19 19 21 21 22 18 20
 8    -.187210E+01  110000   0 3 4 5 5 6 3     8 23 24 27 27 33 23
 9    -.187351E+01  101000   2 2 4 5 5 6 0    23 23 25 28 28 34  9
10    -.187210E+01  011000   1 0 4 5 5 6 1    23 10 26 29 29 35 23
11    -.185343E+01  100100   2 2 3 5 6 6 3    24 24 25 30 36 36 25

≈

62    -.458664E+00  101111   2 2 0 0 0 0 0    64 64 62 62 62 62 62
63    -.440073E+00  011111   1 0 0 0 0 1 1    64 63 63 63 63 64 64
64     .497441E+00  111111   0 0 0 0 0 0 0    64 64 64 64 64 64 64
```

Figure A.2: OUTDATA: Variability of Optimal Policy with Respect to Traffic Density, $\alpha = 0.1$



```
optm - Optimal Channel Selection

M= 6  E=  7  Nb= 2  load=  .900
AG( 1) =  ( 1 2 )   lambda( 1) = .7714
AG( 2) =  ( 2 3 )   lambda( 2) = .7714
AG( 3) =  ( 3 4 )   lambda( 3) = .7714
AG( 4) =  ( 4 5 )   lambda( 4) = .7714
AG( 5) =  ( 5 6 )   lambda( 5) = .7714
AG( 6) =  ( 6 1 )   lambda( 6) = .7714
AG( 7) =  ( 1 3 )   lambda( 7) = .7714

Optimal Blocking Probability =  .353764E+00

 i       v(i)         S(i)       f(i,A(k))              r(i,A(k))
 1     -.187542E+01  000000     2 2 4 5 5 6 3     3  3  5  6  6  7  4
 2     -.150660E+01  100000     2 2 4 5 5 6 3     8  8 11 14 14 18  9
 3     -.148910E+01  010000     1 3 4 5 5 6 3     8 10 12 15 15 19 10
 4     -.150660E+01  001000     2 2 4 5 5 6 1    10 10 13 16 16 20  9
 5     -.147463E+01  000100     2 2 3 5 6 6 1    12 12 13 17 21 21 11
 6     -.147302E+01  000010     2 2 3 4 6 1 3    15 15 16 17 22 14 16
 7     -.147463E+01  000001     2 2 4 4 5 1 3    19 19 21 21 22 18 20
 8     -.118534E+01  110000     0 3 4 5 5 6 3     8 23 24 27 27 33 23
 9     -.120129E+01  101000     2 2 4 5 5 6 0    23 23 25 28 28 34  9
10     -.118534E+01  011000     1 0 4 5 5 6 1    23 10 26 29 29 35 23
11     -.112756E+01  100100     2 2 3 5 5 6 3    24 24 25 30 30 36 25

≈

62     -.269099E+00  101111     2 2 0 0 0 0 0    64 64 62 62 62 62 62
63     -.214957E+00  011111     1 0 0 0 0 1 1    64 63 63 63 63 64 64
64      .392554E+00  111111     0 0 0 0 0 0 0    64 64 64 64 64 64 64
```

Figure A.3: OUTDATA: Variability of Optimal Policy with Respect to Traffic Density, $\alpha = 0.9$

Bibliography

[1] G. L. Abbas, *Frequency Allocation for Fiber Optic Integrated Services Communication Networks*, M.I.T. Electrical Engineering Ph. D. Thesis, Feb 1988.

[2] A. S. Acampora, "A Multichannel Multihop Local Lightwave Network", Submitted to the *IEEE Trans. Commun.*, May 1987.

[3] C. E. Bell and S. Stidham, Jr., "Individual Versus Social Optimization in the Allocation of Customers to Alternative Servers," *Management Science*, Vol. 29, No. 7, Jul 1983.

[4] J. Bellamy, *Digital Telephony*, Wiley (1982).

[5] D. P. Bertsekas, *Dynamic Programming, Deterministic and Stochastic Models*, Prentice-Hall Inc. (1987)

[6] V. E. Benes, "Programming and Control Problems Arising from Optimal Routing in Telephone Networks," *Bell System Technical Journal*, Vol. XLV, No. 1, Nov 1966.

[7] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall Inc. (1987).

[8] S. P. Bradley, A. C. Hax and T. L. Magnanti, *Applied Mathematical Programming*, Addison-Wesley, Reading Mass. (1977).

[9] M. Fine and F. A. Tobagi, "Demand Assigment Multiple Access Schemes in Broadcast Bus Local Area Networks," *IEEE Trans. Computers*, Vol. c-33, No. 12, pp. 1130-1159, Dec 1984.

[10] G. Y. Fletcher, *et. al.*, "A Queueing Network Model of a Circuit Switching Access Scheme in an Integrated Services Environment," *IEEE Trans. Commun.*, Vol. COM-34, # 1, pp. 25–29, Jan 1986.

[11] G. J. Foshini, "On Heavy Traffic Diffusion Analysis and Dynamic Routing in Packet Switched Networks," *Computer Performance*, K. M. Chandy and M. Reiser (EDS.), North Holland Publishing Company (1977)

[12] Howard, *Dynamic Programming and Markov Processes*, M.I.T. Press (1960).

[13] J. Y. Hui, "Pattern Code Modulation and Optical Decoding – A Novel Code-Division Multiplexing Technique for Multifiber Networks," *IEEE Journal on Selected Areas in Comm.*, Vol. SAC-3 NO. 6, Nov 1985.

[14] J. Y. Hui and E. Arthurs, "A broadband Packet Switch for Integrated Transport," *IEEE J. Select. Areas Commun.*, Vol. SAC-5, Oct 1987.

[15] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley (1979).

[16] L. Kleinrock, *Queueing Systems Vol. I: Theory*, Wiley (1975)

[17] L. Kleinrock, *Queueing Systems Vol. II: Computer Applications*, Wiley (1976)

[18] R. W. Klessig, "Overview of Metropolitan Area Networks," *IEEE Communications Magazine*, Vol. 24, pp. 9–15, 1986.

[19] B. Kraimeche amd M. Schwartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks," *IEEE Trans. Commun.*, Vol. COM-33, # 10, pp. 1085–1093, Oct 1985.

[20] H. J. Kushner and C. H. Chen, "Decomposition of Systems Governed by Markov Chains," *IEEE Trans. on Automatic Control*, Vol. AC-19, No. 5, Oct 1974.

[21] S. C. Liew, *Topologies and Power Division Problem of Fiber Optic Networks*, M.I.T. Electrical Engineering Master's Thesis, Feb 1986.

[22] S. C. Liew, *Capacity Assignment In Non-Switching Multichannel Networks*, M.I.T. Electrical Engineering Ph.D's Thesis, Feb 1988.

[23] M. E. Marhic, Y. Birk, and F. A. Tobagi, "Selective Broadcast Interconnection: A Novel Scheme for Fiber-Optic Local-Area Networks," *Optics Letters*, Vol. 10, No. 12, Dec 1985.

[24] M. A. Mason and D. Roffinella, "Multichannel Local Area Network Protocols," *IEEE J. Select. Areas Commun.*, Vol. SAC-1, No. 5, Nov 1983.

[25] P. Nain and K. W. Ross, "Optimal Priority Assignment with Hard Constraint," *IEEE Trans. on Automatic Control*, Vol. AC-31, No. 10, Oct 1986.

[26] K. Nosu, "Fiber-Optic Wavelength-Division-Multiplexing Technology and Its Application," *Japan Annual Reviews in Electronics, Computers and Telecommunications*, Vol. 5, 1983.

[27] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, NJ (1982).

[28] Z. Rosberg, P. P. Varaiya, and J. C. Walrand, "Optimal Control of Service in Tandem Queues," *IEEE Trans. on Automatic Control,* Vol. AC-27, No. 3, Jun 1982.

[29] E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering*, Prentice-Hall Inc. (1976)

[30] D. R. Smith and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," *Bell System Technical Journal*, Vol. 60, No. 1, Jan 1981.

[31] M. G. Smith, *Laplace Transform Theory*, D. Van Nostrand Company Ltd. (1966)

[32] Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver & Boyd, Edinburgh (1960).

[33] Daniel T. W. Sze, "A Metropolitan Area Network," *IEEE Journal on Selected Areas in Communications*, pp. 815–824, Nov 1985.

[34] A. S. Tanenbaum, *Computer Networks*, Prentice Hall, Eaglewood Cliffs, NJ (1981).

[35] S. S. Wagner, *Multiplexing Methods for Fiber Optic Local Communication Networks*, M.I.T. Electrical Engineering Ph.D's Thesis, June 1985.

[36] O. J. Wasem, *Topologies for Fiber Optic Local Communication Networks with No Switch*, Proposal for Ph.D. Thesis Research, M.I.T. EECS Department, Sept 1987.

[37] R. Watanabe, "Optical Multiplexer and Demultiplexer," *Japan Annual Reviews in Electronics, Computers and Telecommunications*, Vol. 5, 1983.

[38] R. R. Weber, "On the Optimal Assignment of Customers to Parallel Servers," *J. Appl. Prob.* **15**, 406-413 (1978)

[39] W. Winston, "Optimality of the Shortest Line Discipline," *J. Appl. Prob.* **14**, 181-189 (1977).

[40] G. Winzer, "Wavelength Multiplexing Components - A Review of Single Mode Devices and Their Applications," *IEEE J. Lightwave Tech.*, Vol. LT-2, No. 4, Aug 1984.

[41] R. W. Wolff, "An Upper Bound For Multi-channel Queues," *J. Appl. Prob.* **14**, 884-888 (1977)

[42] A. K. Wong, *Channel Scheduling for Optical Communication Networks with Frequency Concurrency*, Proposal for Ph.D. Thesis research, M.I.T. EECS Department, Jun 1986.

[43] C. Yeh and M. Gerla, "High Speed Fiber Optic Local Networks," *Proc. of the Tenth Anniv. Meeting of the NSF Grantee-User Group in Optical Communications Systems*, pp. 69–80, June 1982.

[44] Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *IEEE J. Slect. Areas Commun.* Vol SAC-5, No. 8, 1987