PH.D. THESIS

Dynamic Routing of Voice/Data-Integrated and ATM-based Hybrid Networks

by S. Chen
Advisor: J.S. Baras

CSHCN Ph.D. 94-1
(ISR Ph.D. 94-15)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Abstract

Title of Dissertation: Dynamic Routing of Voice/Data-Integrated and ATM-based Hybrid Networks

Shihwei Chen, Doctor of Philosophy, 1994

Dissertation directed by: Professor John S. Baras
Department of Electrical Engineering

A hybrid network consisting of a satellite network and a terrestrial network will increase the overall network efficiency considerably by using all available resources and media. This dissertation considers dynamic routing in both voice/data-integrated and ATM-based hybrid networks.

Optimal dynamic routing in such mixed-media (voice/data-integrated) networks under Markov Queueing Modeling has been developed and solved. Routing problems in such a domain usually lead to a weighted-sum minimization or a minimax problem. A new approach to obtain the trade-off curve of multiple-objective optimization is outlined. With a numerical optimization package, we can plot the trade-off curve exactly.

Both a centralized and a distributed implementation of this problem with Kalman filter techniques and Equilibrium Programming are presented. These techniques allow the control of a large and stochastic network by communicating with a group of communication managers in a parallel manner.

For ATM-based hybrid networks, an economic model with the objective of maximizing the "social welfare" has been adapted. This model can be developed

into a form of a two-player game. With a little modification and adaptation, we will be able to solve the joint problem of access control and routing in both the weighted-sum formulation and the economic formulation.

Dynamic Routing of Voice/Data-Integrated and ATM-based Hybrid Networks

by

Shihwei Chen

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1994

Advisory Committee:

Professor John S. Baras, Chairman/Advisor
Professor Mike Ball
Professor Anthony Ephremides
Professor Armand Makowski
Professor André L. Tits

Dedication

To my beloved parents, brothers, sister, and my lovely wife

Acknowledgements

I would like to express my gratitude to my advisor, Dr. John S. Baras, for guiding me and directing my research topic through his busy schedule. Through many meetings and classes, I was able to learn a lot from him. His achievements and attitudes have inspired me in many ways. His advice is always pertinent; it has led me to new ideas for solving the problems I have faced.

There are many people who have made this dissertation possible; without them it would have been impossible for me to finish it. Though it is difficult to name them without omitting someone, I am happy to acknowledge them here, and I apologize if I have forgotten anyone. First of all, I would like to thank Dr. Jian L. Zhou and Dr. André L. Tits for providing the FSQP program to assist solving the optimization problem. Dr. Tits also gave me much useful advice when needed. I am obliged to members of my thesis committee as well.

My fellow students and office mates here at Institute for Systems Research and staff of Center for Satellite and Hybrid Communication Networks have been very helpful in discussions and inspirations on my work. Among them, Dr. I-hung Lin and Dr. Wen-bin Yang are especially appreciated. I am also very thankful to Mr. Donald Hirsch, Dr. Bill Sutherland, Dr. Samuel Baum, and Dr. Jack Kleiman for editing my thesis.

The Network Management Group is another contributor to my dissertation. Dr. Jayant Haritsa gave me many hints and ideas. He and Dr. Anindya Datta offered me their graphics to be used in Chapter 2 and Chapter 4. Dr. Steve Low at AT&T provided me his papers (see the bibliography) as a basis of Chapter 5.

My family has always been very supportive of my studies, I am certainly most

grateful to them. Last but not the least, I want to thank my wife, Kuo-hua, for her love and support which have helped to complete this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A mixed-media [1] network is an integrated, heterogeneous network which consists of several subnetworks. The transmission media for these subnetworks are different. For example, these subnetworks could be roughly classified either as a terrestrial network or a radio network. A terrestrial network could be a telephone network or a computer network. A radio network could be comprised of a satellite network or a cellular mobile network, etc. To make it distinctive from multi-media traffic, we also term such a mixed-media network a "hybrid network".

A hybrid network consisting of satellites and terrestrial networks will offer connections to areas where optic fibers are hard to deploy, like remote districts, or mobile terminals. Such hybrid networks will provide broadcast functions as well as more efficient and flexible use of network resources.

The study of hybrid networks is important because the overall network efficiency can be considerably increased by using all available resources and media. In some situations, the combined use of all media may provide connectivity,

[1] This term should be made distinctive from multimedia which means multiple traffic of voice, data, video, etc.

whereas use of a single medium may not. The main problems to be addressed in this thesis are the designs of the dynamic routing algorithms for hybrid networks including ATM-based hybrid networks.

A voice/data-integrated hybrid networks is a hybrid network that carries both voice traffic and data traffic within the same transmission media. Data networks (computer networks) and voice networks (telephone networks) were developed independently. They were originally designed to carry only one traffic type, i.e., either voice or data traffic, but not both. However, as more multi-media applications are emerging, there are substantial needs for voice/data integration techniques. For example, the telephone networks can now send data traffic with various modems and the computer networks can transmit voice traffic by using digital techniques such as quantization, compression, digital signaling processing, etc.

Since some adaptation or interface devices and techniques are required for current networks to carry multi-media traffic, new technique such as Asynchronous Transfer Mode (ATM) was invented to be an universal architecture for multi-media traffic including voice and data and more other types such as fax, image, video, graphic, etc. ATM is a fast packet switching technique based on a connection-oriented method. ATM will be deployed gradually in next five to ten years. Meanwhile the current voice and data networks will still exist for some time. Therefore, substantial research efforts, especially in routing problems, are required for multi-media applications for both voice/data-integrated networks and ATM-based networks at current time.

This thesis concentrates on the routing problem of hybrid networks for both voice/data-integrated traffic and ATM traffic. We begin with the routing prob-

lem in the domain of a general, existing network architecture, that is, a terrestrial network consisting of voice and data sub-networks and there is one satellite in the model. Then we extend the problem to ATM-based hybrid networks in which the terrestrial network under consideration is a broadband ATM network.

1.1 Routing problem for voice/data-integrated networks

The routing problem in a single network domain is well-studied. For example, in a telephone network where voice calls are transmitted and switched mainly by circuit-switched methods, the routing algorithms have actually been put to use in our daily lives for years [3, 4, 25]. Some optimal results have been obtained in the theoretic research [6, 24, 36, 37, 38]. On the other hand, in a computer network where data messages are mostly put in a packet format and then transmitted and switched by the packet-switching methods, the same fruitful results have been achieved both in industrial [18, 47, 54] and research [9, 11, 15, 16, 33, 34] sectors.

The study of the routing problems in both of the above distinctive network types is fundamental to achieve new methodologies in an integrated network. With the advent of the ISDN age, the boundary between the above terrestrial networks has been blurred. As a matter of fact, the routing problem in an integrated packet-switching and circuit-switching network calls for more advanced considerations than mere combination of the existing routing schemes. Innovative schemes have been proposed in different voice/data integrated systems [21, 41, 45], which are currently an active research area. However, in a mixed-media network with voice/data integrated traffic, the routing control problem is even more complex and complicated, and thus deserves a more complete and

concrete study.

The routing problem is a network function closely related to flow control, access control, bandwidth allocation, priority assignment, etc. If we try to envision the future Broadband ISDN era which is now under planning, the routing problem virtually becomes a bandwidth allocation problem. Indeed, if the bandwidth is wide enough to accommodate the usual load of the traffic, we can put all the messages in the same direct link without conflict. Under normal conditions, we do not have to deviate the flow through the alternate routes which are used as backups in a failure situation [5].

In summary, with the imminent advent of BISDN deployment and increasing use of satellites to share the daily information transmission load of ground traffic [40, 59], routing considerations in the hybrid network domain have become more important and meaningful. Therefore, the routing problem in mixed-media networks with integrated voice/data traffic is significant and plays a major role in the success of future communication systems.

1.1.1 Previous work

There have been several related articles in the routing design of a mixed-media network. Huynh et al. [28] presented the optimal design of routing and capacity assignment in mixed-media packet-switched networks consisting of a ground subnet and a satellite subnet. In their algorithm, they assumed linear cost-capacity functions for both terrestrial and satellite links and a fixed-split routing policy in their terrestrial subnetwork. The first assumption makes their capacity assignment problem mathematically tractable, and thus a closed-form solution was obtained with analytic procedures involving Lagrange multipliers. The second

assumption eliminates the routing problem within the terrestrial subnetwork. It also reduces the flow assignment problem to one of determining the optimal amount of traffic going through satellite links for each pair of source-destination nodes.

Using routing algorithms originally developed for terrestrial subnetworks, Gerla et al. [20] have evaluated network throughput and delay of mixed-media networks for satellite access methods such as point-to-point access and multi-access with or without channel contention.

Maruyama reviewed this problem with an extension to a combined problem of capacity, priority, and flow assignment (CPFA) [46]. He assumed discrete cost-capacity functions for nodes, terrestrial links and satellite links, and non-bifurcated routing within each packet class. He limited the problem to dedicated satellite channel access methods for which exact delay computations are possible. The basic idea of his algorithms to obtain the optimal solutions is by trial-and-error and by iterative numerical methods.

More recently, Yuan and Baras focused on the control of multimedia [2] communication processors where the transmission of messages has a time constraint given by the user [62]. They approximated the delay distribution by using a product-form network model. A DFT (Discrete Fourier Transform) algorithm was used to compute a time threshold from the delay distribution so that 99 percent of messages arrived before the threshold. From this threshold, they derived the "Gittins index" which guides the processor to switch the packets to the subnetwork with the minimum discounted cost (delay).

[2]The multimedia here actually means mixed-media, referring to different transmission media (networks).

However, these papers considered only one uniform transmission and switching mode, i.e., only the packet-switching method for data transmission among source-destination pairs.

To consider a voice/data integrated system, we must modify the objective cost function to include a performance measure for voice traffic. In a telecommunication network, the blocking probability of the voice transmission is the major concern for setting up phone calls and should be minimized, because voice traffic must be transmitted in a continuous stream with very low variability of the time delay, while calls which do not get the resources to transmit are blocked and cleared. In contrast, data traffic, which may be either bursty or regular in nature, can be delayed and buffered for later transmission. It is the delay for packet transmission of data traffic that we want to minimize. Thus, the overall objective function of an integrated voice/data system is usually a weighted sum of blocking probability for voice traffic and delay for data traffic.

For example, Gerla and Pazos-Rangel [21] considered the bandwidth allocation and routing problem in ISDN's, using a linear combination of the blocking probability and packet delay as their objective function. They formulated the problem as a constraint nonlinear programming problem, which has a special structure to be exploited and solved by the Frank-Wolfe steepest descent algorithm [44].

A similar type of objective function is used by Viniotis-Ephremides. They use the theory of Markov decision process and dynamic programming to obtain the optimal admission and routing strategy at a simple ISDN node [58]. The results are characterized by the same series of "switching curves". Results of the same kind have been obtained by Lambadaris-Narayan for a circuit-switched

node [39]. However, these results are limited to a system with a low degree of dimension, that is, a network of one or two queues (or a simple ISDN node).

More commonly, the previous authors consider the voice/data integration system in a single domain, not in the mixed-media network domain. In conclusion, the problem is a two-level integration. On the first level, we have voice/data integration and mixed-media domains integration. On the higher level, we consider the further integration of the problems in the first level. Namely, the optimal routing design in the mixed-media network with integrated voice and data traffic.

We will examine the dynamic routing problem in Chapter 2 and discuss the trade-off phenomenon in voice/data integration in Chapter 3. Based on these methodologies, we propose both a centralized and a distributed implementation architectures using Kalman filter techniques and Equilibrium Programming in Chapter 4.

1.2 Dynamic routing for ATM-based hybrid networks

ATM is a connection-oriented way of switching. A virtual circuit is established before the information packets (cells) are pumped into the channel. Since the processing overheads are reduced to a minimum by taking some network layer functions, such as routing and error control at the intermediate nodes, to the higher layers of the end nodes, extremely low delay is attainable. This makes time/delay sensitive traffic, such as voice and video, transportable over ATM links. Such fast packet switching is feasible in optical fiber networks with considerably low bit error rate.

The basic principles of Asynchronous Transfer Mode (ATM) were evolved and

developed through a couple of phases (See, for example, De Prycker [14]). However, the design goal of this new transfer/switching technique is to pave a bridge to the future National Information Infrastructure (NII) which makes multi-media services (voice, video, data, images, fax, ..) available and transportable through a common backbone structure of upcoming Broadband Integrated Services Digital Networks (BISDNs).

To achieve this goal, fast packet switching like ATM was proposed. ATM, based on fixed-length cells, can transport traffic up to 150 Mega Bits Per Second (MBPS) or higher. Therefore, we can transmit various bit rate motion pictures along with high bandwidth graphics, low data rate voice and data all over the same switching fabrics, which is not possible through the existing networks.

To differentiate the types of services available through ATM networks, we commonly associate a set of Quality of Service (QOS) parameters to each class of services. Network service providers must guarantee the QOS requested by the users throughout the connection duration. The main QOS parameters are delay and cell loss probability.

The goal of the dynamic routing problem for ATM-based networks is the same as before. However, the terrestrial network under consideration is now an ATM network. Because the exact analytical expressions for cell delay or cell loss ratio, etc., are not already available or not already verified, another approach must be adapted. We formulate the dynamic routing problem of mixed-media networks with ATM terrestrial subnets based on a minimax problem. This will be presented in Chapter 5. As an extension of this formulation, we can solve the access control problem at the same time. This is also addressed in Chapter 5.

Chapter 2

Routing Problem As Multi-Objective Optimization

In this chapter, we formulate our routing problem in mixed-media networks as a multi-objective optimization problem. There are four objectives which should be minimized: delay, blocking probability in the terrestrial network, delay, and blocking probability in the satellite network. A common approach to such a multi-objective optimization problem is to optimize the weighted-sum of the objectives. This results in a nonlinear programming problem with nonlinear constraints.

2.1 Model

A mixed-media network could comprise several subnetworks. However, to simplify the problem, we will consider a communication network composed of two subnetworks: one is the ground subnet, the other the satellite subnet. One such network is shown in Fig. 2.1, which shows a mixed-media network over the continental US with one satellite. A mixed-media network with more than two subnetworks will be considered in the future, though the problem may become harder. The nodes are the locations of interface message processors (IMPs) linked together by landlines. There are special nodes called satellite IMPs (SIMPs)

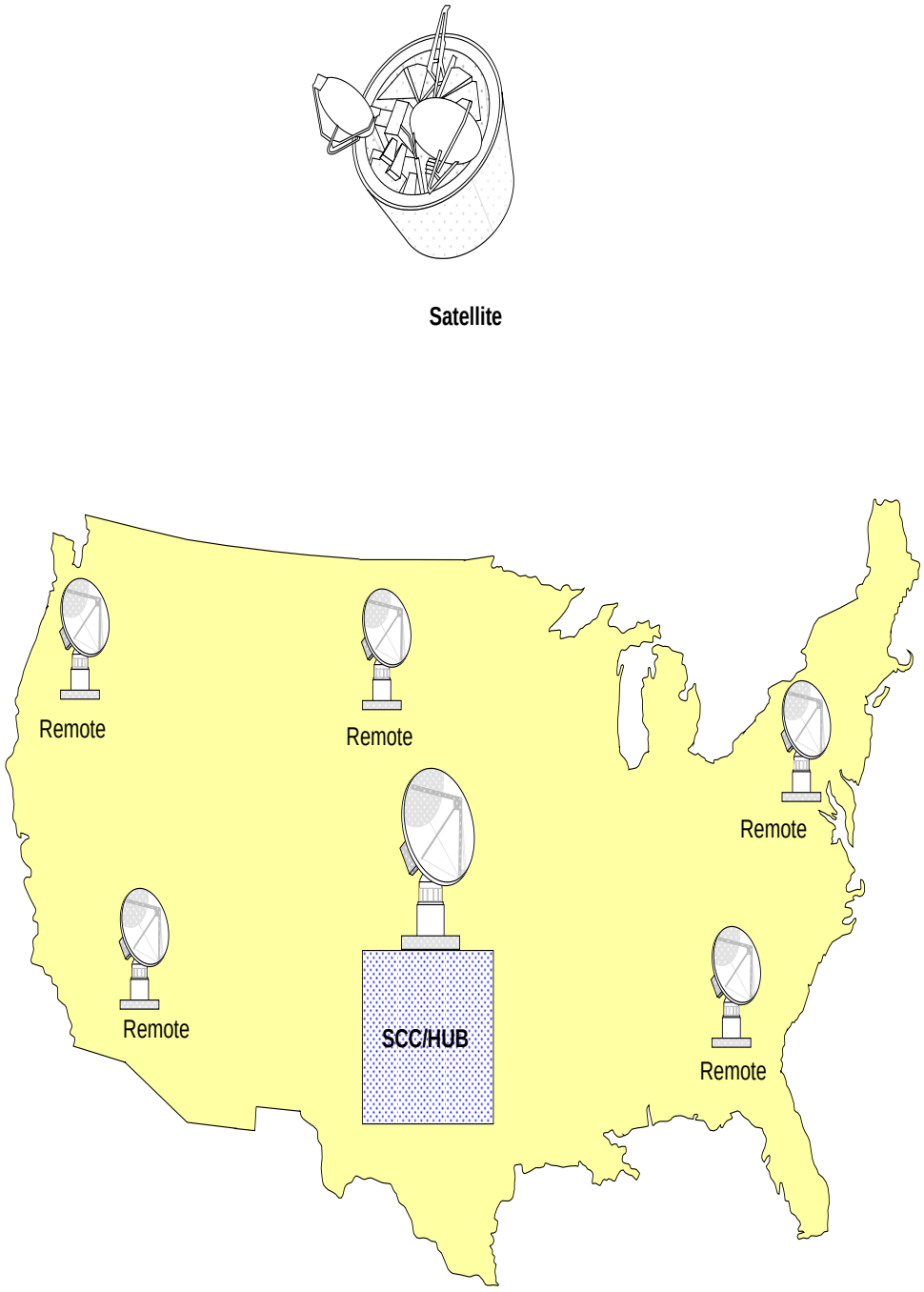

Figure 2.1: A mixed-media network. SCC: Satellite Control Center

which are interface message processors between the satellites and the ground links.

Routing in a mixed-media network consists of two major portions: (1) splitting the input traffic at SIMPs between ground and satellite subnets; (2) routing in the ground subnet. The traffic of the ground subnet consists of voice traffic and data traffic. Our design problem can be stated as follows. Given a network topology, a set of input traffic rates, a ground routing procedure, and link capacities, we want to minimize a weighted sum of the average delays of data traffic in both the ground subnet and the satellite subnet and the blocking probabilities of voice traffic in both the ground subnet and the satellite subnet over the set of traffic splitting factors for data and voice respectively.

Let g_{ij} (s_{ij}) be the splitting factor of data (voice) traffic which specifies the fraction of data (voice) traffic, originating at node i and destined for node j, and going through the ground subnetwork. In the following, we will derive the overall objective function.

2.1.1 Data delay in the ground subnet: $/M/M/1$ model

Suppose that we are given a data subnetwork in the ground of N nodes linked by L ground links of capacities C_{gdl} (bits/sec), $l = 1, 2, ..., L$, in a specified topology. The network is partitioned into M regions, each having a SIMP. These M SIMPs are linked via a satellite channel of capacity C_s (bit/sec). A traffic rate matrix $[r_{ij}]$ specifies, in packets/sec, the average rates of messages flowing between all possible IMP pairs i and j, where $i, j = 1, 2, ..., N$.

If we make the following (Kleinrock independence) assumptions: Poisson arrivals at nodes, exponential distribution of packet length, independence of ar-

rival processes at different nodes, and independence of service times at successive nodes, then we have the following expression for the average data delay in the ground subnet [32],

$$D_g = \frac{1}{\gamma} \sum_{l=1}^{L} \lambda_{gdl} T_l, \qquad T_l = \frac{1}{\mu_d C_{gdl} - \lambda_{gdl}} \tag{2.1}$$

where T_l is the delay on link l, $\gamma = \sum_{i,j=1}^{N} \gamma_{ij}$ = the total data traffic rate in the data subnetwork, λ_{gdl} = the traffic rate on link l, and $\frac{1}{\mu_d}$ = the average length of a packet.

2.1.2 Data delay in the ground subnet: $/M/D/1$ model

We know that the independence assumptions made in the previous section are not true in general, but they are reasonably good approximations for such systems (see p. 165, [9]). In many networks, the assumption of exponentially distributed packet length is not appropriate, for instance, in data networks where packets are of fixed length. Because packets in the satellite network are of fixed length, we may assume that data packets in the terrestrial network are of fixed length, too. In such a case, we replace the $M/M/1$ model by the $M/D/1$ model. The average delay T_l on link l is

$$T_l = \frac{1}{\mu_f C_{dl}} + \frac{\lambda_{dl}}{2\mu_f C_{dl}(\mu_f C_{dl} - \lambda_{dl})} \tag{2.2}$$

The average data delay D_g in the ground subnet remains the same formulation as in (2.1). We can either use the $/M/M/1$ model or the $/M/D/1$ model, however, the characteristics and the nature of the approach and the results are basically the same.

2.1.3 Voice blocking in the ground subnet

Suppose that we are given a telephone subnetwork in the ground which may use the same transmission links and switching facilities as the data subnetwork in the ground. This voice subnetwork has N_1 nodes linked by L_1 trunks (links) of capacities $C_{gvl}, l = 1, 2, ..., L_1$ in a specified topology. The capacity C_{gvl} of link i can be divided into N_{gvl} channels. A traffic rate matrix $[, _{ij}]$ specifies, in calls/min, the average rates of call requests between all possible IMP pairs i and j, where $i, j = 1, 2, ..., N_1$. We can model each trunk by an $M/M/N_{gvl}/N_{gvl}$ system and the average blocking probability P_b is [9]

$$P_b = \frac{1}{,} \sum_{l=1}^{L_1} \lambda_{gvl} P_l, \qquad P_l = \frac{(\lambda_{gvl}/\mu_v)^{N_{gvl}}/N_{gvl}!}{\sum_{n=0}^{N_{gvl}} (\lambda_{gvl}/\mu_v)^n/n!} \tag{2.3}$$

where P_l is the blocking probability of trunk l, $, = \sum_{i,j}^{N_1} , _{ij}$ = the total voice traffic in the voice subnetwork, λ_{gvl} = the traffic rate on link i, and $\frac{1}{\mu_v}$ = the average holding time of a phone call.

2.1.4 Satellite channel

The performance analysis of the satellite subnet depends on the multi-access protocol used in the satellite channel. The common multi-access (MA) schemes are frequency division MA (FDMA), time division MA (TDMA), and code division MA (CDMA). A complete discussion of the pros and cons of these schemes and the detailed performance analysis of such systems are not our focus (see [8, 29, 55, 60]). Rather, we would like to obtain the performance measures for voice and data traffic in the satellite subnet.

The protocols used can be further classified as random access, demand as-

signment, or a hybrid of the above two schemes. The random access protocols require no reservation operations and channels, also known as ALOHA schemes (slotted or pure ALOHA) [9]. In contrast, the users must use a reservation channel to request a transmission channel for the demand assignment schemes, which has a better throughput than the ALOHA system when the traffic is high. The maximum throughput of the slotted ALOHA is 0.368 [1], and this scheme has a shorter delay than demand assignment schemes when the traffic is light. The hybrid scheme, the combined scheme of these two protocols, behaves like a random access scheme when the traffic is light, and like a reservation scheme when the traffic is heavy [60].

Voice traffic must be transmitted in a stream, and thus the whole channel must be established using a reservation scheme before the transmission starts, while the data traffic can be adapted to these three different protocols.

To integrate voice and data traffic in the satellite channel, we can have two strategies: a fixed boundary strategy or a movable boundary strategy. In the fixed boundary strategy, the data packets are not allowed to use the voice channels, even if some of them are idle. In the movable boundary strategy, the data packets can occupy any of the voice channels not currently in use. However, an arriving call has higher priority to preempt the data packets serviced in the voice channels.

We make the following assumptions in our model:

1. The SIMPs collectively generate Poisson data traffic at rate Λ_d packets/sec and Poisson voice traffic at rate Λ_v calls/min (excluding the retransmission due to collision). The overall transmission rates (the original rate plus retransmission rate) of data and voice traffic into the satellite channel are

denoted as Λ'_d and Λ'_v respectively.

2. The data packets are of fixed length. The voice call duration is exponentially distributed with mean $1/\mu_v$ seconds.

3. Channels are slotted. Let T denote the slot length which equals the transmission time of a packet and S be the round-trip delay of the channel measured in slots.

4. The retransmission delay for a request or a random access data packet is uniformly distributed between 0 and K slots.

Data delay in the satellite channel

Suppose that the satellite channel has capacity C_s which is divided into two parts: C_{sd} for the data and C_{sv} for the voice. There are R reservation channels and N_s message channels which are further divided as N_v voice channels and N_d data channels. The word "message" here refers to either voice calls or data packets. All channels are slotted. The length of a slot time is equal to the transmission time of a data packet. A time slot is further divided into n minislots for message reservation and the length of each minislot (T/n) is equal to the transmission of a request packet. There are N_s $(= nR)$ minislots in the R reservation channels. Among the N_s minislots, the first N_v are used for voice requests and the other $(N_s - N_d)$ are used for data requests; see Fig. 2.1.4.

Define p_{suc} to be the probability that a data packet will be successful on a data channel. Then p_{suc} is $\frac{\Lambda'_d}{m} T e^{-\frac{\Lambda'_d}{m}T}$ (see [51] page 430). Here we assume the data packets can go to any of the m data channels and the service at each channel is independent of other channels' services. The throughput of the m (could be

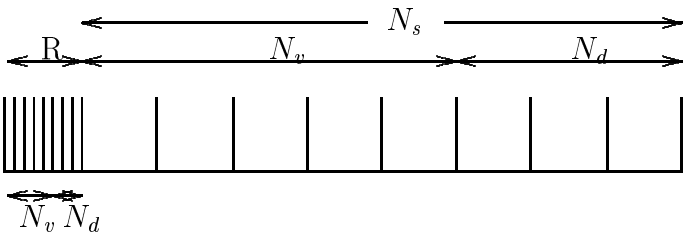

Figure 2.2: The satellite channel assignment

N_d in the fixed boundary scheme or a variable in the movable boundary scheme) data channels system η_{RAD} is $m \times p_{suc}$. The average delay under a random access data (RAD) protocol is [60]:

$$D_{RAD}(m) = [1.5 + S + (e^{\frac{\Lambda'_d}{m}T} - 1)t_{rx}]T \tag{2.4}$$

where $e^{\frac{\Lambda'_d}{m}T} - 1$ is the average number of retransmissions required for the data packet, $t_{rx} = 1.5 + S + \frac{K-1}{2}$ is the average retransmission time measured in slots, and $\Lambda_d = \Lambda'_d e^{-\Lambda'_d T}$ [51]. According to the same reference, the queueing delay at each SIMP is neglected. Since the throughput of such a system is only a low portion of channel capacity, the probability that a message might have to wait for transmission would be small.

For the demand assignment protocol, m is the number of data request minislots available for requesting a data channel in the system. Define p_{req} to be the probability that a data request packet will be successful on a minislot whose length is $\frac{T}{n}$. Then, $p_{req} = \frac{\Lambda'_d}{m}\frac{T}{n} \times e^{-\frac{\Lambda'_d}{m}\frac{T}{n}}$, and the throughput of the reservation channel η_{req} is $m \times p_{req}$. The delay of a successful request is $t_{req} = [1.5 + nS + (e^{\frac{\Lambda'_d}{m}\frac{T}{n}} - 1)t_{rx}]T/n$. The average delay for a data packet transmission is the sum of the request packet delay t_{req}, the queueing delay t_q and

the propagation delay S.

$$D_{DAD} = [\frac{1.5}{n} + 2S + \frac{\Pi'(1)}{\Lambda'_d \frac{T}{n} e^{-\frac{\Lambda'_d}{m}\frac{T}{n}}} + (e^{\frac{\Lambda'_d}{m}\frac{T}{n}} - 1)t_{rx}]T \tag{2.5}$$

where $\Pi'(1)$ is the average queue length [12], and $t_q = \Pi'(1)/\eta_{req}$, according to Little's formula which is also negligible due to the same reason.

In this scheme, we note that the delay is at least "two hops" (one hop is a round trip from an earth station to the satellite, and back to the earth station). Therefore, it is not suitable for real time traffic.

Voice blocking in the satellite channel

For a voice channel, this is an $M/G/N_v/N_v$ system, and the average blocking probability P_s is given by

$$P_s = \frac{\Lambda_v}{,} P_v, \qquad P_v = \frac{(\Lambda_v/t_v)^{N_v}/N_v!}{\sum_{k=0}^{N_v} (\Lambda_v/t_v)^k/k!} \tag{2.6}$$

where P_v is the blocking probability of the satellite channel, $\Lambda_v = \sum_{\sigma=1}^{M} \Lambda_{v\sigma}$ is the overall voice traffic rate into the satellite channel, $\Lambda_{v\sigma}$ = the voice call arrival rate from SIMP σ into the satellite channel, and $1/t_v$ is the average call duration plus the round-trip delay ST plus the call request and set-up time.

Since the typical call duration is much longer than the round-trip delay and call set-up time, we can further simplify the system to an $M/M/N_v/N_v$ queue and the probability that a system with N_v channels has n active voice calls is given by

$$\pi_v(n) = \frac{(\frac{\Lambda_v}{\mu_v})^n/n!}{\sum_{k=0}^{N_v} (\frac{\Lambda_v}{\mu_v})^k/k!} \tag{2.7}$$

Fixed boundary scheme

Under the fixed boundary strategy, the data packets are not allowed to use the voice channel. The transmissions of voice calls and data packets do not affect each other. Thus, the performance analysis is the same as shown above in this section.

Movable boundary integrated protocol

The data packets can use the idle voice channels in this strategy. To simplify the calculations, we can assume that the data queues reach their stationary state when k $(0 \leq k \leq N_v)$ voice calls are active. This is reasonable because the average call duration is much longer than call request and set-up time, which includes propagation delay and random retransmission delay in the satellite channel. The average packet delay is

$$D_{MB} = \sum_{k=0}^{N_v} \pi_v(k) d_{data}(N_v - k) \tag{2.8}$$

where d_{data} is obtained from either equation (2.4), or (2.5), and N_v could be as large as N_s which is a constant. In this case, all the channels are used by the voice traffic, and data traffic uses the channels not occupied.

Average delay in the satellite channel

The overall average delay D_s in the satellite channel is

$$D_s = \frac{\Lambda_d}{\gamma} D_d \tag{2.9}$$

where D_d can be D_{RAD}, D_{DAD}, D_{MB}, or other analytical expressions of other data access schemes, $\Lambda d = \sum_{\sigma=1}^{M} \Lambda_{d\sigma}$, and $\Lambda_{d\sigma}$ = the data arrival rate from SIMP

σ into the satellite channel.

2.2 Problem formulation

The overall objective function which we want to minimize is thus given as

$$f = A \times D_g + B \times P_b + C \times D_s + D \times P_s$$

where $A,\ B,\ C,\ D$ are the weighting factors which can be adjusted according to different network topologies. For example, $A = 0$ ($B = 0$) means that the ground subnetwork doesn't have any data (voice) traffic, or the delay (blocking probability) in the ground subnetwork is not an important factor in our consideration. Therefore, our formulation is more general than previous work. Huynh et al. considered only data traffic in packet-switched networks, in which the problem is a special case of ours by setting $A = C = 1, B = D = 0$.

To summarize, denoting $x = \{(g_{ij}, s_{ij}), \forall i, j\}$, the design of the routing problem can be formulated as follows:

$$\min_x \quad f(x)$$

subject to

$$0 \leq g_{ij}, s_{ij} \leq 1,\ \lambda_{gdl} \leq C_{gdl},\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_d \leq C_{sd},\ \Lambda_v \leq C_{sv},\ \forall l \qquad (2.10)$$

This means that the control variables are in the range of $[0, 1]$ and arrival functions of the variables must satisfy the capacity constraints so that the arrival rate on a link will be less than its capacity.

The arrival rates $(\lambda_{gdl}, \lambda_{gvl})$ of the ground links and $(\Lambda_d, \Lambda v)$ of the satellite channel are functions of $\gamma_{ij},\ \gamma_{ij},\ g_{ij},$ and s_{ij}. λ_{gdl} (λ_{gvl}) are related linearly with g_{ij} (s_{ij}) by the following expression:

$$\lambda_{gdl} = \sum_{ij} \gamma_{ij} [g_{ij} \sum_{k=1}^{n_{ij}} p_{ij}^k (\delta_{ij}^k)_l + \overline{g_{ij}} (\sum_{k=1}^{n_{i\sigma}} p_{i\sigma}^k (\delta_{i\sigma}^k)_l + \sum_{k=1}^{n_{\sigma' j}} p_{\sigma' j}^k (\delta_{\sigma' j}^k)_l)] \tag{2.11}$$

where n_{ij} ($n_{i\sigma}$) is the number of paths between source i (i) and destination j (σ), σ (σ') is the SIMP of the region in which IMP i (IMP j) is located, p_{ij}^k ($p_{i\sigma}^k$) is the probability (or fraction) of the SD (Source-Destination) pair ij ($i\sigma$) traffic that will be assigned to ij's ($i\sigma$'s) kth route, and δ_{ij}^k ($\delta_{i\sigma}^k$) is an indicator function with the value of 1 if the kth route of the SD pair ij ($i\sigma$) passes through link l, and 0 otherwise. λ_{gvl} has the same expression with γ_{ij} and g_{ij} replaced by , $_{ij}$ and s_{ij}.

The derivation of this equation is as follows: for IMPs i, j in the same region, $g_{ij} = 1$, i.e., all the traffic goes through the ground network; for i and j in different regions, $g_{ij} \in [0, 1]$, is the control variable to be decided. We assume that every node can be reached from every other node in the network. A fraction g_{ij} of traffic requirement goes through the ground network, while the rest $(1-g_{ij})$ of the traffic requirement must go to the regional SIMP before being sent to the satellite channel, and from the SIMP to the destination j.

For a satellite channel, the total arrival rate for Λ_d (Λ_v) is the sum of the arrival rates $\Lambda_{d\sigma}$ ($\Lambda_{v\sigma}$) from SIMP σ of the M SIMPs. $\Lambda_{d\sigma}$ ($\Lambda_{v\sigma}$) is a linear function of g_{ij} (s_{ij}) as

$$\Lambda_{d\sigma} = \sum_{i \in R_\sigma} \sum_{j \in \overline{R_\sigma}} \overline{g_{ij}} \gamma_{ij} \tag{2.12}$$

where R_σ is the region containing SIMP σ, $\overline{R_\sigma}$ is the region not containing SIMP σ, and $\overline{g_{ij}} = 1 - g_{ij}$. The expression for $\Lambda_{v\sigma}$ is the same with γ_{ij} and g_{ij} replaced by , $_{ij}$ and s_{ij}.

2.3 Numerical examples

The optimization problem can be solved by different algorithms. Among those are primal-dual methods [44, 50], and Sequential Quadratic Programming (SQP) [56]. We have used the FSQP (Feasible SQP) subroutines developed by J. Zhou and A. Tits at the University of Maryland [64]. This program has been proved to be powerful and fast for our problems. For a combined voice and data network with 8 nodes, 20 links and 64 control variables, we can obtain a solution in less than 3 minutes. Examples 1 and 2 whose problem sizes are halved can be solved in a minute. To guarantee the solution, we must investigate the convexity of the objective function. The Erlang B formula is not convex [26], therefore we may get a local minimum by using FSQP. The delay in a single ground link is convex in the range of $[0, C_{gdl}]$, where C_{gdl} is the capacity of link l. The summation of such function may result in non-convex function. Nonetheless, in our example, we assume that each ground link has the same capacity. The summation of these delays will be convex in the same capacity range. The complexity of FSQP algorithm depends on the complexity of the objective function. For our problem formulation, we will have linear running time on the average. Therefore, we can extend our example to a large network.

[Example 1:] First of all, we consider a network with data traffic only. This example is taken from [28]. This network has eight nodes (IMPs) and 20 links as shown in Fig. 2.3. In this network, there are two regions consisting of nodes {1, 2, 3, 4} and {5, 6, 7, 8} respectively; and the regional SIMPs are located at nodes 1 and 7, which are also IMPs. The traffic demand matrix is

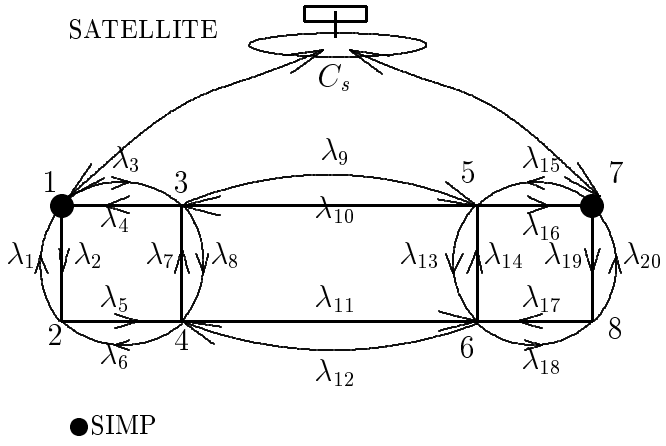

Figure 2.3: A mixed-media network

assumed to be uniform with $\gamma_{ij} = 12$ (packets/sec) [1] for all $i \neq j$ and $\gamma_{ii} = 0$ for all $i = 1, 2, ..., 8$. The average packet length is assumed to be 512 bits on all ground channels. The packet length on the satellite channel is fixed and equals 1 kbits. The ground link capacities (C_l) are all assumed to be 50 kbits/sec (5×10^4 bits/sec), and the satellite capacity to be $C_s = 1.5 \times 10^6$ bits/sec.

The ground routing we used in this example is the split traffic routing (or alternate routing) which is based on the minimum number of hops required to transmit packets from a given source node to a destination node. For example, if we want to send packets from IMP 1 to IMP 8 via the ground node, the minimum number of hops between IMPs 1 and 8 is four, and there are four alternate paths of our hops: they are path (1) $1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8$, path (2) $1 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 8$, path (3) $1 \rightarrow 3 \rightarrow 4 \rightarrow 6 \rightarrow 8$, and path (4) $1 \rightarrow 2 \rightarrow 4 \rightarrow 6 \rightarrow 8$.

[1] Huynh et al. used 20 packets/sec, which is too large and would be over the capacities of the some ground links when summed by eq. 2.11.

At any node along the paths selected above, if there are two links of the selected paths emanating from the node, then the traffic rate is bifurcated equally on each of these two links. For instance, the traffic coming into IMP 3 will be split equally into links l_8 and l_9. Using this ground sun-network routing algorithm, we find that the traffic assignment between IMPs 1 and 8 is 1/8 of total traffic $\gamma_{1,8}$ over path (1), 1/8 over path (2), 1/4 over path (3), and 1/2 over path (4). After obtaining these split fractions for each node pair and adding them up using equation 2.11, we obtain the link traffic rate λ_{gdl} for each $l = 1, 2, ..., 20$.

The routing indexes g_{ij} are given in Table 2.1 after running the FSQP. The overall objective function is the combination of average delays in ground links and the satellite link. Due to the symmetry of the given network topology and the uniformity of the traffic requirements among all SD pairs, we will expect the results will demonstrate symmetry, too. However, the results in the Huynh's original paper didn't show any symmetry. This is why we repeat the procedure. The correctness of numerical data is important, because with further experiments, we can obtain more insights from the correct results. For instance, if we consider the delay as a function of requirement traffic matrix γ_{ij} only, we have the following observations.

[**Observation 1:**] The delay is a convex function of γ_{ij}. When the rate is small, all the traffic goes through ground links, since the delay of the ground subnetwork is small. As the rate increases, the ground links become loaded, and eventually saturated. When the arrival rates approach link capacities, more and more traffic will go through satellite link. The relation between γ_{ij} and g_{ij} is shown in Fig. 2.4. The result shows that the SD pairs with more number of hops (links) will deviate their traffic to satellite first (e.g. SD pair 1, 8), and then the

Table 2.1: Splitting ratio g_{ij} for a data network

		destination			
		5	6	7	8
origin	1	1	1	0.482	0
	2	1	1	0	0.677
	3	1	1	1	1
	4	1	1	1	1
		objective=0.0842 sec.			

SD pair with fewer number of hops (SD pair 1,7, then pair 2,7), etc. This is the so-called "farthest end routing" in [48]. The last to sent their data through the satellite are the nodes one-hop away (for example, SD pair 3,5 and 4,6). This observation is also true for other network topologies.

[**Observation 2:**] The symmetry of our example can be represented in equivalence classes. For example, nodes {1, 2, 7, 8} and {3, 4, 5, 6} are equivalence classes. The traffic from IMP 8 to IMP 1 is the same as that of from IMP 7 and IMP 2. Furthermore, $g_{1,8}$ will be the same as $g_{8,1}$. Symmetric results like this can be used to check the correctness of the programming. Of course, just symmetry is not sufficient to guarantee the correctness of the results. This observation is only true when applied to the networks with symmetric topologies and uniform traffic requirement rates among nodes.

[**Example 2:**] In this example, we consider a network with voice traffic only, the objective function is the sum of blocking probabilities of ground links and satellite links. The network is the same as that in Example 1. However,

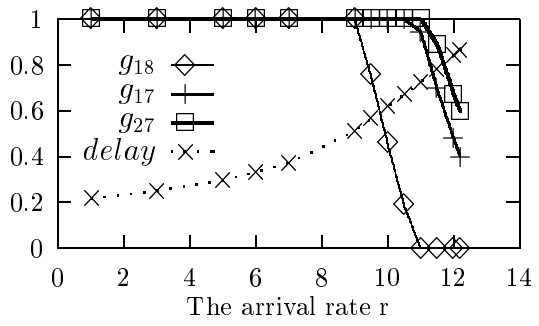

Figure 2.4: The effects of increasing arrival rate r_{ij}

the capacities are converted into channels for voice transmission. The traffic requirement matrix is again assumed to be uniformly 1 calls/min, the average call duration is 4 minutes per call, the satellite has 50 channels, and the ground links have 5 channels per link. The voice splitting ratios s_{ij} are obtained as in Table 2.2.

From Fig. 2.5, we note that the difference of voice traffic and data traffic is that some voice traffic goes to satellite channel even at low arrival rates, for instance, traffic between two inter-region nodes such as that between node 2 and 8, etc. This is because the satellite has much larger capacity than any of the ground links, and delay of satellite is not a big factor to consider. If both the source node and the destination node are in the same region, the traffic will only use the ground links. If they are not in the same region, some traffic will go to the satellite link. For the traffic from nodes located between two regions to the nodes in the other region (also located between two regions), it only uses the ground links.

Table 2.2: Splitting ratio s_{ij} for a voice network

		destination			
		5	6	7	8
origin	1	0	0	0	0
	2	0.559	1	0	0
	3	1	1	0.177	0.322
	4	1	1	0	1
		objective=0.98			

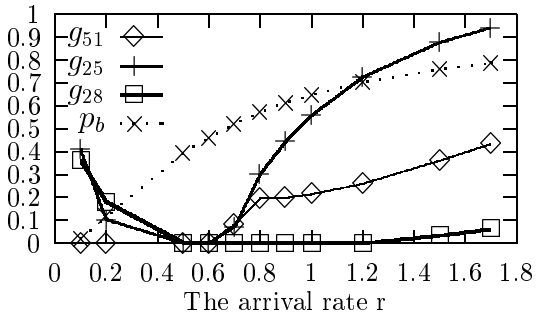

Figure 2.5: The effects of increasing arrival rate r_{ij}

Since the objective function is not convex (following a shape of the Erlang B formula, which can be obtained by running Mathematica), the result shows that splitting ratios are not monotone functions of arrival rate as in the data case; see Fig. 2.4 and Fig. 2.5

[**Example 3:**] In this example, we consider a ground subnetwork capable of transmitting and switching both voice and data traffic through the same IMPs and transmission links. Using the same network and the same parameters as in the previous examples. For fixed boundary scheme, we would expect the splitting ratios of voice and data will be the same as those of Example 2 and 1 respectively. This is because the voice traffic and the data traffic are independent of each other. For movable boundary, we assume the following parameters of the system: the number of voice channels in the satellite is 20, the number of data channels is 10, and the ground capacity in each link is 5 channels/link. The traffic requirement for data is uniformly 6 packet/sec, and that for data is 1 call/min. The delay expression used for satellite channel is D_{RAD}. The overall objective function is the (unweighted) sum of delays and blocking probabilities in the ground links and the satellite channel. Minimizing this objective in this example is valid because both values are less than 1. We have programmed the results in Table. 2.3 and Table 2.4. In case the delay is much larger than 1, we may have to weight between these individual objectives. Or we can use objective functions that have the same units, for instance, the number of customers requesting data services in the system, and the number of voice calls being blocked. The other approach for optimizing the multiple objectives is to minimize the delay (blocking probability) under the blocking probability (delay) constraints, which will be explored in next Chapter.

Table 2.3: Splitting ratio g_{ij} for the data subnetwork

		destination			
		5	6	7	8
origin	1	1	1	0.165	0
	2	1	1	0	0.487
	3	1	1	1	1
	4	1	1	1	1

Table 2.4: Splitting ratio s_{ij} for the voice subnetwork

		destination			
		5	6	7	8
origin	1	0		0	0
	2	0.257	1	0	0
	3	1	1	0.176	0
	4	1	1	0	1
		objective=0.191			

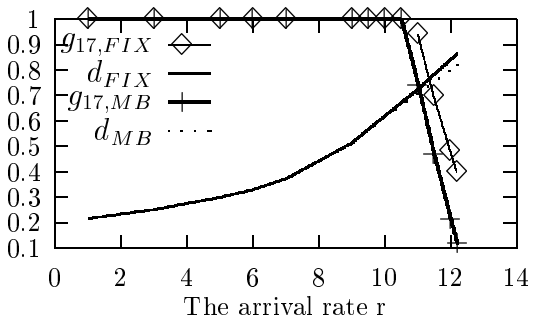

Figure 2.6: The effects of movable boundary scheme

We note the results are quite similar to the previous examples. The comments in each example apply here. From the illustration of Fig. 2.6, we confirm that the delay of the movable boundary scheme is lower than that of fixed boundary scheme, when the arrival rate is large enough so that some traffic goes to the satellite channel. We also find that more traffic will go to the satellite channel under the movable boundary scheme. This is because more data channels are available in the satellite channel under the movable boundary scheme. Therefore, the moving boundary can increase the utilization.

Chapter 3

Trade-off Curves

In the previous chapter, we have obtained the optimal routing ratios for mixed-media networks with integrated voice and data traffic. This was done by optimizing a weighted sum of multiple-objective function [13]. The physical meaning of the weighting coefficients is the relative importance of each objective to the overall objective. This statement will be proved rigorously. The delay could be as large as infinity, while the upper limit for the blocking probability is one. Thus these two measures have different units, and a linear combination of these two functions is likely to be biased. Therefore, we must be careful of the way these weights are found. In the following, we will develop a systematic procedure for finding the weighting coefficients.

Based on the systematic analysis, the trade-offs among individual objectives are clear. However, we would like to propose a new approach to obtain the trade-off curve of multiple objective functions. In our case, the trade-off is between optimal performance value of voice traffic and that of data traffic.

3.1 Linearization of multiple objectives

In multiple objective function optimization, linearization is desirable because we can utilize the the existing algorithm for single objective function optimization to solve the problem. One straightforward way of linearization is the weighted sum approach as shown in the previous chapter.

Suppose that the number of decision (routing) variables is n, and these are denoted by a vector $p = (p_1, ..., p_n)$, so that p is a point in $\mathcal{R}^n$, the input space. In the same way if there are m objectives $\phi = (\phi_1, ..., \phi_m)$ then ϕ is a point in $\mathcal{R}^m$, the output space. Since $\phi_i,\ i = 1, ..., m$ is a function of p, ϕ is a mapping from the input space to the output space

$$\phi : \mathcal{R}^n \Rightarrow \mathcal{R}^m \tag{3.1}$$

In general, ϕ is nonlinear. Our objective is to minimize each of the ϕ's.

Definition 1 *A point $p \in \mathcal{R}^n$ is a Pareto point for a mapping ϕ if there exists no other point $\hat{p} \in \mathcal{R}^n$ such that*

$$\phi_i(\hat{p}) \leq \phi_i(p),\ i = 1, ..., m \tag{3.2}$$

where none of the above inequalities are equalities.

Definition 2 *A point $p \in \mathcal{R}^n$ is a Pareto critical point for a mapping ϕ if no small change Δp exists, such that*

$$\phi_i(p + \Delta p) \leq \phi_i(p),\ i = 1, ..., m \tag{3.3}$$

where none of the above inequalities are equalities.

Thus the Pareto point, which is the global optimal point, is also the Pareto critical point, the local optimal point. The set of all Pareto points will be denoted by P and the set of all Pareto critical points will be denoted by P_c. The set of points $q \in \mathcal{R}^m$ such that there exists $p \in \mathcal{R}^n$ where $q = \phi(p)$ is called the realizable set of outputs and is denoted by Q. That is

$$Q \equiv \{q \in \mathcal{R}^m | \phi(p) = q\} \tag{3.4}$$

In general $Q \neq \mathcal{R}^m$, so that Q has a boundary, ∂Q. It is well known that the image of Pareto critical points reside in the boundary of Q [10].

It can be shown in general that all points of $\phi(P)$ can not be reached by forming a weighted sum of the ϕ's ([10], p. 277) unless each ϕ_i is a convex function. Then $P = P_c$ and $\phi(P)$ lie in the boundary of the convex hull of the realizable set in the output space defined by the weighted sum of ϕ's.

Another way of linearization of multi-objective optimization is the *minimax* approach which circumvents the difficulty of interpreting the meaning of weighting coefficients. The *minimax* approach is to minimize the maximum of the multiple objectives. In our case, we have a nonlinear programming problem due to nonlinear constraints as following

$$\min_x \max_k \quad f_k(x)$$

subject to

$$0 \leq g_{ij}, s_{ij} \leq 1,\ \lambda_{gdl} \leq C_{gdl},\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_d \leq C_{sd},\ \Lambda_v \leq C_{sv},\ \forall l \tag{3.5}$$

where $f_k,\ k = 1, 2, 3, 4$ are the multiple objective functions to be minimized. Note that from time to time we will use f in place of the mapping ϕ and use x in place of p to denote the set of routing variables $\{g_{ij}, s_{ij},\ \forall i, j\}$.

This approach is widely adapted to numerous optimization problems. Minimax optimization is a more natural setting for obtaining Pareto points than minimizing weighted sums. The results from the minimax approach give us an indication of the best one can achieve in the worst case. In comparison, the weighted sum approach only gives the best values of the sum of objectives without information on any particular one.

It can be shown in the same reference ([10], p. 325) that any Pareto point can be reached by a minimax optimization, whereas only those Pareto points on the convex part of the boundary of the realizable set can be obtained by the weighted sum approach.

However, the weighted sum method can be transformed to achieve this apparent advantage for optimization.

Theorem 1 *Let λ_i be the Lagrange multipliers obtained in the constrained minimization*

$$\min \gamma$$

such that

$$\gamma - f_i(x) \geq 0, \ i = 1, ..., m \tag{3.6}$$

which is equivalent to $\min_x \max_i f_i$. *Let x^* be the solution to (3.6) and suppose $f(x^*) \in$ the boundary of the convex hull of the realizable set. Then*

$$\min_{\mathrm{x}} \max_{\mathrm{i}} \ f_i(x) = \sum_{i=1}^{m} \lambda_i f_i(x^*) = \min_x \sum_{i=1}^{m} \lambda_i f_i(x) \tag{3.7}$$

In other words, if x^* is reachable by a weighted sum minimization, then the Lagrange multipliers obtained from the minimax problem (3.6) are the correct weights to reach x^*

Proof: According to the Kuhn-Tucker (KT) conditions, $\gamma,\ x^*$, and λ satisfy

$$\lambda_i(\gamma^* - f_i(x^*)) = 0 \tag{3.8}$$

$$\lambda_i \geq 0 \tag{3.9}$$

$$1 - \sum_{i=1}^{m} \lambda_i = 0 \tag{3.10}$$

$$\gamma^* - f_i(x^*) \geq 0 \tag{3.11}$$

$$\sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial x}(x^*) = 0 \tag{3.12}$$

Since $\gamma^* = \min_x \max_i f_i(x)$, then the KT conditions give

$$\min_x \max_i f_i(x) = \gamma^* = \gamma^* - \sum_{i=1}^{m} \lambda_i(\gamma^* - f_i(x^*)) \tag{3.13}$$

$$= \sum_{i=1}^{m} \lambda_i f_i(x^*) \tag{3.14}$$

The condition

$$\sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial x}(x^*) = 0 \tag{3.15}$$

means that the vector λ is normal to the plane tangent to Q at $f(x^*)$. Since $f(X^*) \in$ convex hull of Q, then this plane is a support plane of Q at $f(x^*)$, i.e., $\sum_{i=1}^{m} \lambda_i f_i(x^*) \leq \sum_i \lambda_i q_i$ for all $q \in Q$. Hence $\sum_{i=1}^{m} \lambda_i f_i(x^*) \leq \sum_{i=1}^{m} \lambda_i f_i(x)$. $\square$

Therefore, these weighting coefficients are actually the Lagrange multipliers of a constrained optimization problem which is equivalent to the minimax problem. Since the Lagrange multipliers sum up to one, the weighting coefficients are the relative importance (or the percentage) of the corresponding objective in overall objectives. Therefore, in our previous setting, all those four objectives are equally important, since a scale factor applied to the objective does not affect the optimal values of the control variables.

In constrained optimization, the Lagrange multiplier can be interpreted as the sensitivities of the optimal value to variation of the components of the constraints. As in ([56], P. 71), the Lagrange multiplier λ is interpreted as the equilibrium price in the sense that if we relax the constraints at the price P per unit, then the optimal value will be gained by an amount of λ per unit. If $P \leq \lambda$, then it is an earning to relax the constraint. If $P \geq \lambda$, then we can gain by tightening the constraints. If $P = \lambda$, neither relaxation nor tightening yields any gain. Hence λ is called the equilibrium price. In our constrained optimization, the constraints are actually the component objectives. Thus the Lagrange multiplier in our case can be interpreted as the sensitivities of the optimal values to variation of the components of the objectives.

3.2 Trade-off curve

The image of the Pareto points, $\phi(P)$, is the optimal trade-off curve, and in general is the optimal trade-off surface. It represents points of optimum system response. Typically it is hard for the designer to specify the relative importance of each objective until he knows what are the best capabilities for the system. After the designer obtains the system capabilities using some ad hoc way of assigning weighting coefficients, he/she may want to adjust the weights. Thus the design process is both interactive and iterative.

The trade-off among different objectives can be seen clearly from the weighted sum approach. Suppose a weighted sum is given as

$$\Phi(p) = \sum_{i=1}^{m} w_i \phi_i \tag{3.16}$$

where w_i's are the relative importance of minimizing ϕ_i.

From the optimality condition (first order), $p \in P_c$ if and only if there exists a set of weights $w = (w_1, ..., w_m),\ w_i > 0$, such that

$$\sum_{i=1}^{m} w_i \frac{\partial \phi_i}{\partial p} = 0, \tag{3.17}$$

i.e., gradient of Φ is zero. For simplicity, consider two objectives, then the above condition gives

$$w_1 \frac{\partial \phi_1}{\partial p} = -w_2 \frac{\partial \phi_2}{\partial p} \tag{3.18}$$

This shows that the gradients which are the deepest decent directions of the two objective are in exact conflict. The direction which one objective increases most is the direction that the other decreases most. From the same equation, we also obtain that the increase of one objective will result a decrease of the other objective. This is the trade-off curve of the optimal image of the Pareto points. The above arguments hold true for a general multiple objective optimization.

The trade-off curve is desirable since we would like to know the best capabilities of the system, but it is usually not analytically available. Some approximation to the optimal trade-off curve can be used if $\phi(p)$ is convex ([10], p. 27). In the following, we will give a procedure to find out the exact trade-off curve in some special cases. To be specific, this is when we have convex objective functions where the global optimal is guaranteed. Otherwise, we might end up with local optimal.

For example, in a two-objective optimization, the procedure is simply to formulate the problem so as to optimize one objective under an additional constraint of the other objective and tighten this additional constraints to find out how the optimal value of the objective is affected. We can plot the result as our trade-off curve.

Since there are two ways to choose one function of the two objectives as our objective and the other as the constraint, there two ways to obtain the trade-off curve. It is true that these two ways will yield the same trade-off curve if we fix the constrained variables in the same range as that of optimal values obtained when the constraint is taken as objective in the other formulation. However, we can have different interpretations of these two formulations.

In some cases, the designer of the system may not want to optimize at the same time both data delay and voice blocking probability. He may just want to optimize the voice blocking probability or residual capacity without sacrificing the data traffic too much. Or the data traffic may have to meet some real time constraints. In these cases, it is more appropriate to optimize the voice traffic's objective under the constraints on data traffic links' delays. Using the same arguments, we may want to guarantee some link or trunk performance of voice traffic while optimizing the data traffic delay. In this case, we must use a single objective (average network delay) and the constraints are that the average voice blocking probabilities are less than some specified values. Since we can specify any link delay or trunk blocking probability, we have more detailed control over the performance of the networks. In either case, we have a non-linear programming problem with non-linear constraints.

In general, if we have more than two objective functions, the proposed procedure will work. One way is to group them as we have done, i.e., group delay of terrestrial networks and the satellite network together, etc. The general procedure is: (1) pick one objective to be optimized; the other functions are constraints,(2) tighten one constraint at one time, (3) record and plot the results. Graphics of more than 3-dimension is hard to imagine, but conceptually the step

(3) is useful if we want to find out the variation locally.

3.3 Problem formulation

3.3.1 Performance constraints

For performance constraints of mixed-media networks, we are given as constraints the blocking probabilities of all links in order to minimize the average delay. The overall objective function which we want to minimize is thus given as

$$\min_x \quad f(x) = D_g + D_s$$

subject to capacity constraints and

$$0 \leq g_{ij}, s_{ij} \leq 1, \quad P_l \leq c_l, \forall l, P_v \leq c_s \tag{3.19}$$

where P_l is the blocking probability of link l, c_l is the constraint of link l, P_v is the blocking probability of the satellite link, and c_s is its upper bound.

3.3.2 Real time constraints

For mixed-media networks with real time delay constraints, we must optimize the voice objective under the real time deadline given to the system. The overall objective function which we want to minimize is thus given as

$$\min_x \quad f(x) = P_b + P_s$$

subject to capacity constraints and

$$0 \leq g_{ij}, s_{ij} \leq 1, T_l, \leq d_l, \forall l, D_d \leq d_s \tag{3.20}$$

where T_l is the delay of link l, d_l is the constraint of link l, D_d is the delay of the satellite link, and d_s is its upper bound.

3.4 Numerical examples

Again, the optimization problem can be solved by different algorithms. Among those, primal-dual methods [44, 50], and Sequential Quadratic Programming (SQP) [64]. We have used the Feasible SQP (FSQP) subroutines developed by J. Zhou and A. Tits at the University of Maryland [64].

The following paragraphs are meant to find the interaction between voice and data traffic under various traffic conditions. We will consider movable boundary scheme only, since there is no coupling between voice and data in the fixed boundary scheme. In another words, the objective function and the constraints are independent in the fixed boundary scheme and we will not find any tradeoff in such scheme.

3.4.1 Performance constraints

In this section we will investigate the optimization of the network system delay under the constraints on the link blocking probabilities.

[**Example 1:**] First of all, we consider the same network topology and ground routing algorithm as in the previous chapter. For the movable boundary, we assume the following parameters of the system: the number of voice channels in the satellite is 25, that for data is 10, and ground capacity in each link is 5 channels/link. The traffic requirement for data is uniformly 12 packet/sec, and that for voice calls is 0.1 call/min. The other values of parameters used in this example are not given here for brevity; see Chapter 2 for details.

The routing indexes g_{ij} are given in Table 3.1 after running the FSQP. Now consider the delay as a function of requirement traffic matrix γ_{ij} only. We have

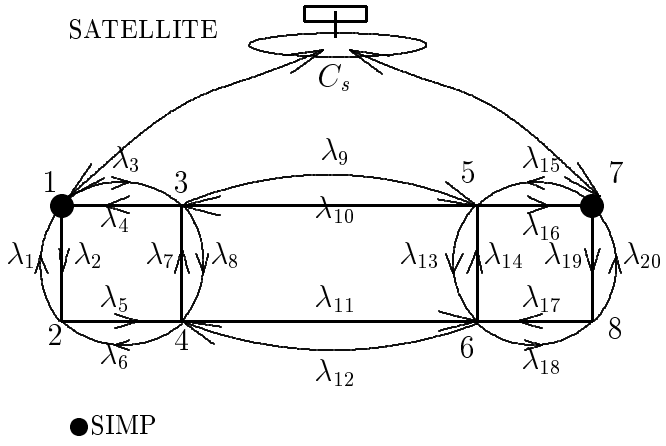

Figure 3.1: A mixed-media network

the following observations.

The effects of increasing arrival rates

To find the effects of increasing arrival rate, we first assume large constraints on all voice links. The objective function is the sum of average ground delay and satellite channel delay. The former is a convex function of γ_{ij}, the data arrival rate. The latter, as shown in Fig. 3.2, is also convex in γ_{ij}. Where x_i are the values of the splitting ratios. The linear summation of such functions is still convex in data arrival rate, as shown in Fig. 3.3. We see that as x_i grows smaller, i.e., more traffic goes to the satellite channel, the delay of the satellite channel grows larger. Therefore, the minimum delay is achieved when all data traffic and all voice traffic go through the terrestrial network.

When the rate is small, all the data traffic go through ground links, since the delay of the ground sun-network is small. At this point, the voice arrival rate has no effect on the objective function. Because the only movable boundary is

Table 3.1: Splitting ratios g_{ij} for a data sub-network

		destination			
		5	6	7	8
origin	1	1	1	0.215	0
	2	1	1	0	0.514
	3	1	1	1	1
	4	1	1	1	1
		objective=0.08 sec.			

Table 3.2: Splitting ratios g_{ij} for a voice sub-network

		destination			
		5	6	7	8
origin	1	1	1	1	1
	2	1	1	1	1
	3	1	1	1	1
	4	1	1	1	1
		objective=0.08 sec.			

the satellite channel, the voice traffic will not affect data traffic in the satellite channel when all the data traffic is on the ground. As data arrival rate increases, some data traffic will move into the satellite channel. Under these circumstances, the increase of voice arrival rate will cause an increase of satellite channel delay, and thus an increase of the average delay of the whole system. As a result, when there is data traffic in the satellite channel, the minimum is achieved when all voice traffic stays on the ground. This is demonstrated in Table 3.2 and Fig. 3.4.

As shown in Fig. 3.4, the delay of satellite channel is not a convex function in voice arrival rate, but follows the shape of Erlang B's formula [1]. As explained above, its minimum, which is a constant, is achieved when $x_i = 1$, for i in voice transmission links. The above is true independent of data arrival rate and voice arrival rate.

Compared to the weighted sum approach, some voice traffic will have to go through the satellite channel in such cases (See [13]). This is because minimum blocking probability is achieved at the points where there is some voice traffic in the satellite channel. If you put the blocking probability as a part of objective function, the FSQP will find these points as the solutions. This is found to be a major difference between these the weighted sum approach and the one proposed here.

The above statement is true for loose constraints. As constraints grow tighter, the minimum is no longer achieved at the point where all voice traffic stays on the ground, as explained in the next section.

We can also assume that some or all the ground channels utilize a movable

[1]The plot can be obtained by running Mathematica

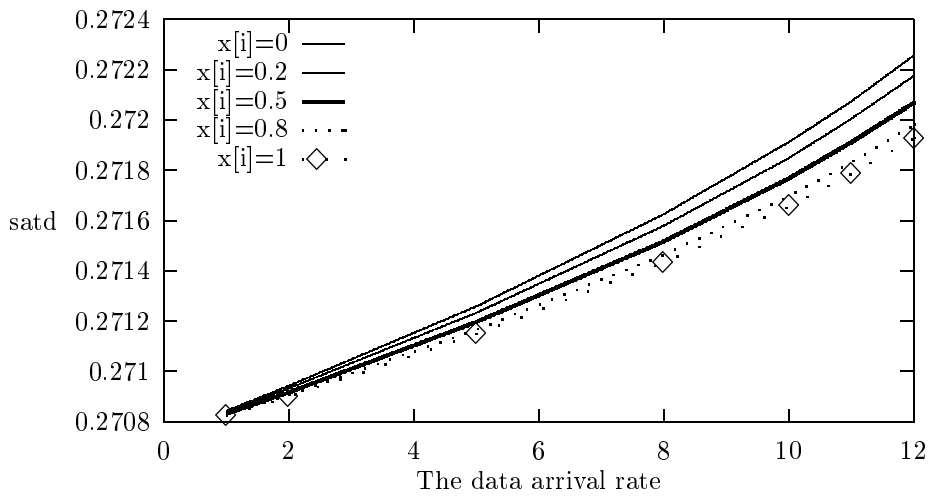

Figure 3.2: Satellite channel delay vs. data arrival rate

boundary scheme. However, it would be computationally more difficult but conceptually not much different. We believe that what we have done is the first step toward this problem and the results here give us insight to the nature of such problems.

The effects of tighter constraints

The first observation is that if we are given a big or loose constraints for the system, then it is equivalent to a system without any constraints. However, as constraints on blocking probabilities grow tighter on some particular ground links, then more voice traffic will go to the satellite channel. This is reasonable since the satellite channel (which has a larger capacity than any ground links) is used as an overflow channel. The relationship is shown in Fig. 3.5 when the constraints on link 9, 10, 11, 12, have been tighten. It is found that the

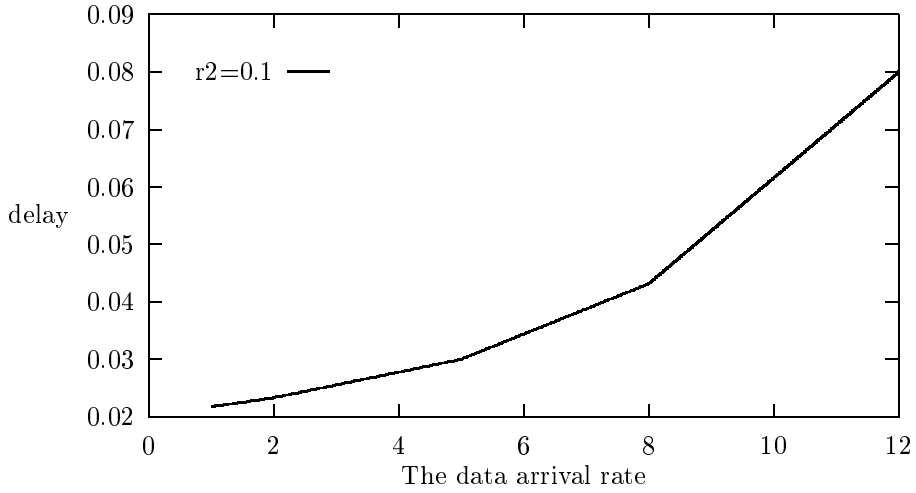

Figure 3.3: Objective (delay) as a function of data arrival rate

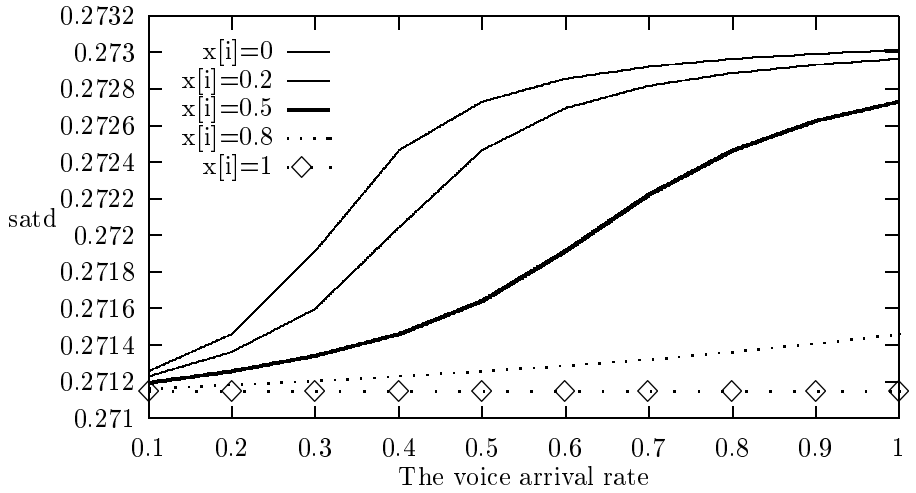

Figure 3.4: Satellite channel delay vs. voice arrival rate

tightening of constraints on voice blocking probabilities have no apparent effects on the data splitting ratios which are affected mainly by the data arrival rate.

We can plot the objective function as a function of constraints. This is the so called trade-off curve, which has the meaning that we must trade delay at the price of constraints. If we want a tighter blocking probability, then we will have a larger delay. On the other hand, if we want to have a smaller delay, then we are forced to have looser performance constraints. This relationship is shown in Fig. 3.6. We notice that the same type of curves are obtained as in Fig. 3.7 if we plot the delay in the satellite channel vs. constraints on blocking probabilities.

Under a tighter constraint, if we try to increase the voice arrival rate, then we might not be able to find a solution, since there are no feasible points in the solution space under such conditions. For a greater voice arrival rate, the links will have larger blocking probabilities. Unless we increase the links' capacities, we can not get feasible points under tighter constraints on blocking probabilities. If we increase the data arrival rate, we will have the same type of curves as before. This is because the data arrival rate affects only the objective values and the data splitting ratios and voice arrival rate affects both the blocking probabilities and the constraints.

Due to the symmetry of the network topology, we tightened the constraints uniformly. If we increase the constraints on some particular links, we may or may not get solutions. However, the rationale is that if we tighten the constraints on one link, then at the same time we have to reduce the flow on that link. Since the flow is conservative, the flow that is reduced from that link must go to the other links which achieve minimum delay and observe their own constraints.

For the relationship between tightening constraints and blocking probability

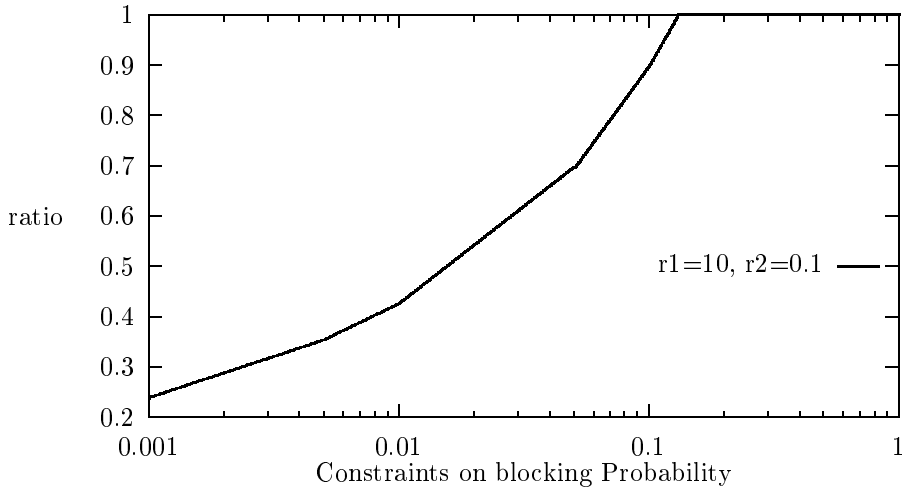

Figure 3.5: Ratios vs. tightening the constraints

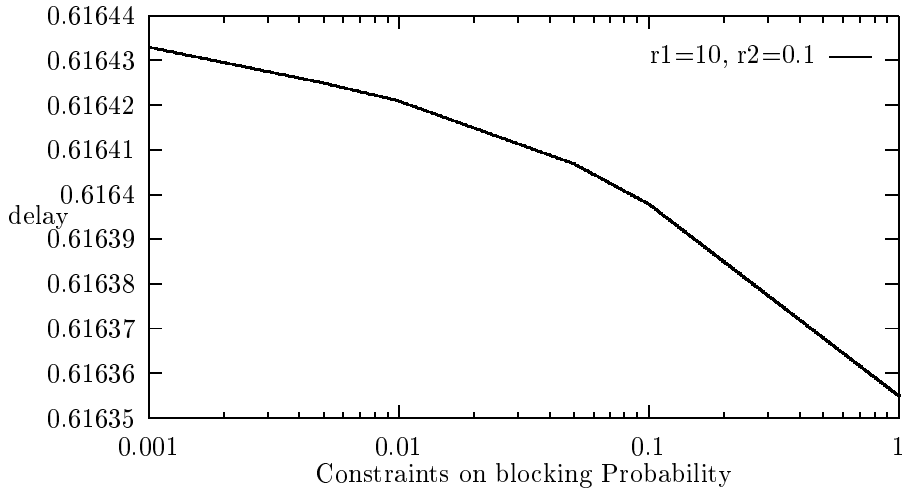

Figure 3.6: The tradeoff curve: System delay vs. tightening constraints

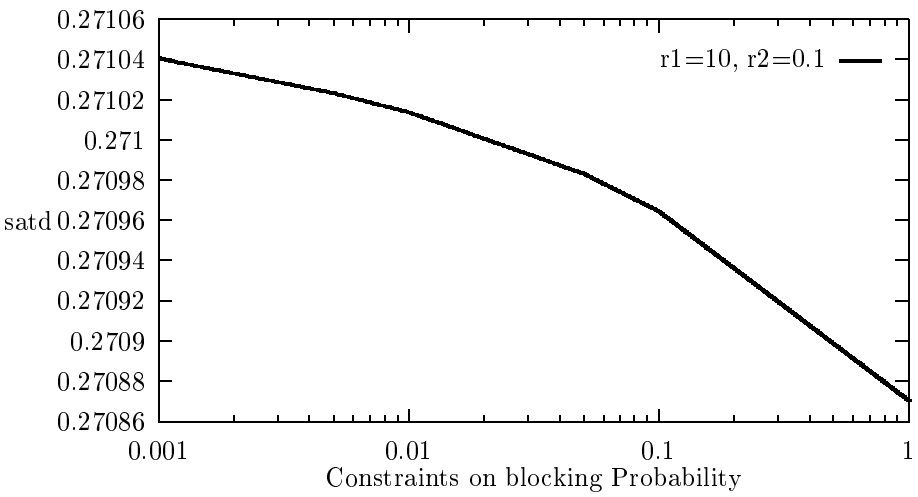

Figure 3.7: The tradeoff curve: Satellite delay vs. tightening constraints

in the satellite channel, we look at the Fig. 3.8. It shows that the tighter the constraints, the larger the blocking probability in the satellite channel. We obtained the same type of figures for blocking probabilities on the ground links. A figure for blocking probability of ground link 1 vs. constraints is shown in Fig. 3.9.

Discussion

One may argue that as gigabit networks are being deployed [35], the ground link capacities are larger than that of the satellite channel. Nonetheless, we feel that the nature of the problem will not change even thought the parameters of the above example change. We have a scenario that the ground links will saturate gradually as data arrival rate increases and will then use satellite link as a backup channel. The above is true independent of the ground links' capacities.

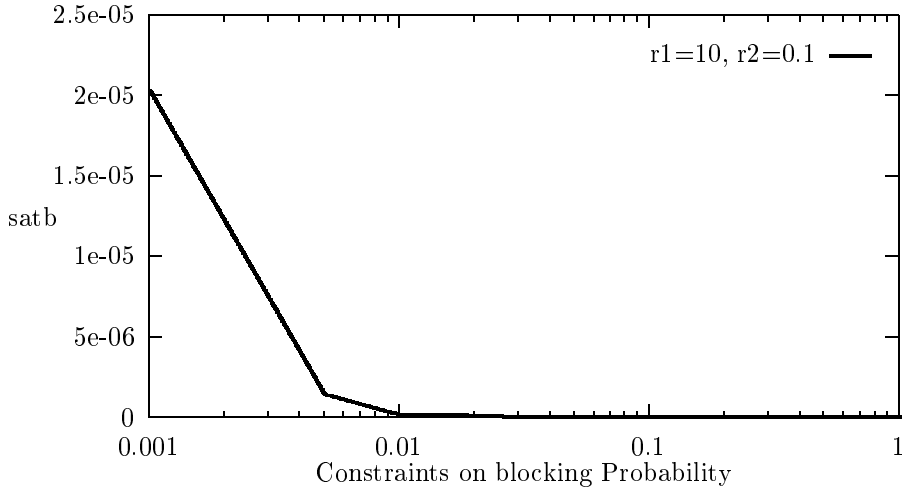

Figure 3.8: Satellite blocking vs. Constraints

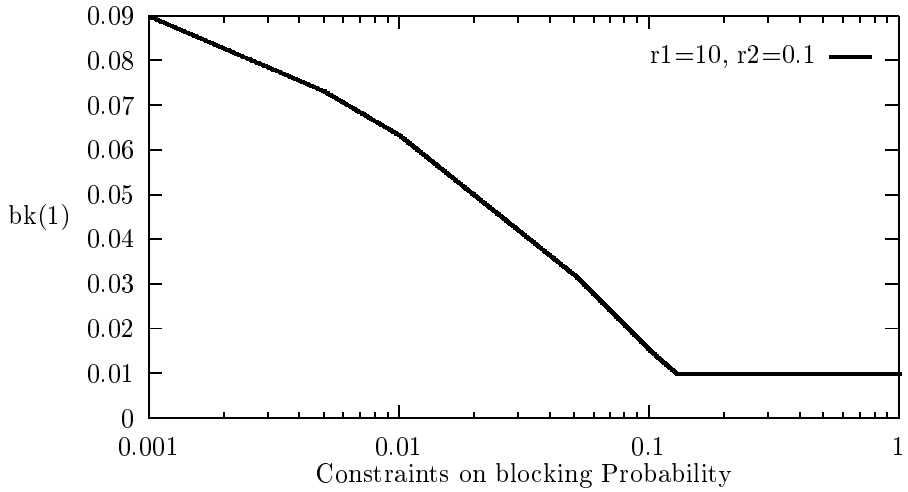

Figure 3.9: Blocking probability of link 1 vs. constraints

Table 3.3: Splitting ratios g_{ij} for a voice sub-network

		destination			
		5	6	7	8
origin	1	0	0	0	0
	2	0.41	1	0	0.36
	3	1	1	0	0.37
	4	1	1	0	1
		objective=0.033			

3.4.2 Delay constraints

In this section, we investigate the optimization of the network system blocking probability under the constraints on link delays. The optimal switching ratios for voice traffic are given in Table 3.3. We found that this result is the same as those obtained in the previous Chapter. Traffic within two hops will all go through a terrestrial network while traffic connected to SIMPs will direct all their traffic through a satellite channel.

The effects of increasing arrival rates

The first observation is that the increase of data arrival rate will not affect the voice traffic (and thus system blocking probability) at all. Since voice traffic has priority over data traffic, the voice can preempt the data traffic. Thus the amount of the data flow is inconsequential. We can plot system objective (blocking probability averaged by the links' flow) vs. voice rate in Fig. 3.10, which is the same type of graph as Fig. 3.4. These graphs are shown to follow

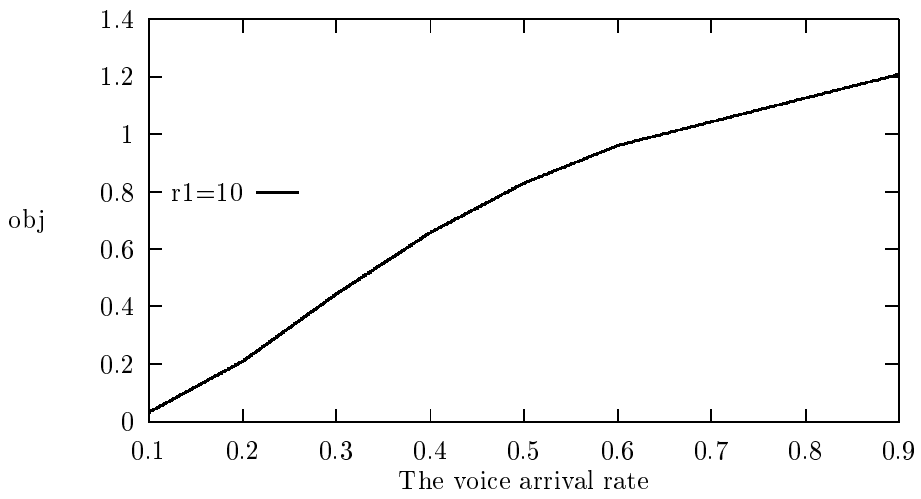

Figure 3.10: Flow averaged system blocking vs. voice arrival rate

the shape of Erlang B's formula, since either the system blocking probability or the satellite channel delay is a linear combination of Erlang B's formula (in which voice arrival rate is an independent variable).

We have also plotted the satellite channel delay as a function of voice arrival rate. The results are shown in Fig. 3.11. As the voice arrival rate increases, the splitting ratios are found to be about the same as those in Table 3.3 except the ratios will become smaller when the voice arrival rate is larger.

The effects of tighter constraints

Since only the movable boundary scheme is used in the satellite channel, we would have to tighten the constraints on the delay in the satellite link. However, due the large propagation delay in the satellite link (which is about 0.27 sec. round trip delay), it is not possible to bound the delay under 0.27 sec. On the

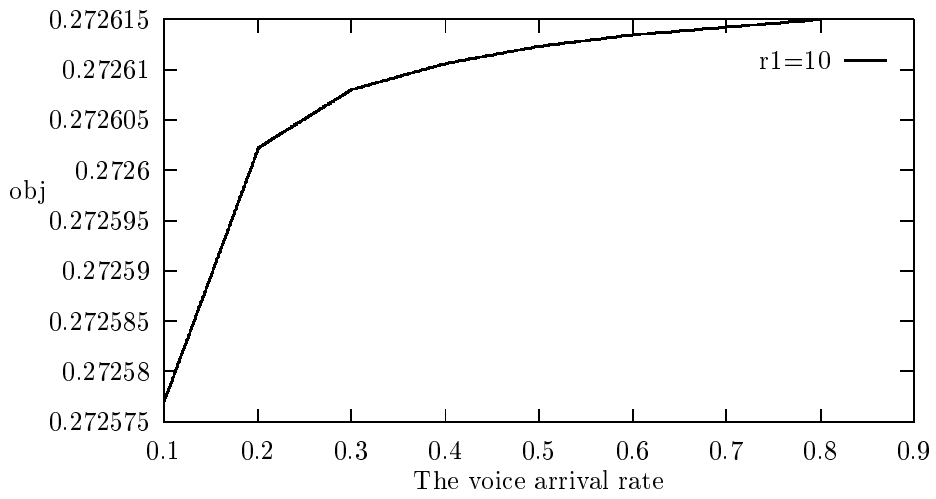

Figure 3.11: Satellite channel delay vs. voice arrival rate

other hand, a bound as 0.273 sec. is found to a upper bound of the delay. Thus, it would not be too meaningful to try to bound the delay on the magnitude of scale in such range. That is, it does not matter too much in reality if we have a delay of 0.003 sec difference in such case, which is 1.1 percent in total delay and can be neglected.

If we want to find the tradeoff curves with constraints in the range of 0.27 and 0.273 sec., we still can find the same type of curves as in Fig. 3.6. The explanation, using same reasoning, is that we must tradeoff between voice blocking probability and data delay. If we want a smaller delay, we will suffer a larger blocking probability. If we can withstand a larger delay, we can have a better performance (smaller blocking probability). The choice is up to the system's specification for different applications.

Chapter 4

Dynamic Traffic Control

In previous chapters, we have proposed several schemes to obtain the traffic routing ratios of mixed media networks. The procedures are static and the traffic parameter profiles are given in numerical calculations. In this chapter, we would like to propose a general network management and control architecture which is real-time and state-dependent so that the network control can cope with the dynamics of traffic and react to real time network traffic accordingly to avoid congestion and to achieve optimal performance. The architecture together with some adaptation of the previous optimization procedures will constitute the basis of the dynamic network traffic control.

4.1 Network control architecture

The network control architecture presented here consists of three separate but interconnected modules: Network Elements (NEs), Operating Systems (OSs), and Interconnection Networks (INs). The classification is by no means a standard nor is it an existing system. This is primarily for clarity of explanation of the concepts which describe the logical view of a network control architecture and how the control operations are executed through these modules. The functions

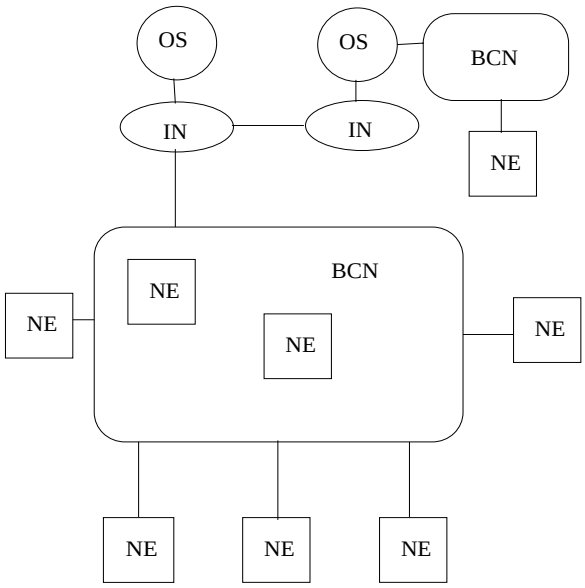

Figure 4.1: The network architecture

of individual modules will be elaborated in the following sections as well as how the control schemes incorporate these modules.

4.1.1 Network Elements

Networks Elements are communication equipment which make up the "body" of the network structure. NEs can be further classified by their functions into the following categories.

- switching elements: are the elements that are responsible for switching messages from one input to its destination output, e.g., circuit switch, packet switch, ATM switch, etc.

- transport elements: are the elements that carry messages in a non-switch way, e.g., multiplexors, digital cross-connect system, digital loop carrier system, etc.

- end equipment: are the elements that serve as interfaces between human users or operators and the network, e.g., workstations, consoles, remote terminals, etc.

- control devices: are the elements that control the call setup, teardown, routing ratios, etc.

- mediation devices: are the elements that perform functions such as protocol conversion, address mapping, concentration, message conversion, and data storage.

- sensors: are the elements that measure the traffic parameters, collect network data, etc. They are called agents in network management parlance.

- signaling information transport elements: are the elements that transmit the signaling information such as call setup request, calling tone, call teardown message, etc.

The collection of some of these NEs constitute a network node, e.g., a node may consist of switching elements, transport elements, control elements, sensors, mediation devices, etc.

On top of each element, we have computer software that make them operational. We may have to define protocols, procedures, interfaces for these NEs in order to achieve the functions we need in an integrated manner.

4.1.2 Interconnection networks

Interconnection networks are used to carry information between NEs or between NEs and OSs or between OSs. These INs could be as simple as one link between

one NE to another NE or to one node. A collection of the links, nodes, and NEs constitute the Backbone Communication Networks (BCN).

We assume our network has out-of-band signaling capability such as the Common Channel Signaling (CCS), that is, the control signals are sent through a dedicated channel or communication link. Therefore, the interconnection network between one OS and NEs or the backbone network is a (conceptually and functionally) separate network from the backbone network whose function is the transmission of network control and management. The information transmitted through this Management Network (MN) from NEs to the OS are the network traffic data, measurements, etc. The information transmitted in reverse is the control data, signaling, decision variables, etc. In a centralized control network, it requires the OS to have a link to every node in order to control its traffic.

From one OS to the other OS information can either go through a direct link between them or go through a different INs. For a distributed implementation of our algorithm, we might divide the network into several subnets, each subnet is controlled by its OS. From time to time, these OSs must exchange traffic information of its own area with the others. Thus, a direct link among OSs is more efficient.

4.1.3 Operating Systems

Operating Systems are the software embodiment of the network "brain". It's responsible for the control and management of the network so that the network can function well and meet users' needs. Besides the routines for network operations, central to the OS is a (conceptually) global database. The database stores information such as network and system configuration, current and his-

toric performance, trouble logs, security codes, accounting information, etc. This database is called the Management Information Base (MIB) in OSI (Open Systems Interconnection) terminology.

The design principle of this database to serve as a "mirror" and control panel of the network. It is designed to reflect the health status of the network and the operator can view the details of the network at any level. At the same time, we can control the network through the control the database. For example, if the network operator would like to change the routing ratios due to traffic changes or link failures, the operator merely updates the appropriate variables in the database. The actual implementation of these changes are triggered by an execution process in the database through the Management Network to the corresponding NEs.

Besides the above functional requirements, the database must be fault tolerant itself to be able to operate 24 hours on-line. We also desire the real time response of the database which enables the OS to respond to network conditions dynamically.

Similar to the design of the database proposed by MANDATE [27], the high level design of a network management system using database technology is incorporated. They presented an object oriented modeling in the design and implementation of such systems. They also advocated a data model in which the network data is classified in three categories: structural data, control data, and sensor data. The structural data refer to slow changing configurational data which outline the network structure. Sensor data are the data collected by sensors which are the information needed for control and the information about the network status. Control data are the signals generated by the OS after the

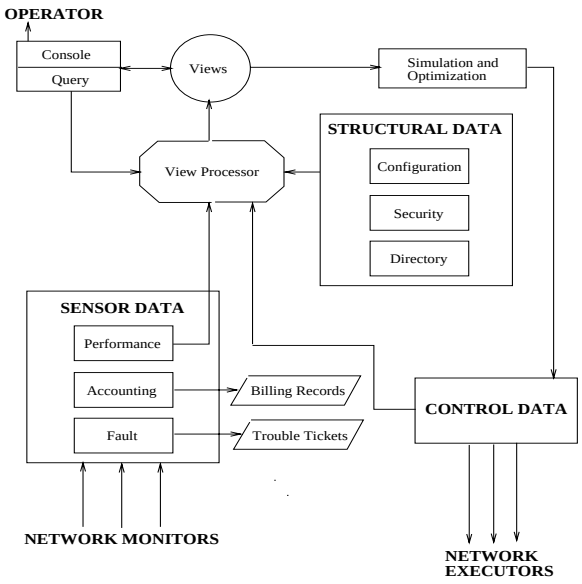

Figure 4.2: The top level view of the OS

feedback from the sensor data. See the Fig. 4.2 for a graphical description.

The design of the OS will include two important features: embedded optimization and a built-in analysis algorithm. Based on the network-wide collected data at the database, the OS can perform sophisticated network analysis and execute predictive control algorithms based on the embedded optimization routines. Optimization algorithms will output a set of values for certain control variables. These values can either be used by the execution process in the database or fed to an analysis algorithm which can evaluate the impact on the network performance of a set of control settings. The simulation program is one such analysis algorithm. The latter mechanism will allow the OS to evaluate the potential impacts of certain network settings before the controls are applied to the network. See Fig 4.3 for illustrating a configuration of these concepts.

Due to the changing nature of the network traffic, we have to limit the time of the above optimization and/or analysis before applying the controls. Otherwise,

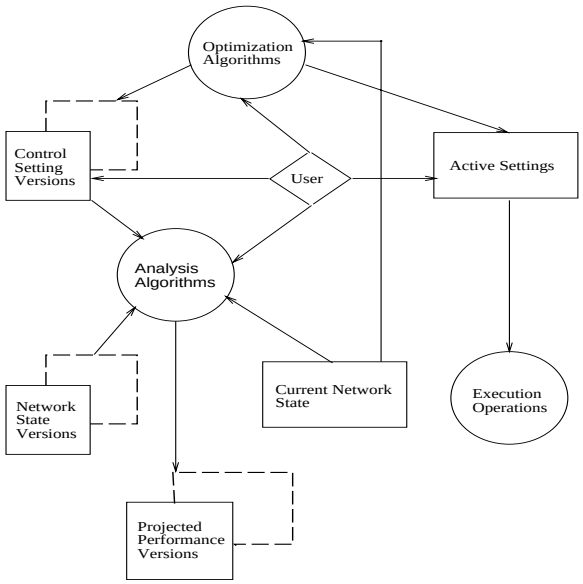

Figure 4.3: Interaction between optimization and analysis

the views of the database to the network might have been changed substantially when the controls apply. Therefore, a precise predictive optimization algorithm is desired. The optimization routines used are also essential. The FSQP we used can produce a feasible solution at every iteration and in most cases it can arrive close to the optimal values quickly. In a time constrained application, we may have to stop the optimization process within a time limit, the output of the control settings are still feasible and close to the optimal which are good enough for our control purpose. On the contrary, if the application is not time constrained, but performance critical, we may have to run the optimization and analysis processes several times before the real controls are applied. In such cases, the performance is very crucial and the consequence of the control may be very expensive.

4.2 Dynamic control procedure

The control procedure operates in several steps: (1) periodically measure and monitor the network traffic, (2) state prediction or estimation based on the real time data measurement, (3) performance optimization and/or analysis, (4) apply controls and repeat the procedure. The traffic control variables, for example, may determine the portion of the incoming traffic to be assigned to different paths of a particular SD pair during the upcoming period.

Every ΔT time unit, each sensors sends the current state of the network and other dynamic network data to the OS through the MN. There are two commonly available protocols for such data collection: Simple Network Management Protocol (SNMP) and Common Management Information Protocol (CMIP) [61]. In SNMP where the raw measurement data are sent to the OSs without processing, we can execute some query-like commands to obtain the state information we want. In CMIP, the data collected is processed before it is sent to the OS. Therefore the state information we need must be defined beforehand in the CMIP.

The state of the network can be defined either as the number of busy circuits on its outgoing trunks, or the occupancy of the buffers of the switches, or the arriving rates of ATM traffic classes, etc, depending on the applications. In addition to the state data, the sensors may transmit other information such as the number of service requests to all network destinations during the preceding time interval. Using this information, the prediction module will compute and project the future occupancy levels indicating congestion levels which can be used for routing and admission control. Based on this projection, the optimization will then compute a control policy that would result in "optimal" network behavior during the upcoming interval. The MN will relay the new control variables from

the OS to the Call Control Module (CCM). For example, these control variables specify the portion of the incoming traffic to be assigned to each route.

4.2.1 Call setup procedure

When a source Call Transport Module (CTM) receives a service request, it transmits this message to its designated CCM for treatment using the CCS-like Protocol. The CCM either recommends a complete route to the destination or rejects the service according to the particular service traffic control variables. If the service request is accepted, then the source CTM inserts the path identifications that comprise the recommended route and transmits the signaling message.

In the event that a CTM on the planned route is unable to locate an idle circuit on the designated path, an Unsuccessful Backward Setup Message (UBSM) is transmitted to the source CTM, which in turn, passes the message to the controlling CCM. The CCM will then either reject the service request or recommend another route to the destination.

4.3 State prediction

The techniques presented here use the arrival rate as the state, nonetheless, they can be generalized to other variables that are used as the state of the system. We have several methods to model the arrival rates as a state depending on the traffic conditions. It is reasonable to model the arrival rate as a function of time, however, the predicted value of the arrival rate will be used as a constant in the following update interval. Therefore, the choice of the update interval is important. If this interval is small, then we have higher accuracy in tracking the

state. Nonetheless, the overhead is higher.

There are two major considerations in the choice of T. The interval T must be small enough to capture and control the transients of the network. However, T can not be too small, otherwise it will cause instability of the network as we know from classical control theory. There are other considerations in picking the length of T. In our example, we can choose T to be on the order of half an hour.

In the slow changing traffic case which is likely to be in the light traffic period, we may assume that the arrival rate changes as an almost linear function of time since the changes in the arrival rate are not abrupt. On the contrary, in the heavy and high changing traffic load, the arrival rate is likely to be a nonlinear function of time. Therefore, the prediction methods will be different in these two cases, namely, a linear filter vs. a nonlinear filter. It is noted that during stable traffic condition, the accuracy of the predictor is not of substantial consequence. In comparison, in heavy traffic condition, the accuracy of the predictor is critical, since the error will result in dire consequence (e.g. congestion.) Thus, care must be taken in the latter case.

4.3.1 Linear traffic prediction

In the period of stable traffic, the arrival rate is assumed to be an almost linear function. We can model it by the following linear state equation (the SD subscripts are omitted for convenience) :

$$x(t) \equiv \begin{pmatrix} \lambda(t+T) \\ \frac{d}{dt}\lambda(t+T) \end{pmatrix} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda(t) \\ \frac{d}{dt}\lambda(t) \end{pmatrix} + \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix} \tag{4.1}$$

where the state is $\lambda(t)$ is captured by the first linear function of its changing rate $\frac{d}{dt}\lambda(t)$, the second equation indicates that the changing rate of $\lambda(t)$ is almost constant except it is varied by a white Gaussian noise, which indicates that the traffic is slow changing, the state noise processes w_1 and w_2 model the uncertainty in the model and are assumed to be white Gaussian with zero mean. Denote the state coefficient matrix as Φ.

$$\Phi(t) = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$$

The goal of this equation is to model the arrival rate by its Taylor expansion to the first order plus some error (noise). Thus, it is suitable for an almost linear function. The measurement equation is

$$y(t) = (1 \quad 0) \begin{pmatrix} \lambda(t) \\ \frac{d}{dt}\lambda(t) \end{pmatrix} + v(t) \tag{4.2}$$

where $y(t)$ denotes the measurement of the state, $v(t)$ is the measurement noise and is also assumed to be a white Gaussian process with zero mean. Denote the measurement coefficient matrix, $(1 \quad 0)$, as H. The measurement noise and the state noise are assumed to be independent. The variances of these noise processes are

$$E\{W(t)W(t)^T\} = Q = \begin{pmatrix} q_{11} & 0 \\ 0 & q_{22} \end{pmatrix} \text{ and } \quad E\{v(t)^2\} = \sigma^2 \tag{4.3}$$

where T superscript denotes the transpose operation.

Using this model, we can form an unbiased estimate $\hat{\lambda}(t)$ of the true arrival rate $\lambda(t)$, which is the minimum mean square error estimator [19]. The solution to this problem is well known to be a discrete Kalman filter where the

new estimate is formed by a weighting sum of the past estimate and a current measurement by

$$\hat{\lambda}(t,t) = [1-\alpha(t)]\hat{\lambda}(t,t-T) + \alpha(t)y(t) \tag{4.4}$$

$$\frac{d}{dt}\hat{\lambda}(t,t) = [1-\beta(t)]\frac{d}{dt}\hat{\lambda}(t,t-T) + \beta(t)\frac{[y(t)-\hat{\lambda}(t-T,t-T)]}{T} \tag{4.5}$$

where $\hat{\lambda}(t,t)$ is the filtered estimate of $\lambda(t)$ using measurement $y(t)$ from 0 up to time t, $\hat{\lambda}(t,t-T)$ is the predicted estimate of $\lambda(t)$ using measurement $y(t)$ from 0 up to time $t-T$, which is obtained by the state estimate extrapolation equation

$$\hat{\lambda}(t,t-T) = \hat{\lambda}(t-T,t-T) + T\frac{d}{dt}\hat{\lambda}(t-T,t-T) \tag{4.6}$$

The weighting coefficients $\alpha(t)$ and $\beta(t)$ are the so called Kalman gains and may be computed recursively using the Kalman Gain equation. Let $K(t) = \begin{pmatrix} \alpha(t) \\ \beta(t) \end{pmatrix}$

$$K(t) = P(t,t-T)H^T[HP(t,t-T)H^T + \sigma^2]^{-1} \tag{4.7}$$

where $P(t,t-T)$ is the error covariance at time t due to the noise from 0 up to time $t-T$ which can be obtained in the error covariance extrapolation

$$P(t,t-T) = \Phi(t-T)P(t-T,t-T)\Phi(t-T)^T + Q(t-T) \tag{4.8}$$

The error covariance is updated by the following equation:

$$P(t,t) = [1-K(t)H]P(t,t-T) \tag{4.9}$$

For the recursion to begin, we must give the initial conditions as

$$E\{x(0)\} = \hat{x}_0, \quad E[(x(0) - \hat{x}_0)(x(0) - \hat{x}_0)^T] = P_0 \tag{4.10}$$

For the above model to work in different networks, we have to have information about the noise processes whose statistics are used in the prediction method. There are two choices. The first one is to use model fitting techniques to mimic the real traffic data. For example, we collect traffic data in a certain period of time (e.g., evening) of a particular telephone network. We then analyze the data to obtain the noise variance due to modeling and other parameters such as weighting coefficients, etc. These parameters are then used to predict the traffic in real time. The other alternate is to perform a simulation assuming certain traffic conditions and traffic arrivals, etc. The simulation is then executed several times, and the parameter used is the average of the simulated results.

Measurement noise accounts for the difference between the measured rate $y(t)$ and the true arrival rate $\lambda(t)$. We can obtain the measurement noise variance for Poisson arrival [31] using the current estimate of the arrival rate. In this case, the measurement rate is defined as the number of Poisson arrivals N measured during the preceding interval of duration T, divided by T. We know the mean and variance of N is $\lambda(t)$ for a given t. Then

$$E[v^2] = E[(\frac{N}{T} - \lambda(t))^2] \tag{4.11}$$

$$= \frac{\lambda(t)}{T} \approx \frac{\hat{\lambda}(t,t)}{T} \tag{4.12}$$

If the arrival is not Poisson, we can apply this technique by computing the mean and variance of the arriving Point processes of the particular distribution and apply the previous equation. In case the distribution is not known at all or we do not make assumption about the arriving processes, we can compute the

statistics from the empirical data analysis of the collected data by model fitting.

4.3.2 Nonlinear state prediction

Even though the previous method can be used to model some fast changing arriving rate cases by allowing the state error (q_{22}) to be larger, we still use the linear Kalman filter to estimate the state. In case of the the state (arrival rate) changes in a nonlinear fashion, then a nonlinear estimation must be used in its place because this will yield a more correct estimation. As we mentioned earlier, in such (heavy traffic and fast changing arrival rate) cases accuracy is often required to avoid congestion.

Now the system in general can be described as

$$\dot{x}(t) = f(x(t), t) + w(t) \tag{4.13}$$

where $x(\cdot)$, $f(\cdot)$, and $w(\cdot)$ could be vectors. The function $f(\cdot)$ is a nonlinear function of the state and time t. In our modeling, this function is to be specified by the system depending on the applications. To be more specific, we have collected enough data from the past to show that in some period of time the arrival rate is going to be changing in a nonlinear fashion. For example, this function is exponential or second (or higher) order in the state, whose exact form can be determined from the collected data by data fitting. Then we can use this estimation skill in such a busy period and switch back to the previous linear procedure in a slow changing traffic period. The measurement equation remains the same.

There are several techniques that can be used to meet our needs. One of the simple and efficient methods is to use the extended Kalman Filter [19] which is

the minimum variance estimator of the state. The idea behind this approach is to approximate the nonlinear function by its Taylor expansion to its first order about a known vector $\bar{x}(t)$ that is close to $x(t)$. In particular, if f is expanded about the current estimate of the state $\hat{x}(t)$, then

$$f(x,t) = f(\hat{x},t) + \left.\frac{\partial f}{\partial x}\right|_{x=\hat{x}} (x-\hat{x}) + ... \tag{4.14}$$

The state estimate propagation equation is then

$$\frac{d}{dt}E[x(t)] = f(\hat{x}(t)), \quad t_{k-1} \le t \le t_k \tag{4.15}$$

The error covariance propagation equation is

$$\dot{P}(t) = F(\hat{x}(t),t)P(t) + P(t)F^T(\hat{x}(t),t) + Q(t), \quad t_{k-1} \le t \le t_k \tag{4.16}$$

where $F(\hat{x}(t),t)$ is the matrix whose ijth element is given by

$$f_{ij}(\hat{x}(t),t) \triangleq \left.\frac{f_i(x(t),t)}{\partial x_j(t)}\right|_{x(t)=\hat{x}(t)} \tag{4.17}$$

The error covariance update equation (4.9) and the Kalman Gain matrix equation (4.7) remain the same, and we can use the same state update equation (4.4) of the previous section.

One of the important differences between this result and the conventional one is that the Kalman gains $K(t)$ are actually random variables depending on the estimates $\hat{x}(t)$ through the matrix $F(\hat{x}(t))$. Hence, the sequence $\{K(t),$ $t = 0,\ T,\ 2T,...\}$ must be computed real time; it can not be precomputed (off-line) before the measurements are collected and stored in the computer memory.

Other methods are available, see [22, 30]. We could also use higher order approximation to $f(\cdot)$. The tradeoff here is that the more complex the model we use, the more accurate the estimate we obtain. The price we pay is more computer time and more subtle implementation.

4.3.3 Centralized implementation

With the above network control architecture and optimization formulation from Chapter 2 and Chapter 3, we can summarize the control procedure for a centralized implementation in which there is only one OS. The OS may reside in a Network Control and Management Center. This Center is the hub for satellite control and has the fast computing power for its processing to achieve dynamic control. The summary is as follows:

1. At the beginning of an update interval of duration T, the sensors of each network node measure the arrival rates of data and voice from this node to every other node (or other arrival rates, for example, images arrival rates, if any, plus other sensor data, for example, queueing delay, buffer occupancies, etc., if necessary) and relay this data to the OS through the MN.

2. The OS chooses the filters automatically to predict the arrival rates from every node to every other node for this coming interval T, depending on the traffic conditions at different time periods: linear filter for slow changing traffic (e.g. in the evening) or nonlinear filter for fast changing traffic (e.g. during day time). The obtained values are the predicted traffic requirements of every SD pair.

3. Based on the prediction values of the arrival rates, the operator may direct the OS to compute the routing ratios using the weighted sum optimization in Chapter 2 or performance constrained optimization and real-time constrained optimization in Chapter 3, depending on varying circumstances.

4. The operator may choose to feed the control to a simulation program if he wants to answer what-if (for example, performance or stability) questions or he/she may choose to feed the control settings directly to the execution processes in the database.

5. The execution processes are triggered to set-up the routing ratios to the corresponding Call Control Modules of network nodes.

6. The traffic is sent splitted to ground stations and to the satellite according to the routing ratios for data and voice respectively.

7. The results of control including delays and blocking probabilities in the satellite links and terrestrial links are relayed to the OS for data analysis.

8. Repeat the procedure for the next update interval T.

4.4 Distributed implementation

In this section, we formulate a distributed implementation of the aforementioned control procedure. The motivation is severalfolds: speeding up of computation, more reliability than centralized control, and more practicality in real applications. The speed-up of the computation is a result of decomposition of the original one big nonlinear programming problem into several subproblems by the so-called equilibrium programming. The subproblems with fewer dimensions and

fewer variables can be executed in parallel. It is more reliable because in a centralized version if one component of the Control Center breaks down, the system is totally out of control. In comparison, in a distributed system if one component fails it will not result in complete system chaos. The distributed implementation is also practical because we have a heterogeneous network where a central controller for such a heterogeneous and multi-vendor network is unlikely to exist. In contrast, the communication of these multi-vendor subnetworks by their OSs which may be different from each other is a natural way of system integration. It can achieve a "global" optimization by doing distributed optimization of each subnet based on global information obtained by exchanging its control information with control information of other subnets. This is the principal of the divide-and-conquer algorithm and system integration.

Consider a general mixed-media network. This network can be divided into several subnets by selected criteria, for example, by geographical locations, or by vendors, or by the size in which we limit the number of nodes in the subnet. Suppose we have M subnets, each with $N_1,\ N_2,\ ,...,\ N_M$ nodes, then the total number of the decision (routing) variables from every node in subnet i to every other node in subnet j with k types (k=2 in our case, voice and data) of traffic is

$$2k \sum_{i=1}^{M} N_i \sum_{i \neq j, j=1}^{M} N_j \tag{4.18}$$

which is on the order of $O(kM^2N_i^2)$ if every node has about the same size of nodes N_i. For example, a medium size network with 10 subnets, each subnet having 40 nodes, two types of traffic, then the total number of decision variables is 576000 which is formidable for any optimization packages to handle at once. As an alternative, each subnet runs a decomposed problem such that each subnet

only optimizes the routing ratios from a node in its subnet to every node in other subnets based on the information obtained from other subnets (the information can be the routing ratios for a node in other subnets sending traffic to the nodes in its subnet). The decision variables for each subnet is on the order of $O(kMN_i)$. This results in an order of reduction in the number of variables. In our example, the total number of the decision variables is 720. Consider a particular optimization algorithm has complexity $O(N^2)$, where N is the number of the input variables. The difference in computational overhead between a centralized implementation and a distributed one is further amplified. This shows that a distributed implementation is the key to solve an optimization problem in a large heterogeneous network.

4.4.1 Equilibrium Programming

The distributed implementation utilizes the theory of Equilibrium Programming (EP). In this section, we present the general formulation of an EP problem. The first-order necessary conditions that are satisfied by a solution to an EP problem is given. A theorem on solution existence, some solution properties, and some solution algorithms are summarized.

Equilibrium Programming is a generalization of nonlinear programming which can be characterized as having M decision-makers (which may be implemented as search algorithms) that interact in a system [17]. Each decision-maker has a nonlinear programming subproblem to solve, and an independent set of design variables to control. To simplify notation, the design variables controlled by decision-maker i are denoted as x^i, the design variables of all M decision-makers are denoted as $x = (x^1, ..., x^m)$, and all design variables not controlled

by decision-maker i are denoted as $\bar{x}^i = (x^1, ..., x^{i-1}, x^{i+1}, ..., x^M)$. Decision-maker i has an objective function to minimize, $f^i(x^i, \bar{x}^i)$, while satisfying a set of constraints. Thus, the general mathematical description of EP is:

$$\min_{x^i} f^i(x^i, \bar{x}^i) \tag{4.19}$$

$$g^i(x^i, \bar{x}^i) \leq 0 \tag{4.20}$$

$$h^i(x^i, \bar{x}^i) = 0 \tag{4.21}$$

for the $i = 1, ..., M$ interacting subproblems. The variables following the comma in any of the functions in statement (4.20) to (4.21) are treated as fixed parameters in that subproblem. Thus, in the nonlinear subproblem of decision-maker i, the design variables from other decision-makers $\bar{x}^i$ enter as parameters.

The solution to all nonlinear subproblems represented by statement (4.20) is $x = x^*$, which is called an equilibrium point. In a manner similar to a nonlinear programming problem, first order necessary conditions are satisfied at the equilibrium point subject to a constraint qualification. Thus, at the equilibrium point $x = x^*$, there exist Lagrange multipliers (λ^i, μ^i) such that the following conditions are satisfied:

$$\frac{\partial f^i(x^i, \bar{x}^i)}{\partial x^i} + (\lambda^i)^T \frac{\partial g^i(x^i, \bar{x}^i)}{\partial x^i} + (\mu^i)^T \frac{h^i(x^i, \bar{x}^i)}{\partial x^i} = 0 \tag{4.22}$$

$$g^i(x^i, \bar{x}^i) \geq 0 \tag{4.23}$$

$$h^i(x^i, \bar{x}^i) = 0 \tag{4.24}$$

$$\lambda^i \geq 0 \tag{4.25}$$

$$(\lambda^i)^T g^i(x^i, \bar{x}^i) = 0 \tag{4.26}$$

for $i = 1, ..., M$.

In utilizing EP for a design decomposition, it is important that a solution exist for the decomposed problem. In fact, the requirements for existence of an equilibrium point can be used as a guide in the decomposition process. A second concern involves the optimality of the equilibrium point. That is, is the equilibrium point resulting from the decomposition of a design problem also an optimal solution of the original nonlinear formulation? This concern can be addressed by showing that EP necessary condition relation (4.22) to (4.26) for the decomposed problem implies the necessary conditions of the original problem. Both the existence and the optimality of an equilibrium point depend on the satisfaction of a constraint qualification. In our problem, if the equilibrium point is different from the optimal point, the equilibrium point is still suitable for controlling the traffic.

One form for the constraint qualification is given [63]. It is satisfied if for all x feasible to the nonlinear subproblem represented by statement (4.20) to (4.21), and for every i : (1) the vectors $h^i_j(x^i, \bar{x}^i)/\partial x^i = 0,\ j = 1, ..., l$, where l is the number of equality constraints. are linearly independent, and (2) there is at least one solution z^i to the relations

$$\frac{\partial g^i_*(x^i, \bar{x}^i)}{\partial x^i} z^i > 0 \tag{4.27}$$

$$\frac{\partial h^i(x^i, \bar{x}^i)}{\partial x^i} z^i = 0 \tag{4.28}$$

where g^i_* is the vector of inequality constraints in subproblems that are active at x. With regard to the inequality constraints, this constraint qualification essentially states that it is always possible to move into the interior of the feasible

region from a point on the boundary of that region. However, because constraint qualification relation (4.27) and (4.28) must be satisfied individually by the EP subproblems, the constraint functions in EP must satisfy more restrictive requirements than those given by the original problems. For example, the constraints from two EP subproblems given by $g_*^1(\cdot, \bar{x}^2)$ and $g_*^2(x^2, \dot{)}$ will not satisfy relation (4.27) and (4.28) individually because $\frac{\partial g_*^1(x^i, \bar{x}^2)}{\partial x^1} z^1 \equiv 0$. However, the relation (4.27) and (4.28) may be satisfied for this example when the constraints and design variables of the subproblems are combined within a single nonlinear problem (i.e., $\frac{\partial g_*}{\partial x} z$ where $g_* = (g_*^1, g_*^2)$ and $x = (x^1, x^2)$) [52].

A very general theorem for the existence of an equilibrium point, given in [63], requires continuity, but not differentiability, of the objective and constraint functions. Other conditions for existence are: (1) the functions satisfy constraint qualification relations (4.27) and (4.28) (actually only a weakened form of the constraint qualification is required, (2) the feasible region is bounded, (3) the functions $f^i(x^i, \bar{x}^i)$ and $g^i(x^i, \bar{x}^i)$ are concave in x^i, and (4) the functions $h^i(x^i, \bar{x}^i)$ are linear in x^i. Other existence theorems are available which relax the concavity and linearity restrictions on f^i, g^i and h^i, respectively, if additional differentiability requirements are satisfied.

Although the differences between an equilibrium point and an optimal point may appear slight, they are important. The following equilibrium point properties illustrate the differences, summarized from [17].

An equilibrium point is, in general, different from an optimal point (i,e., the solution of a nonlinear programming problem), even if the same constraints are satisfied and each EP subproblem has the same objective function. This difference can occur because the coupling of the constraint derivatives in the

respective necessary conditions is generally less for an EP formulation then for a nonlinear formulation. Thus, the subproblem in EP can be called "loosely coupled." Larger differences between the EP and the nonlinear programming problem solutions can occur when M subproblems in EP have M different, and possibly conflicting, objective functions. An equilibrium point has a "stability" property which states that a solution to the EP problem does not change for a perturbation of the design variables of a single subproblem from equilibrium values. Additional constraints can effect EP solutions differently than the nonlinear programming problem. In a nonlinear programming problem, additional constraints generally increase the value of the objective function for a minimization problem. However, in EP it is possible for additional constraints to force a "cooperation" that reduces the objective function of all the EP subproblems.

It should also be noted that EP is different from a multiple-objective function problem. The multiple-objective problem has a single decision-make trying to optimize several different objectives at once. EP, in contrast, has several different decision-makers (each perhaps with multiple objectives), and most critically the decision-makers interact in a system. EP thus stresses the system aspect and that there are M decision-makers interacting and interrelating in a system.

The EP solution algorithm can be obtained in several ways. The most straightforward method is to solve sequentially all the subproblems in some predetermined cyclic order [52]. When the solutions to all the subproblems do not change from the previous iteration (within some numerical tolerance), the equilibrium point has been reached. Using the solution of the set of nonlinear necessary condition equation [17] is also possible and may be advantageous when convergence is difficult to achieve by the sequential solution method. Another

method is used by [23] where these subproblems are executed in parallel and repeat until a convergence occurs, which is the method that we have adapted.

4.4.2 EP formulation

One starting point of the EP formulation of our nonlinear programming problem (NLP) problem in Chapter 2 and Chapter 3 is to decompose the nonlinear programming problem in a systematic way. There could be many ways of decomposition. We should try to resolve the problem of the "optimal" decomposition.

For example, in a weighted sum approach of multiple objectives, there are four objectives to optimize: average delay in the satellite channel, average delay in the terrestrial links, average blocking probability in the satellite channel, and average blocking probability in the terrestrial links. We rewrite the equation to show the dependence of individual objective on the variables (arrival rates which is a function of the routing variables) for the fixed boundary scheme, denote $g = \{g_{ij}\}$ and $s = \{s_{ij}\}$

$$\min f(x) = AD_g(\lambda_{gd}) + BP_b(\lambda_{gd}) + CD_s(\Lambda_d) + DP_s(\Lambda_v)$$

subject to

$$0 \leq g_{ij}, s_{ij} \leq 1,\ \lambda_{gd}(g) \leq C_{gd},\ \lambda_{gv}(s) \leq C_{gv},\ \Lambda_d(g) \leq C_{sd},\ \Lambda_v(s) \leq C_{sv},\ \forall l \tag{4.29}$$

that is, the objective function is

$$\min f(g_{ij}, s_{ij}) = AD_g(g) + BP_b(s) + CD_s(g) + DP_s(s)$$

In a movable boundary scheme, the average delay in the satellite will be a function of both the arrival rates of voice (Λ_v) and data (Λ_d), thus a function of

both the voice traffic routing ratios s_{ij} and data traffic routing ratios g_{ij}

$$\min f(x) = AD_g(g) + BP_b(s) + CD_s(g,s) + DP_s(s)$$

In a fixed boundary scheme, we can decompose the NLP into two subproblems by the traffic type subject to the same constraints as

$$\min f^1(g) = AD_g(g) + CD_s(g)$$

subject to

$$0 \leq g_{ij} \leq 1,\ \lambda_{gd} \leq C_{gd},\ \Lambda_d \leq C_{sd},\ \forall l \tag{4.30}$$

$$\min f^2(s) = BP_b(s) + DP_s(s)$$

subject to

$$0 \leq s_{ij} \leq 1,\ \lambda_{gv} \leq C_{gv},\ \Lambda_v \leq C_{sv},\ \forall l \tag{4.31}$$

The formulation actually divides the total routing decision variables (variables for short) into two halves. Each subproblem f^1 and f^2 has half of the variables according to the traffic types: voice or data. These two subproblems can be executed in parallel without exchanging information because the variables in two subproblems are independent. Note that we can't further decompose each subproblem, because D_g and D_s are inter-related.

In a movable boundary scheme, we have the following problem

$$\min f^1(g,s) = AD_g(g) + CD_s(g,s)$$

subject to

$$0 \leq g_{ij}, s_{ij} \leq 1,\ \lambda_{gdl} \leq C_{gdl},\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_d \leq C_{sd},\ \Lambda_v \leq C_{sv},\ \forall l \tag{4.32}$$

$$\min f^2(s) = BP_b(s) + DP_s(s)$$

subject to

$$0 \leq s_{ij} \leq 1,\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_v \leq C_{sv},\ \forall l \tag{4.33}$$

In this formulation, the decision-maker 1 needs information from decision-maker 2 while decision-maker 2 relies no information from decision-maker 1. Thus, this is one-way communication. Once given the information from decision-maker 2, subproblem 1 has the same number of variables, thus the same complexity, as in the fixed boundary scheme. The overhead in the movable boundary scheme is the one-way communication.

EP formulation with M subnets using weighted sum approach

However, these decompositions do not lead us to much computer-time saving (the number of control variables is reduced from $O(N)$ to $O(N/2)$, where N is the number of control variables). Besides, the decomposition itself does not lead to a distributed implementation, either. According to the analysis in section 4.4, we would like to obtain complexity reduction by decomposing the network geographically (or by vendor, or by its size, etc.) into M subnets. Dividing the large network geographically is natural because the LANs or WANs are usually connected in their covering areas. A satellite to connect these geographically distributed LANs or WANs is a common practice as we have modeled it in previous Chapters. Thus we prefer that LANs or WANs be able to compute independently the routing ratios for nodes in their areas to other nodes in other areas with information obtained from other subnets. We can further decompose the LANs or WANs by their vendors, because networks of different vendors have different protocols and operating systems or we can decompose by the size to limit the computational overheads. Of course, this will result in many

combinatorial ways of decomposition. We will not explore all these alternatives. Instead we give an example to show how the distributed scheme is implemented.

Define set x^k as the vector of control variables for subnet k

$$x^k = \{(g_{kj}, s_{kj}) | \text{ node } k \text{ is in subnet } k,\ j \in \text{ other subnet}\} \tag{4.34}$$

Define set g^k as the vector of data control variables for subnet k

$$g^k = \{(g_{kj}) | \text{ node } k \text{ is in subnet } k,\ j \in \text{ other subnet}\} \tag{4.35}$$

Define set s^k as the vector of voice control variables for subnet k

$$s^k = \{(s_{kj}) | \text{ node } k \text{ is in subnet } k,\ j \in \text{ other subnet}\} \tag{4.36}$$

For the fixed boundary scheme

$$\min_{x^i} f^i(x^i, \bar{x}^i) = AD_g(g^i, \bar{g}^i)) + BP_b((s^i, \bar{s}^i)) + CD_s((g^i, \bar{g}^i)) + DP_s((s^i, \bar{s}^i)) \tag{4.37}$$

subject to

$$0 \leq g_{ij} \leq 1, 0 \leq s_{ij} \leq 1,\ \lambda_{gdl} \leq C_{gdl},\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_d \leq C_{sd},\ \Lambda_v \leq C_{sv},\ \forall l$$

where the objective function is a weighted sum of multiple objectives.

For the movable boundary scheme

$$\min_{x^i} f^i(x^i, \bar{x}^i) = AD_g(g^i, \bar{g}^i)) + BP_b((s^i, \bar{s}^i)) + CD_s((x^i, \bar{x}^i)) + DP_s((s^i, \bar{s}^i)) \tag{4.38}$$

subject to

$$0 \leq g_{ij} \leq 1, 0 \leq s_{ij} \leq 1,\ \lambda_{gdl} \leq C_{gdl},\ \lambda_{gvl} \leq C_{gvl},\ \Lambda_d \leq C_{sd},\ \Lambda_v \leq C_{sv},\ \forall l$$

where the objective function is a weighted sum of multiple objectives.

The decomposed M sub-problems are executed in parallel by M decision-makers for the $i = 1, ..., M$ interacting subsystems ($M = 2$ in our numerical examples given in Chapter 2). By this decomposition, the number of variables of each subnet to decide has been reduced to $O(\sqrt{N})$, where N is the total number of variables before decomposition.

We now examine the optimality of the equilibrium point. Due to a finite range limit $[0, 1]$ on all the decision variables, we found that the Lagrange multipliers exist for equation (4.22). To simplify the problem, we assume the traffic requirement is such that the capacity constraints are not violated. The first order necessary condition is

$$\frac{\partial f^i(x^i, \bar{x}^i)}{\partial x^i} + (\lambda_1^i)^T \frac{\partial g^i(x^i, \bar{x}^i)}{\partial x^i} = 0 \tag{4.39}$$

since the element in $g^i(x^i, \bar{x}^i)/\partial x^i$ is either -1 or 1, we can find positive Lagrange multipliers for this problem.

The constraint qualification is satisfied by the calculation of $\partial g^i(x^i, \bar{x}^i)/\partial x^i$, as we can find solutions z^i to the equation

$$\frac{\partial g_*^i(x^i, \bar{x}^i)}{\partial x^i} z^i > 0 \tag{4.40}$$

by

$$z_j^i = \begin{cases} 1 & \text{if } \partial g_*^i(x^i, \bar{x}^i) = 1 \\ 0 & \text{if } \partial g_*^i(x^i, \bar{x}^i) = -1 \end{cases} \tag{4.41}$$

This assures that the existence of the equilibrium point.

Though this section demonstrates how to decompose the weighted-sum NLP, the above procedure can be applied to minimax NLP or the NLP to get the trade-off curves in Chapter 3.

4.4.3 Summary of the distributed implementation

In the distributed implementation, we have M subnets, and each subnet has its own OS which is responsible for the optimization of its subproblem and communication with other OSs. It's clear though the problem size has been reduced, the communication overhead is increased substantially. There is a trade-off for choosing M between problem-size reduction and the communication overhead.

One of the advantages of the mixed-media networks is that the communication overhead can be alleviated when using EP to solve the NLP. This is because we can utilize the ubiquitous nature of satellite broadcasting capability. Besides, there is a synchronization problem. Suppose there are no satellites, then every OS must wait for all other OSs' information and at the same time send information to other OSs. With the satellite, assuming each OS sends its information in a TDMA fashion, the collected information of all OSs can be then broadcasted at next time frame. Hence, the EP problem solving procedure utilizing satellite is preferred.

The Distributed implementation is summarized as follows:

1. At the beginning of an update interval of duration T, the sensors of each subnetwork node measure the arrival rates of data and voice from this node to every other node, and the amount (ratio) of traffic that goes through terrestrial links and satellite channel. The data are then relayed to the OS through the MN.

2. The OS chooses the filters to predict the arrival rates from every node to every other node for this coming interval T, depending on the traffic conditions of different time periods: linear filter for slow changing traffic

(e.g., in the evening) or nonlinear filter for fast changing traffic (e.g., during day time). The obtained values are the predicted traffic requirements of every SD pair.

3. Each OS sends two pieces of filtered information to the satellite for broadcasting: the traffic requirement of every SD pair and the corresponding routing ratios of every SD pair.

4. Based on the broadcasted values of the traffic requirement and routing ratios of other OS, the operator may direct the OS to compute the routing ratios using weighted sum optimization in equations (4.37) for fixed-boundary scheme or (4.38) for movable-boundary scheme or the decomposed problem of performance constrained optimization and real-time constrained optimization in Chapter 3, depending on differing circumstances.

5. Each OS sends the obtained results (routing ratios only) of computation to the satellite in the same fashion for broadcasting and repeats the computational procedure (Step 3) until the obtained results of one iteration of computation is not different from the previous one "significantly".

6. The operator may choose to feed the control to a simulation program if he wants to answer what-if (for example, performance or stability) questions or he/she may choose to feed the control settings directly to the execution processes in the database.

7. The execution processes are triggered to set-up the routing ratios to the corresponding Call Control Modules of network nodes.

8. The traffic is sent splitted to ground stations and to the satellite according to the routing ratios for data and voice respectively.

9. The results of control including delays and blocking probabilities in the satellite links and terrestrial links are relayed to the OS for data analysis.

10. Repeat the procedure for the next update interval T.

Chapter 5

Dynamic Routing of ATM-based Hybrid Networks

In the previous Chapter, we formulated a distributed implementation using Equilibrium Programming. There are M decision makers in such a system. Each decision maker cooperates with others by communication regularly in order to minimize his (or her) individual "cost". It's a form of game theoretic approach, i.e., a cooperative M players game.

In this Chapter, we formulate the routing problem of mixed-media networks with ATM terrestrial subnets based on a different game. In this game, there are two players: the network and the user. They cooperative optimize a common objective (the to-be-defined Social Welfare). From this game, we can formulate the routing problem for ATM-based hybrid networks. We can not formulate the routing problem of ATM-based hybrid networks as we have presented in previous Chapters because the exact analytical expressions for cell delay or cell loss ratio, etc., are not already available or not already verified. Therefore another approach must be adapted.

5.1 Introduction

In an inspirational paper by Low and Varayia [43], a game theoretic formulation of service provisioning in ATM networks was introduced using theory and concepts from Economics. One of the key observations in their paper is that bandwidth and buffers are "substitute resources" to meet users service quality requirements. In this thesis, we can consider that the satellite subnet is "another resource" to provide services as the terrestrial ATM subnets can not due to resource shortage.

As in [43], an ATM service is specified by a one-way virtual connection (source, destination, route) and two sets of service parameters negotiated before connection is established. A connection is used to transport a data stream or message from the source to the destination. One set of parameters specifies constraints on a user's traffic "burstiness", and the other specifies the QOS constraints such as cell loss rate and end-to-end delay. The network offers different types of services by a pricing mechanism which is to be explored. There are message flows, depending on the service cost, of user requests for these services. A service request is admitted if sufficient resources can be allocated along the connection's path, either through the satellite channel or terrestrial ATM networks, to guarantee the QOS. Network resources include bandwidth and buffers of the terrestrial networks and the satellites.

To be more specific, a network offers a set of S types of services. A unit of type s service is provided by a type s connection with the associated traffic and quality parameters, and is sold for a unit price of w_s. The network can produce any amount of services if the required resources do not exceed its capacity. Users request services according to the resource price. The network adjusts the resource

price and the user adjust the allocation of bandwidth and buffers to maximize the Social Welfare (to be defined).

The routing problem to be formulated is to further optimize the Social Welfare by adjusting the amount of services that goes to the satellite given the resource price and user allocation.

5.1.1 Model

As in the previous Chapters, we consider a hybrid network, but the terrestrial network is now an ATM network. The ATM network has a set L of links. Link $l \in L$ consists of a link bandwidth of C_l cells per second (cps), and buffers B_l. The satellite has C_S bandwidth and B_S buffers. A set $R = R_A \cup R_S$ of routes is specified which is the union of a set R_A of routes consisting of terrestrial link only and a set R_S of routes consisting of both terrestrial links and the satellite link. An element $r \in R$ comprises the set of links (of terrestrial networks or of terrestrial networks and the satellite) along that route. This mixed-media network offers a set S of services. These services are differentiated by both their source-destination (SD) pairs and the traffic types (voice, data, video, etc.) with different QOS. To simplify the problem, we assume the routing for type s service is fixed. Both routes for type s service are offered: a fixed-route through ATM networks only and a route consisting of both terrestrial links and the satellite link. This is to say that for every SD pair there is two pre-assigned routes. Since ATM network is a connection-based network, one service call can take only one of these two paths. No fractional traffic of a connection is allowed in our modeling. The decision problem is not for a particular arrival request to split its traffic between ATM subnet and the satellite channel. But the routing

problem is how many of the total arriving requests in a period of time for type s service should take the first route (using the ATM resources), and how many (the rest) should take the other route which utilizes the satellite. This service differentiation concept is similar to the the multi-commodity flow problems [11] where each SD is a commodity. However, the difference between a service and a commodity is that service type for each SD pair is further differentiated by its traffic type (voice, data, video, etc., each with different QOS). With this differentiation, we are able to assign routing for service type s to route r_s.

A unit of type s of service is sold at a price of w_{sA} (w_{sS}) if it is provided by a connection over route r_{sA} (r_{sS}). The price may be some fictitious currency for control purposes that the network can adjust at will. We shall assume that user demand, or requests, for type s service is given by the aggregate demand function for type s service $D_s(w_s) = \nu_s exp(-w_s)$, where $\nu_s \stackrel{\Delta}{=} \lambda_s T_s$, T_s is the average duration of a type s connection, and λ_s is a parameter that associates with the arrival rate of type s service.

The aggregate demand for type s service consists of a portion (U_s) of the total demand that use path r_{sA} (this portion is denoted by $D_s(w_{sA})$) and the other portion $(1 - U_s)$ of the total demand that use the path r_{sS} (this portion is denoted by $D_s(w_{sS})$). Therefore,

$$D_s(w_s) = D_s(w_{sA}) + D_s(w_{sS}) \tag{5.1}$$

where $D_s(w_{sA}) = U_s D_s(w_s)$, and $D_s(w_{sS}) = (1 - U_s) D_s(w_s)$.

Since the prices for two routes of the service will be different, then

$$D_s(w_s) = \begin{cases} \nu_s exp(-w_{sA}) & \text{if } r_s = r_{sA} \\ \nu_s exp(-w_{sS}) & \text{if } r_s = r_{sS} \end{cases} \tag{5.2}$$

Substituting equation (5.2) into (5.1), we conclude

$$D_s(w_s) = U_s\nu_s exp(-w_{sA}) + (1-U_s)\nu_s exp(-w_{sS}) \quad (5.3)$$

$$= T_s[U_s\lambda_s exp(-w_{sA}) + (1-U_s)\lambda_s exp(-w_{sS})] \quad (5.4)$$

where $\lambda_s U_s exp(-w_{sA}) + \lambda_s(1-U_s)exp(-w_{sS})$ is the average arrival rate of type s request, $U_s\lambda_s exp(-w_{sA})$ is the average arrival rate to route r_{sA}, $(1-U_s)\lambda_s exp(-w_{sS})$ is the average arrival rate to route r_{sS}, U_s is the control variable that decides the ratio of arrival requests sent through a path in ATM only subnets for service type s to the requests sent through a path with satellite link, $0 \leq U_s \leq 1,\ s \in S$, and $w_s \triangleq (w_{sA}, w_{sS})$.

The demand function is such that the higher the price, the lower the demand. Therefore, other types of demand function with negative slope can be used. In general, using a satellite is more expensive now. But this could change in some areas where the optical fiber is difficult to deploy. The control variable U_s is the portion of the requests in a time period that uses ATM links only. One way to enforce the portion is to use U_s as a probability of choosing the route that uses ATM subnets only for a single service s connection. With this probabilistic choosing interpretation, the portion of average aggregate demand D_s that use path r_{sA} in a period of time will be U_sD_s. This implementation also avoids the fractional connections which are not allowed in ATM paradigm for fluid-flow model when U_sD_s is not an integer.

Therefore, a type s request is admitted and assigned a connection with probability U_s a fixed-route $r_s = r_{sA}$ provided there is available bandwidth of μ_{sl} cps and spare buffers of b_{sl} cells, in each link $l \in r_{sA}$.

A type s request is admitted and assigned a connection with probability $1-U_s$ a fixed-route $r_s = r_{sS}$ provided there is available bandwidth of μ_{sk} cps, available

satellite bandwidth μ_{sS} cps, spare buffers of b_{sk} cells, and available satellite buffers b_{sS} cells, in each link $k \in r_{sS} \setminus \{l_S\}$, where l_S denotes the satellite link. We shall use r'_{sS} to denote $r_{sS} \setminus \{l_S\}$, the hybrid route for service s excluding the satellite link.

Besides the control variables U_s, the allocation for type s service

$$A_1 : \ s \rightarrow \{(\mu_{sl}, b_{sl}), \ l \in r_{sA}\}$$

$$\text{or} \ \ A_2 : \ s \rightarrow \{(\mu_{sk}, b_{sk}), \ k \in r'_{sS} \text{ and } (\mu_{sS}, b_{sS})\}$$

can be freely chosen provided the service quality constraint is met. Note that this is not a function mapping since it is one-to-many. It is a list of allocation variables which must be decided by the users.

We will use the following vector notation for the rest of this Chapter. The bandwidth allocation for a type s connection

$$\mu_s = \begin{cases} \{\mu_{sl}, \ l \in r_{sA}\} & \text{if } r_s = r_{sA} \\ \{\mu_{sk}, \ k \in r'_{sS}, \ \mu_{sS}\} & \text{if } r_s = r_{sS} \end{cases} \tag{5.5}$$

and μ denotes the vector $\{\mu_s, \ s \in S\}$. Similarly, the buffers allocation for a type s connection

$$b_s(\mu_s) = \begin{cases} \{b_s(\mu_{sl}), \ l \in r_{sA}\} & \text{if } r_s = r_{sA} \\ \{b_s(\mu_{sk}), \ k \in r'_{sS}, \ b_{\mu_{sS}}\} & \text{if } r_s = r_{sS} \end{cases} \tag{5.6}$$

and $b(\mu)$ denotes the vector $\{b_s(\mu_s), \ s \in S\}$. Let U denotes the vector $\{U_s, \ s \in S\}$.

If there is enough network resource, a type s request is admitted, with a connection set up. The user sends the message which can be modeled as a "fluid flow" $m_s(t), \ 0 \leq t \leq T$, where $m_s(t)$ is the instantaneous rate in cells per

second, and T is the duration. A type s message must satisfy two constraints denoted by $(\bar{\rho}_s, b_s(\mu_s))$. The parameter $\bar{\rho}_s$ is a positive real number that upper bounds the average message rate of the user message. It is also the minimum bandwidth allocated by the network for a type s message at each node along its route. The parameter $b_s(\mu_s)$ is a non-negative, decreasing, convex function that upper bounds the user message "bustiness" [43]. The bustiness of a message has different implications. Here, the bustiness is a function curve of the allocated bandwidth which depicts the buffers needed as a function of allocated bandwidth.

We consider the cases where both the network and the user are compliant. A type s message m_s is said to be compliant [43] if

$$\rho_s \triangleq \frac{1}{T}\int_0^T m_s(\tau)d\tau \leq \bar{\rho}_s \tag{5.7}$$

and

$$\max_{0\leq q\leq t\leq T}\int_q^t [m_s(\tau) - \mu_{sl}]d\tau \leq b_s(\mu_{sl}) \quad \mu_{sl} \geq \rho_s,\ l \in r_s \tag{5.8}$$

That is, the compliant average user message rate, ρ_s, cannot exceed $\bar{\rho}_s$, the maximum average message rate. $b_s(\mu_{sl})$ is the maximum backlog if m_s is transmitted over a link l with a constant rate $\mu_{sl} \geq \rho_s$. Hence, inequality (5.8) says that if m_s is allocated a bandwidth of μ_{sl}, then a buffer size $b_s(\mu_{sl})$ is sufficient to prevent cell loss. Note that the large is μ_{sl} the smaller is $b_s(\mu_{sl})$. Thus the function b_s, called the bustiness curve, gives the bandwidth-buffers trade-off for zero cell loss. In general, it is hard to know the bustiness curve. However, we can perform traffic shaping which bounds the bustiness curve before the message is sent out. The traffic shaping devise like leaky bucket scheme [14] can reduce the bustiness to a shape of curve which we can know of. The above statement is proven in [42].

A network service provisioning decision μ is said to be compliant if it the allocation $\{\mu_{ls}, b_{ls}, l \in r_s\}$ is such that

$$\mu_{sl} \geq \underline{\mu_s} \;\; b_{sn} \geq b_s(\mu_{sl}) \;\; l \in r_s \tag{5.9}$$

i.e. at each link including satellite link, the allocated bandwidth exceeds the minimum bandwidth $\underline{\mu_s}$, and the allocated buffers exceed the burstiness constraint.

We require that $\underline{\mu_s} \geq \bar{\rho_s}$ so that the QOS is guaranteed. For simplicity, we can just set $\underline{\mu_s} = \bar{\rho_s}$. However, the $\geq$ means that we can choose the bandwidth allocation in such range as a trade-off between allocation for bandwidth and buffers. An allocation with $b_{sl} = b_s(\mu_{sl})$ is enough to prevent cell loss, which will be assumed hereafter. An allocation includes the bandwidth allocation and the buffers allocation. If the allocation is as in equation (5.9), from [42], we conclude (1) no cells will be lost at any link $l \in r_s$, and (2) the end-to-end delay is at most $\frac{b_s(\mu_s)}{\mu_s}$ + propagation and processing delay

We use $\underline{\mu}$ to denote the vector $\{\underline{\mu_s},\ s \in S\}$, so are variables without subscript like $U,\ x,\ w$, using $\mu \geq \underline{\mu}$ to mean $\{\mu_{sl} \geq \underline{\mu_s},\ l \in r_s,\ s \in S\}$.

To summarize, the allocation problem is to find the "optimal" allocation $\{(\mu_{sl}, b_{sl}),\ l \in r_{sA}\}$ and $\{(\mu_{sk}, b_{sk}), k \in r'_{sS}$ and $(\mu_{sS}, b_{sS})\}$ that guarantees the QOS for each $s \in S$ for the compliant user and the compliant network. The routing problem is to find the "optimal" ratio of services that utilize the satellite, U_s, for each service type s.

5.2 Social Welfare and Social Expenditure

From our model, the network can produce any amount supply x_s of type s service, $U_s x_s$ using the route r_{sA} and the rest using the route r_{sS}, provided that sufficient resources are available,

$$\sum_{s,\ l\in r_{sA}} U_s x_s \mu_{sl}\delta_{lj} \quad + \quad \sum_{s,\ k\in r'_{sS}} (1-U_s) x_s \mu_{sk}\delta_{kj} \le C_j \quad j \in L \tag{5.10}$$

$$\sum_{s,\ l\in r_{sA}} U_s x_s b_s(\mu_{sl})\delta_{lj} \quad + \quad \sum_{s,\ k\in r'_{sS}} (1-U_s) x_s b_s(\mu_{sk})\delta_{kj} \le B_j \quad j \in L \tag{5.11}$$

$$\sum_s (1-U_s) x_s \mu_{sS} \le C_S \tag{5.12}$$

$$\sum_s (1-U_s) x_s b(\mu_{sS}) \le B_S \tag{5.13}$$

and expects a revenue of $\sum_s [U_s x_s w_{sA} + (1-U_s) x_s w_{sS}]$. We shall term the constraints (5.10) to (5.13) as the resource capacity constraints.

For a type s service after the price is set at w_s, the demand that would sell at a price higher (from w_s to ∞) is the user surplus or the user's utility. Thus, the aggregate demand function summarizes the user's utility for service type s as $\int_{w_s}^{\infty} D_s(v)dv$. The result of the maximization is summarized by an aggregate demand function $D_s(w_s)$ which can be easily measured by counting the number of service requests, both admitted and rejected. The (total) user surplus is thus $\sum_s \int_{w_s}^{\infty} D_s(v)dv$ [57].

Take the Social Welfare as the sum of user surplus and network revenue

$$\begin{aligned} W(w,x,\mu,U) \ &\triangleq \ \sum_s \int_{w_{sA}}^{\infty} D_{sA}(v)dv + \sum_s \int_{w_{sS}}^{\infty} D_{sS}(v)dv \\ &+ \ \sum_s [U_s x_s w_{sA} + (1-U_s) x_s w_{sS}] \end{aligned} \tag{5.14}$$

so that the problem is to maximize $W(w,x,\mu,U)$ over the variable set (w,x,μ,U) subject to constraints (5.10) to (5.13). Consider initially a fixed allocation $\mu \ge \underline{\mu}$ and a fixed set of routing ratios U. Denote $\pi = (\mu, U)$.

Definition 3 *A set of service prices and amounts of services produced* $\{(w_{sA}, w_{sS}), x_s,\ s \in S\}$ *form an equilibrium if*

$$x_s(\pi) = D_s(w_s(\pi)) = U_s \nu_s exp(-w_{sA}(\pi)) + (1 - U_s)\nu_s exp(-w_{sS}(\pi)) \qquad (5.15)$$

and

$$\sum_s [U_s x_s(\pi) w_{sA}(\pi) + (1 - U_s) x_s(\pi) w_{sS}(\pi)] \geq \sum_s [U_s x_s w_{sA}(\pi) + (1 - U_s) x_s w_{sS}(\pi)] \qquad (5.16)$$

for all $\{x_s\}$ *satisfying constraints (5.10) to (5.13)*

The condition states that, in equilibrium, user demand is met and the network maximizes revenue. Standard price-adjustment scheme [57] can be used to reach an equilibrium, in which the network posts a set of price $\{(w_{sA}, w_{sS}),\ s \in S\}$ and produces supply $\{x_s,\ s \in S\}$, users announce their demand $D_s(w_s)$, and the network increases or decreases the price accordingly as demand is greater or less than supply. The process is repeated until the equilibrium is achieved.

Proposition 1 $(w_{sA}(\pi), w_{sS}(\pi), x_s(\pi)),\ s \in S$ *is an equilibrium if and only if there exists* $(\alpha_j(\pi), \beta_j(\pi), \alpha_S(\pi), \beta_S(\pi)) \geq 0,\ j \in L$ *such that*

1. The supply equals demand

$$x_s(\pi) = U_s \nu_s exp(-w_{sA}(\pi)) \quad + \quad (1 - U_s)\nu_s exp(-w_{sS}(\pi)) \qquad (5.17)$$

2. The total price of the service s is

$$\begin{aligned} U_s w_{sA}(\pi) \quad &+ \quad (1 - U_s) w_{sS}(\pi) = \\ \sum_{j \in L,\ l \in r_{sA}} U_s \alpha_j(\pi) \mu_{sl} \delta_{lj} \quad &+ \quad \sum_{j \in L,\ k \in r'_{sS}} (1 - U_s) \alpha_j(\pi) \mu_{sk} \delta_{kj} \\ + \sum_{j \in L,\ l \in r_{sA}} U_s \beta_j(\pi) b_s(\mu_{sl}) \delta_{lj} \quad &+ \quad \sum_{j \in L,\ k \in r'_{sS}} (1 - U_s) \beta_j(\pi) b_s(\mu_{sk}) \delta_{kj} \\ + (1 - U_s) \alpha_S(\pi) \mu_{sS} \quad &+ \quad (1 - U_s) \beta_S(\pi) b_s(\mu_{sS}) \end{aligned} \qquad (5.18)$$

3. The network maximizes revenue which is

$$\begin{aligned}\sum_s [U_s x_s(\pi) w_{sA}(\pi) &+ (1-U_s) x_s(\pi) w_{sS}(\pi)] = \\ \sum_{j\in L} \alpha_j(\pi) C_j &+ \sum_{j\in L} \beta_j(\pi) B_j + \alpha_S(\pi) C_S + \beta_S(\pi) B_S \end{aligned} \tag{5.19}$$

also, the inequalities from (5.10) to (5.13) become equalities (capacity constraints are met) if $(\alpha, \beta) \triangleq (\alpha_j(\pi), \beta_j(\pi), \alpha_S(\pi), \beta_S(\pi)) > 0,\ j \in L$, *i.e., strictly greater than zero.*

Proof

(5.17) follows from the definition.

By the duality of linear programming $\max_x \sum_s [U_s x_s w_{sA} + (1-U_s) x_s w_{sS}]$ is equivalent to $\min_{\alpha,\beta} \sum_{j\in L} \alpha_j(\pi) C_j + \sum_{j\in L} \beta_j(\pi) B_j + \alpha_S(\pi) C_S + \beta_S(\pi) B_S$. The extreme values of these two optimization are the same if there is a unique value to this problem. This proves (5.19).

By duality, the constraint set for the dual problem is,

$$\begin{aligned} \sum_{j\in L} \sum_{l\in r_{sA}} U_s \alpha_l \mu_{sl} \delta_{lj} &+ \sum_{j\in L} \sum_{k\in r'_{sS}} (1-U_s) \alpha_k \mu_{sk} \delta_{kj} + \\ \sum_{j\in L} \sum_{l\in r_{sA}} U_s \beta_l b_s(\mu_{sl}) \delta_{lj} &+ \sum_{j\in L} \sum_{k\in r'_{sS}} (1-U_s) \beta_k b_s(\mu_{sk}) \delta_{kj} + \\ (1-U_s)\alpha_S \mu_{sS} + (1-U_s)\beta_S b(\mu_{sS}) &\geq U_s w_{sA} + (1-U_s) w_{sS},\ s \in S \end{aligned} \tag{5.20}$$

By the complimentary slackness property of linear programming, since $x_s \neq 0$, the slackness in the dual problem must be zero, and equation (5.18) follows. □

In equation (5.18) if we set $U_s = 1$ and 0, we have the following expressions for equilibrium price.

$$w_{sA}(\pi) = \sum_{j\in L,\ l\in r_{sA}} \alpha_j(\pi) \mu_{sl} \delta_{lj} + \sum_{j\in L,\ l\in r_{sA}} \beta_j(\pi) b_s(\mu_{sl}) \delta_{lj}$$

$$= \sum_{j \in r_{sA}} \alpha_j(\pi)\mu_{sj} + \sum_{j \in r_{sA}} \beta_j(\pi) b_s(\mu_{sj}) \quad (5.21)$$

$$\begin{aligned} w_{sS}(\pi) &= \sum_{j \in L,\ k \in r'_{sS}} \alpha_s(\pi)\mu_{sk}\delta_{kj} + \sum_{j \in L,\ k \in r'_{sS}} \beta_j(\pi) b_s(\mu_{sk})\delta_{kj} \\ &+ \alpha_S(\pi)\mu_{sS} + \beta_S(\pi) b_s(\mu_{sS}) \\ &= \sum_{j \in r'_{sS}} \alpha_s(\pi)\mu_{sj} + \sum_{j \in r'_{sS}} \beta_j(\pi) b_s(\mu_{sj}) \\ &+ \alpha_S(\pi)\mu_{sS} + \beta_S(\pi) b_s(\mu_{sS}) \end{aligned} \quad (5.22)$$

Note these equations mean that the equilibrium price equals the resource cost for providing that service. The cost for the service s to use route r_{sA} is given by taking as the summation of the "shadow" price (or rent) of $\alpha_j(\pi)$ per cps of bandwidth and rent of $\beta_j(\pi)$ per cell of buffer over the links j in its route r_{sA}. The cost for the service s to use route r_{sS} is given by taking as the summation of the "shadow" price (or rent) of $\alpha_j(\pi)$ per cps of bandwidth, rent of $\beta_j(\pi)$ per cell of buffer over the links j in its route r'_{sS}, rent of $\alpha_S(\pi)$ per cps of satellite bandwidth, and rent of $\beta_S(\pi)$ per cell of satellite buffer. If $(\alpha, \beta) > 0$ at the equilibrium, the capacity limits are achieved by all users demand, i.e., the facilities are fully utilized.

It can be verified that there is a unique equilibrium provided that users can choose their demand freely and other continuity conditions of the supply and price functions are satisfied [57], and that the equilibrium maximizes $W(w, x, \mu, U)$ over $x \geq 0$, subject to capacity constraints (5.10) to (5.13), and that the maximum welfare is (by integral of demand function and the Proposition 1)

$$\begin{aligned} W(\mu, U) \triangleq \max_x W(w, x, \mu, U) &= \sum_s x_s(\pi) + \\ \sum_{j \in L} \alpha_j(\pi) C_j + \sum_{j \in L} \beta_j(\pi) B_j &+ \alpha_S(\pi) C_S + \beta_S(\pi) B_S \end{aligned} \quad (5.23)$$

We obtained Proposition 1 by forming the duality of the maximizing problem. Therefore at the equilibrium by duality

$$W(\mu, U) = \max_{x} W(w, x, \mu, U) = \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \tag{5.24}$$

where $G(\alpha, \beta, \mu, U)$ is defined as

$$\begin{aligned} G(\alpha, \beta, \mu, U) \triangleq \sum_{s} \{ U_s \nu_s exp[- \sum_{j \in r_{sA}} \alpha_j(\pi)\mu_{sj} \quad - \quad \sum_{j \in r_{sA}} \beta_j(\pi) b_s(\mu_{sj})] + \\ (1 - U_s)\nu_s exp[- \sum_{k \in r'_{sS}} \alpha_k(\pi)\mu_{sk} - \sum_{k \in r'_{sS}} \beta_k(\pi) b_s(\mu_{sk}) \quad - \quad \alpha_S(\pi)\mu_{sS} - \beta_S(\pi) b_s(\mu_{sS})] \} \\ + \sum_{j \in L} \alpha_j(\pi) C_j + \sum_{j \in L} \beta_j(\pi) B_j \quad + \quad \alpha_S(\pi) C_S + \beta_S(\pi) B_S \end{aligned} \tag{5.25}$$

To get this definition, in equation (5.23), $x_s(\pi)$ is substituted by demand function (5.17) and equilibrium prices are further substitutes by equations (5.22) and (5.21).

One major reason to solve the minimization problem ($\min_{\alpha,\beta} G$) instead of the maximization problem ($\max_x W$) is that the maximization problem has a quite complex constraints, (5.10) to (5.13), which may cause a demanding computation. By duality, the constraint set for the minimization problem is as in (5.20). If the number of service types is less than two times the number of links in the network plus two, it is more computational efficient to solve the minimization problem.

Also, in doing so, we can have more insight into the interpretation of the Social Welfare. From the above definition of G, which is the summation of the total supply and total users payment, G can be interpreted as the Social Expenditure. The first term is derived from maximizing the user's surplus. At equilibrium, this value can be interpreted as the network's labor, that is, the total labor that the network has to work to provide the supply. Thus by duality,

to maximize the Social Welfare is to minimize the Expenditure. At equilibrium, from maximizing the Social Welfare's point of view, the network's revenue is maximized, and the user's surplus is also maximized. From minimizing the Social Expenditure point of view, the network's labor is minimized, and the user's payment is also minimized. From both angles, it is beneficial for both the network and the user to achieve equilibrium. This is why the network and the user would like to work cooperatively in this game formulation.

5.2.1 Optimal routing and optimal allocation

Now suppose $\mu \geq \underline{\mu}$ can be chosen freely and U can be chosen freely after $(x, w)/(\alpha, \beta)$ have been properly chosen. To choose μ as Low and Varayia is to find the trade-off between the allocation of bandwidth and that of buffers for the user to maximize the Social Welfare, since bandwidth and buffers are two resources that can substitute for each other. To choose U, the routing ratios, is the main goal of our formulation. By adjusting U, we can make better utilization of the already existing resources in the satellite and the terrestrial network.

Definition 4 *An allocation μ is optimal, or welfare-maximizing, if it maximizes the welfare $W(\mu, U)$ in (5.24) given a fixed set of routing ratios U.*

The user can adjust the allocation to achieve the maximal Social Welfare. μ^* is a welfare-maximizing allocation given U if

$$W(\mu^*, U) = \max_{\mu \geq \underline{\mu}} W(\mu, U) = \max_{\mu \geq \underline{\mu}} \min_{\alpha, \beta} G(\alpha, \beta, \mu, U) \tag{5.26}$$

This is the two players non-zero-sum dynamic game we want to formulate which has an interpretation of "network economics". One player is the network

who adjusts the rent (α, β) to minimize G over, and the other player is the collective users as one player who chooses allocation to maximize the Social Welfare. From this, we would like to derive the routing problem to our interests.

Definition 5 *A set of routing ratios U is optimal, or welfare maximizing if it optimize the welfare fixed allocation given a fixed set of allocation μ.*

That is U^* is a welfare-maximizing routing given fixed μ if

$$W(\mu, U^*) = \max_{0 \le U \le 1} W(\mu, U) = \max_{0 \le U \le 1} \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \tag{5.27}$$

The saddle points of both these two max-min games exist can be shown rigorously as in [42] or by Von Neumann's [7]. By inspection into (5.25) shows that $G(\alpha, \beta, \mu, U)$ is convex in α, β, thus a minimum exists for $\alpha \ge 0, \beta \ge 0$. $G(\alpha, \beta, \mu, U)$ is also convex in μ and linear in U, but if we take the maximum over a finite range, the maximum exists. That is, the maximum exist for $\underline{\mu} \le \mu, 0 \le U \le 1$.

Put this all together, we have the following Social Welfare function to be optimized

$$\begin{aligned} W(\alpha^*, \beta^*, \mu^*, U^*) &= \max_{0 \le U \le 1} \max_{\underline{\mu} \le \mu} W(\mu, U) = \\ \max_{0 \le U \le 1} \max_{\underline{\mu} \le \mu} \max_{x} W(x, w, \mu, U) &= \max_{0 \le U \le 1} \max_{\underline{\mu} \le \mu} \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \end{aligned} \tag{5.28}$$

This is a very complex function to achieve the optimal value, thus we have to do several optimization computations, which may not be feasible in some applications. We propose a divide and conquer methodology and concentrate on the routing problem.

In summary, this welfare-maximizing problem can be divided into three steps as follows

1. The network adjusts the supply and the price in the price adjustment procedure to achieve the equilibrium. This step is equivalent to the network adjust the shadow rents in order to achieve the equilibrium. That is, fix μ, U, and perform the following

$$W(\alpha^*, \beta^*, \mu, U) = \max_x W(x, w, \mu, U) = \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \tag{5.29}$$

2. The user chooses the allocation to trade-off the bandwidth and the buffers allocation and maximize the following

$$W(\mu^*, U) = \max_{\underline{\mu} \le \mu} W(\mu, U) = \max_{\underline{\mu} \le \mu} \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \tag{5.30}$$

3. We can perform the routing optimization as follows

$$W(\mu^*, U^*, \alpha^*, \beta^*) = \max_{0 \le U \le 1} W(\mu^*, U) = \max_{0 \le U \le 1} \max_{\underline{\mu} \le \mu} \min_{\alpha,\beta} G(\alpha, \beta, \mu, U) \tag{5.31}$$

To concentrate on the routing, we can assume that both the rent and allocation are given (that is, step 1 and 2 have been solved already). These can be optimized off-line or obtained from a historical database. The routing problem for the ATM-base hybrid networks can then be formulated as

$$\max_{0 \le U \le 1} W(\mu, U) = \max_{0 \le U \le 1} G(\alpha, \beta, \mu, U) \quad \text{given} \quad \alpha, \beta, \mu \tag{5.32}$$

The centralized and distributed implementation in the previous Chapter can then be applied to this problem when the network at hand is an ATM-based hybrid network.

5.2.2 Other routing problem formulations

Following the divide-and-conquer methodology and to concentrate on the routing problem, the routing problem for ATM-based hybrid networks can have other formulations. We assume that only U is the decision variable set, all other variable sets are given from the beginning. In comparison to the previous formulation (5.32), where routing variable set is solved after the duality and other decision variable sets are solved, routing variable set in this section is the only decision variable set in the original Social Welfare formulation and other variable sets are given from the beginning.

This problem is to find the ratio of each service request using the satellite. We can have two formulations: the primal problem and the dual problem. The primal problem is to maximize the Social Welfare and formulated as the following given $w,\ x,\ \mu,$

$$\begin{aligned}
&\max_{0\le U\le 1} W(w,x,\mu,U) \\
= &\sum_s U_s\nu_s exp(-w_{sA}) + \sum_s (1-U_s)\nu_s exp(-w_{sS}) + \sum_s [U_s x_s w_{sA} + (1-U_s)x_s w_{sS}] \\
= &\sum_s \{[\nu_s exp(-w_{sA}) - \nu_s exp(-w_{sS}) + x_s w_{sA} - x_s w_{sS}]U_s + x_s w_{sS} + \nu_s exp(-w_{sS})\}
\end{aligned}$$

this is equivalent to

$$\max_{0\le U\le 1} \sum_s [\nu_s exp(-w_{sA}) - \nu_s exp(-w_{sS}) + x_s w_{sA} - x_s w_{sS}]U_s \tag{5.33}$$

since the last two terms in the sum is a constant.

This problem has the following constraints:

$$\begin{aligned}
\{\sum_{s,\ l\in r_{sA}} x_s\mu_{sl}\delta_{lj} \ - \sum_{s,\ k\in r'_{sS}} x_s\mu_{sk}\delta_{kj}\}U_s + \sum_{s,\ k\in r'_{sS}} x_s\mu_{sk}\delta_{kj} \le C_j \quad j\in L \\
\{\sum_{s,\ l\in r_{sA}} x_s b_s(\mu_{sl})\delta_{lj} \ - \sum_{s,\ k\in r'_{sS}} x_s b_s(\mu_{sk})\delta_{kj}\}U_s + \sum_{s,\ k\in r'_{sS}} x_s b_s(\mu_{sk})\delta_{kj} \le B_j \quad j\in L
\end{aligned}$$

$$\sum_s x_s \mu_{sS} - \sum_s x_s \mu_{sS} U_s \leq C_S$$
$$\sum_s x_s b(\mu_{sS}) - \sum_s x_s b(\mu_{sS}) U_s \leq B_S$$

Move the constants in the left hand side of equations to the right hand side, we have

$$\{\sum_{s,\ l \in r_{sA}} x_s \mu_{sl} \delta_{lj} \quad - \quad \sum_{s,\ k \in r'_{sS}} x_s \mu_{sk} \delta_{kj}\} U_s \leq C'_j \quad j \in L \tag{5.34}$$

$$\{\sum_{s,\ l \in r_{sA}} x_s b_s(\mu_{sl}) \delta_{lj} \quad - \quad \sum_{s,\ k \in r'_{sS}} x_s b_s(\mu_{sk}) \delta_{kj}\} U_s \leq B'_j \quad j \in L \tag{5.35}$$

$$\sum_s -x_s \mu_{sS} U_s \leq C'_S \tag{5.36}$$

$$\sum_s -x_s b(\mu_{sS}) U_s \leq B'_S \tag{5.37}$$

where $C'_j = C_j - \sum_{s,\ k \in r'_{sS}} x_s \mu_{sk} \delta_{kj}$, $B'_j = B_j - \sum_{s,\ k \in r'_{sS}} x_s b_s(\mu_{sk}) \delta_{kj}$, $C'_S = C_S - \sum_s x_s \mu_{sS}$, and $B'_S = B_S - \sum_s x_s b(\mu_{sS})$.

The dual of this problem is to minimize the " Social Expenditure", that is,

$$\min_{\alpha', \beta'} G'(\alpha', \beta', \mu) = \sum_{j \in L} \alpha'_j(\pi) C'_j + \sum_{j \in L} \beta'_j(\pi) B'_j + \alpha'_S(\pi) C'_S + \beta'_S(\pi) B'_S \tag{5.38}$$

given x, w, μ, subject to

$$\begin{aligned}
\sum_{j \in L} \{\sum_{l \in r_{sA}} x_s \mu_{sl} \delta_{lj} \quad - \quad \sum_{k \in r'_{sS}} x_s \mu_{sk} \delta_{kj}\} \alpha'_j + \\
\sum_{j \in L} \{\sum_{l \in r_{sA}} x_s b_s(\mu_{sl}) \delta_{lj} \quad - \quad \sum_{k \in r'_{sS}} x_s b_s(\mu_{sk}) \delta_{kj}\} \beta'_j + \\
-x_s \mu_{sS} \alpha'_S \quad - \quad x_s b(\mu_{sS}) \beta'_S \geq \\
\nu_s exp(-w_{sA}) - \nu_s exp(-w_{sS}) \quad + \quad x_s w_{sA} - x_s w_{sS} \quad s \in S
\end{aligned} \tag{5.39}$$

Note that this "Social Expenditure" has a different formulation from the previous one, therefore different variable symbols are used, but the interpretation remains the same.

It is more efficient to solve the dual problem when the number of constraints in the dual problem is less than that of primal problem.

5.2.3 A comparison of game theoretic formulation and the weighted-sum approach

In looking back, we compare the differences between the game theoretic formulation and the weighted-sum formulation. Although in this Chapter the problem domain is in the ATM-based network and it is different in the previous Chapters, we approach the comparison from the angle of programming setups. By doing this, we may improve both methods and gain some insight into the nature of the routing problem in mixed-media networks.

1. One of the most important differences between the game theoretic formulation and the formulation in previous Chapters is that we use cost (and the constraints) models that are linear (e.g., network revenue, Social Welfare, G') or convex (e.g., G) in the game formulation. The linearity of the problem enables us to derive a simplified dual problem which is crucial in the formulation of the game. But in the previous Chapters, the cost models (and the constraints) are non-linear and do not have a property such as convexity.

2. The notion of economics equilibrium and gaming nature of the problem is quite general. Even for a non-linear programming problem, we can form a dual problem and use the game playing (minimax) procedure to achieve the saddle point as we shall proceed in next section using the non-linear problems of previous Chapters.

3. More resources are used in the game theoretic formulation which allow us to trade-off among these resources as the work in [43].

5.2.4 A game formulation of the original problem

In this section, we aim to formulate a new game to the routing problem in mixed-media networks. We examine the structure of the problem and reveal that the cost function is increasing in the arrival rate of each link. The arrival rate is a function of the users traffic requirements and the routing ratios and is increasing in the traffic requirement. Denote $x_{ij} = (g_{ij}, s_{ij}),\ x = \{x_{ij},\ \forall i, j\}$, where g_{ij} (s_{ij}) is the splitting factor of data traffic which specifies the fraction of data (voice) traffic, originating at node i and destined for node j, going through the ground sub-network. Denote $y_{ij} = (r_{ij}, , _{ij}),\ y = \{y_{ij}, \forall i, j\}$, where r_{ij} specifies, in packets/sec, the average rates of messages flowing between all possible SD pairs $[, _{ij}]$ specifies, in calls/min, the average rates of call requests between all possible SD pairs i and j. Therefore, we can formulate a two players dynamic game by the following.

$$\min_x \max_y f(x, y) = \min_x \max_y \sum_i A_i f_i(x, y) \tag{5.40}$$

subject to the capacity constraint where A_i is the weighting coefficients for objective f_i.

If the saddle point $(x^*, y^*$ for such game exists then

$$\min_x \max_y \sum_i A_i f_i(x, y) = \sum_i A_i f_i(x^*, y^*) = \max_y \min_x \sum_i A_i f_i(x, y) \tag{5.41}$$

subject to the capacity constraints.

Thus the two players can play the game as the users try to maximize the objective to increase their demand or the traffic requirement y and the network trys to minimize the objective the choosing x. What the users do is similar to the previous network economics game in which network try to produce the amounts of services that maximize its revenue. This will result in a saturation (or equilibrium) of the network capacity. The minimizing player which is the network will then choose the optimal routing ratios x to minimize the objective. Of course, in reality, the system is not in equilibrium and the users will not choose the requirement to saturate the network.

5.3 Access control and optimal routing

One advantage of our formulation of mixed-media networks by either a weighted sum approach or a "economic" approach is that we can easily extend the formulation into $N \geq 2$ subnets (or more than two fixed routings). Here the subnet means the medium that we take to transmit the traffic, namely, satellite or terrestrial. We can have other media to carry the traffic. For instance, a cellular network as another subnet to transport traffic for given SD pairs. This corresponds to a third path in our economic formulations.

Instead of formulating problems for many media, we consider a joint access control and optimal routing problem where access control is applied before the optimal routing is executed. To achieve access control, we can imagine for every SD pair, there is another "trash" path which is used to carry the "rejected" traffic when the capacity constraints are violated. Therefore, for the problem is weighted sum approach, we decide for every SD pair to carry a given traffic requirement how much should take the first subnet (path, i.e. the one utilize

terrestrial networks only), how much should take the second path (the one utilize the satellite channel), and how much should take the third "trash" path (how much should be rejected). We can also extend the number of different types of traffic other than voice and data if the corresponding performance measures of the types of traffic is available.

Note, this concept of "trash path" not only applies to the weighted-sum approach, but also applies to the minimax or the trade-off approach.

5.3.1 Weighted sum formulation

For an access control only problem, we can add a set $A = \{a_{ij}, \forall i, j\}$ of access control variables a_{ij} for every SD pair in equation (2.11) as

$$\lambda_{gd} = \sum_{ij} a_{ij}\gamma_{ij}[g_{ij}\sum_{k=1}^{n_{ij}} p_{ij}^k(\delta_{ij}^k)_l + \overline{g_{ij}}(\sum_{k=1}^{n_{i\sigma}} p_{i\sigma}^k(\delta_{i\sigma}^k)_l + \sum_{k=1}^{n_{\sigma' j}} p_{\sigma' j}^k(\delta_{\sigma' j}^k)_l)] \qquad (5.42)$$

How to perform access control now becomes a decision process to solve a non-linear programming problem with $0 \leq a_{ij} \leq 1$.

For joint access control and routing, we have to introduce one more set of variables for every SD pair. Let g_{ij} (s_{ij}) be the splitting factor of data traffic which specifies the fraction of data (voice) traffic, originating at node i and destined for node j, going through the ground sub-network. Let $\overline{g_{ij}}$ be the splitting factor of data traffic which specifies the fraction of data (voice) traffic, originating at node i and destined for node j, going through the ground and the satellite sub-network. Now $g_{ij} + \overline{g_{ij}} \neq 1$ if the trash path is used. Thus $1 - g_{ij} + \overline{g_{ij}}$ denotes the factor that must be rejected in order to satisfy the capacity constraints.

Define $x := (\lambda_{gd}, \lambda_{gv}, \Lambda_d, \Lambda_v)$ and reformulate the weighted sum problem

$$\text{minimize} \quad f(x)$$

subject to

$$\begin{aligned} 0 \quad &\leq \quad g_{ij}, \overline{g_{ij}} \leq 1, \quad 0 \leq s_{ij}, \overline{s_{ij}} \leq 1, \ \lambda_{gd} \leq C_{gd}, \ \lambda_{gv} \leq C_{gv}, \\ \Lambda_d \quad &\leq \quad C_{sd}, \ \Lambda_v \leq C_{sv}, \ 0 \leq g_{ij} + \overline{g_{ij}} \leq 1, \ 0 \leq s_{ij} + \overline{s_{ij}} \leq 1 \end{aligned} \tag{5.43}$$

By this formulation, the number of variables is doubled. That is $\overline{g_{ij}}$ ($\overline{s_{ij}}$) must be determined, since $\overline{g_{ij}} \neq 1 - \overline{g_{ij}}$. However, the rejected fraction can be easily computed by $1 - g_{ij} - \overline{g_{ij}}$.

Among all the possible access control schemes, one seeks the one that is "fair" which means that the rejected fractions shall be approximately equal for all SD pairs.

5.3.2 Service admission control

From the same line of reasoning, we can perform an admission control before the services are admitted. Once admitted, the aggregate demand will yield an allocation that is always compliant to the capacity constraints.

To do this we have to introduce one more set of decision variables for every SD pair. Let U_s be the splitting factor of data traffic which specifies the fraction of aggregate demand for service type s connection going through the ground sub-network. Let $\overline{U_s}$ be the splitting factor of data traffic which specifies the fraction of aggregate demand for service type s connection, going through the ground and the satellite sub-network. Now $U_s + \overline{U_s} \neq 1$ if the trash path is used. Thus $1 - U_s + \overline{U_s}$ is the factor that must be rejected in order to satisfy the

capacity constraints. Denote $\overline{U} = \{\overline{U_s} = 1 - U_s,\ s \in S\}$, the (primal) problem for joint access control and routing becomes the following given x, w, μ:

$$\max_{0 \le U,\ \overline{U} \le 1} \sum_s \{[\nu_s exp(-w_{sA}) + x_s w_{sA}]U_s + [\nu_s exp(-w_{sS}) + x_s w_{sS}]\overline{U_s} \tag{5.44}$$

subject to

$$\sum_{s,\ l \in r_{sA}} U_s x_s \mu_{sl} \delta_{lj} \quad + \quad \sum_{s,\ k \in r'_{sS}} \overline{U_s} x_s \mu_{sk} \delta_{kj} \le C_j \quad j \in L \tag{5.45}$$

$$\sum_{s,\ l \in r_{sA}} U_s x_s b_s(\mu_{sl}) \delta_{lj} \quad + \quad \sum_{s,\ k \in r'_{sS}} \overline{U_s} x_s b_s(\mu_{sk}) \delta_{kj} \le B_j \quad j \in L \tag{5.46}$$

$$\sum_s \overline{U_s} x_s \mu_{sS} \le C_S \tag{5.47}$$

$$\sum_s \overline{U_s} x_s b(\mu_{sS}) \le B_S \tag{5.48}$$

$$0 \quad \le \quad U_s + \overline{U_s} \le 1 \quad s \in S \tag{5.49}$$

Chapter 6

Conclusions and Future Research

We have considered an important issue: dynamic routing in both voice/data-integrated and ATM-based hybrid networks. We have also extended the problem to a joint routing and access control problem. This is a problem domain requiring system integration concepts and theory.

Much of the work carried out here does not represent the end points of the research. On the contrary, this work indicates possible extensions to other related problems. Thus, it postulates continuing research efforts in the applications of the results already obtained. For example, we have assumed a fixed topology throughout this dissertation. However, doing dynamic routing in a large heterogeneous network with a changing topology is an extremely challenging problem. Especially, the topology may change due to random faults or new configurations. Re-routing under faults is a very interesting problem, yet little work has been done on it. Thus, the changing topology problem is a challenging one.

The solutions to the problems we outlined are not the only solutions. What we presented is something that needs to be designed and implemented. For instance, the Kalman filter techniques can be replaced by many other estimation techniques which might lead to more accurate and faster computation. Among

those, learning tools, and intelligent control in such applications are important. To name a few, using the Learning Automata to do the routing control has been proposed in [49]. Neural Networks can be trained to estimate the arrival rates and Fuzzy Control can be used to deal with events in uncertain conditions. A self-organization map can be used to learn of traffic management [2], etc.

The networks under consideration could be varied for different circumstances. One important case is the incorporation of wireless networks, mobile computing, or Global Personal Communication Networks, all of which are now in the design and planning stage. As we can see, the integration of such networks is essential. Hence, it remains an interesting and significant task for all of us to work on such a global integration.

Bibliography

[1] N. Abramson. "Packet Switching With Satellites". In *Proc. Nat. Computer Conf.*, pages 696–702, 1973.

[2] N. Ansari and Y. Chen. "A Neural Network Model to Configure Maps for a Satellite communication network". In *IEEE GLOBECOM'90*, pages 1042–1046, 1990.

[3] G. Ash. "Use of a Trunk Status Map for Real-Time DNHR". In *Proc. ITC-11*, pages 795–801, 1985.

[4] G. Ash. "Design and Control of Network with Dynamic Nonhierarchical Routing". *IEEE Commun. Mag.*, vol. 28(no. 10):pp. 34–40, Oct. 1990.

[5] G. Ash and S. Schwartz. "Network Routing Evolution". In *Proc. Network Management Control Conf*, pages 357–368, 1990.

[6] H. Ash, B. Blake, and S. Schwartz. "Integrated Network Routing and Design". In *Proc. ITC-12*, pages 640–647, 1989.

[7] T. Basár. *Dynamic Noncooperative Game Theory.* Academic Press, 1982.

[8] G. Benelli, E. Del Re, R. Fantacci, and F. Mandelli. "Performance and Uplink Random-access and Downlink TDMA Techniques for Packet Satellite Networks". *Proc. IEEE*, vol. 72(no. 11):pp. 1583–1593, Nov. 1984.

[9] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, 1987.

[10] R. K. Brayton. *Sensitivity and Optimization.* Elsevier Scientific Publishing, 1980.

[11] D. Cantor and M. Gerla. "Optimal Routing in Packet-Switched Computer Networks". *IEEE Trans. Comput.*, vol. 37(no. 11):pp. 1062–1069, Oct. 1974.

[12] J. Chang and L. Lu. "High Capacity Low Delay Packet Switching via Processing Satellite". In *Proc. Int. Commun. Conf. '82*, pages 1E.5.1–1E.5.5, 1982.

[13] S. Chen and J. S. Baras. "Optimal Routing in Mixed Media Networks with Integrated Voice and Data Traffic". In *Proc. GLOBECOM'92*, pages pp. 335–339, 1992.

[14] M. de Prycker. *Asynchronous Transfer Mode: Solution for Broadband ISDN.* Ellis Horwood, 1991.

[15] L. Fratta, M. Gerla, and L. Kleinrock. "The Flow Deviation Method: An Approach to Store-and-forward Communication Network Design". *Networks*, vol. 3:pp. 97–133, 1973.

[16] R. Gallager. "A Minimum Delay Routing Algorithm Using Distributed Computation". *IEEE Trans. Commun.*, vol. COM-25(no. 1):pp. 73–85, 1977.

[17] C.B. Garcia and W.I. Zangwill. *"Pathways to Solutions, Fixed Points, and Equilibria"*. Prentice-Hall, 1981.

[18] B. Gavish and S. Hantler. "An Algorithm for Optimal Route Selection in SNA Networks". *IEEE Trans. Commun.*, vol. COM-31(no. 10):pp. 1154–1161, Oct. 1983.

[19] A. Gelb. *Applied Optimal Estimation*. the MIT Press, 1974.

[20] M. Gerla, W. Chou, and H. Frank. "Cost Throughput Trends in Computer Networks Using Satellite Communications". In *Proc. Int. Conf. Commun.*, pages 21c1–21c5, Minneapolis, MN, 1974.

[21] M. Gerla and R. Pazos-Rangel. "Bandwidth Allocation and Routing in ISDN's". *IEEE Commun. Mag.*, vol. 22(no. 2):pp. 16–26, Feb. 1984.

[22] A. Gersht and S. Kheradpir. "Integrated Traffic Management in Sonet-Based Multi-service Networks". In A. Jensen and V.B. Iverson, editors, *Teletraffic and Datatraffic in a Period of Change, ITC-13*, pages pp. 67–72. Elsevier Science Pub., 1991.

[23] A. Gersht, S. Kheradpir, and Anda Friedman. "Real-Time Traffic Management by a Parallel Algorithm". *IEEE Tran. Commu*, vol. 41(no. 2):pp. 351–361, 1993.

[24] A. Gersht and A. Shulman. "Optimal Routing in Circuit Switched Communication Networks". *IEEE Trans. Commun.*, vol. COM-37(no. 11):pp. 1203–1211, Nov. 1989.

[25] A. Girand. *Routing and Dimensioning in Circuit-Switched Networks*. Addison-Wesley, 1990.

[26] A. Harel. "Convexity Properties of the Erlang Loss Formula". *Operations Research*, vol. 38(no. 3):pp. 499–505, 1990.

[27] J.R. Haritsa, M.O. Ball, N. Roussopoulos, A. Datta, and J.S. Baras. "MANDATE: MAnaging Networks Using Database Technology". Technical Report TR 92-98, Inst. Systems Research, U. Maryland, 1992.

[28] D. Huynh, H. Kobayashi, and F. Kuo. "Optimal Design of Mixed-Media Packet-Switching Networks: Routing and Capacity assignment". *IEEE Trans. Commun.*, vol. COM-25(no. 1):pp. 156–187, Jan. 1977.

[29] K. Johannsen. "Code Division Multiple Access Versus Frequency Division Multiple Access Channel Capacity in Mobile Satellite Communication". *IEEE Tran. Veh. Tech.*, vol. 39(no. 1):pp. 17–26, Feb. 1990.

[30] C. R. Johnson. *Lectures on Adaptive Parameter Estimation*. Prentice Hall, 1988.

[31] S. Kheradpir. "PARS: A Predictive Access-Control and Routing Strategy for Real-Time Control of Telecommunication Networks". In *Proc. Network Management and Control, ed. by A. Kershenbaum*, pages 389–413, 1990.

[32] L. Kleinrock. *Communication Nets: Stochastic Message Flow Delay*. McGraw-Hill, 1964.

[33] L. Kleinrock. *Queueing Systems*, volume 1. Wiley, 1975.

[34] L. Kleinrock. *Queueing Systems*, volume 2. Wiley, 1975.

[35] L. Kleinrock. "The Latency/Bandwidth Tradeoff in Gigabit Networks". *IEEE Commun. Mag.*, vol. 30(no. 4):pp. 36–0, Apr. 1992.

[36] K. Krishnan. "Markov Decision Algorithms for Dynamic Routing". *IEEE Commun. Mag.*, vol. 28(no. 10):pp. 66–69, 1990.

[37] K. Krishnan and T. Ott. "State-dependent Routing for Telephone Traffic: Theory and Results". In *Proc. 25th Control Decis. Conf. (CDC)*, pages 2124–2128, 1986.

[38] K. Krishnan and T. Ott. "Forward-Looking Routing: A New State-Dependent Routing Scheme". In *Proc. ITC-12*, pages 1026–1031, 1989.

[39] I. Lambadaris. *Admission Control and Routing Issues in Data Network.* PhD thesis, Dept. of Elect. Eng., U. of Maryland, 1991.

[40] J. Lee. "Symbiosis Between a Terrestrial-based Integrated Services Network and a Digital Satellite Network". *IEEE Sel. Areas. Commun.*, vol. SAC-1(no. 1):pp. 103–109, 1983. Space communication.

[41] I. Lin, E. Geraniotis, and W. Yang. "Joint Voice Scheduling and Data Routing in Networks via Iterative Methods". In *Proc. 25th CISS*, pages 417–422, Johns Hopkins U., 1991.

[42] S. Low and P. Varayia. "A Simple Theory of Traffic and Resource Allocation in ATM". In *GLOBECOM'91*, pages 1633–1637, 1993.

[43] S. Low and P. Varayia. "Service Provisioning in ATM Networks". In *Interdisciplinary Workshop on Coordination and Complexity*, 1993.

[44] D. Luenberger. *Linear and Non-Linear Programming.* Addison-Wesley, 1984.

[45] B. Maglaris, R. Boorstyn, S. Panwar, T. Spirtos, P. O'Reilley, and C. Jack. "Routing of Voice and Data in Burst Switched Networks". *IEEE Trans. Commun.*, vol. COM-38(no. 6):pp. 889–897, June. 1990.

[46] K. Maruyama. "Optimization of Mixed-Media Communication Networks". *Computer Networks*, vol. 2:pp. 168–178, 1978.

[47] J. McQuillan, I. Richer, and E. Posen. "The New Routing Algorithm for the ARPANET". *IEEE Trans. Commun.*, vol. COM-28(no. 5):pp. 711–719, May 1980.

[48] Y. Morihiro, S. Okasaka, and H. Nakashima. "A Dynamic Channel Assignment and Routing Satellite and Digital Networks-DYANET". In *Proc. GLOBECOM'87*, pages 10.1.1–10.1.5, 1987.

[49] K.S. Narendra and M.A.L. Thathachar. "On the Behavior of a Learning Automaton in a Changing Environment with Application to Telephone Traffic routing". *IEEE T. on system, Man, and Cybernetics*, vol. SMC-10(no. 5):pp. 262–269, 1980.

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

[51] M. Schwartz. *Telecommunication Networks: Protocols, Modeling and Analysis.* Addison Wesley, 1987.

[52] S. Scotti. "Structural Design Using Equilibrium Programming". Technical report, NASA-TM-107593, 1993.

[53] M. Soronshnejad and E. Geraniotis. "Multi-Access Strategy for Integrated Heterogeneous Mixed-Media Packet Radio Networks". In *Proc. 1990 Conf. Int. Sciences Systems (CISS)*, 1990.

[54] A. Tanenbaum. *Computer Networks, 2nd ed.* Prentice-Hall, 1988.

[55] S. Tasaka. "Multiple-access Protocols for Satellite Packet Communication Networks: a Performance Comparison". *Proc. IEEE*, vol. 72(no. 11):pp. 1573–1582, Nov. 1984.

[56] A. Tits. *Classnotes of Optimal Control.* Electrical Eng. Dept., U. Of Maryland, 1993.

[57] H.R. Varain. *Microeconomic Analysis, 2nd edition.* W.W. Norton & Company, 1984.

[58] I. Viniotis. *Optimal Control of Integrated Communication Systems via Linear Programming Techniques.* PhD thesis, Dept. of Elect. Eng., U. of Maryland, 1988.

[59] Y. Watanabe and H. Mori. "Dynamic Routing Schemes for International ISDNs". In *ITC seminar*, pages 299–308, 1988. Routing in ISDN.

[60] C. Wu and V. Li. "Integrated Voice and Data Protocols for Satellite Channels". In *Proc. Mobile Satellite Conference, Jet Propulsion Lab.*, pages 413–422, May 1988.

[61] S.F. Wu, S. Mazumdar, and S. Brady. "On Implementing A Protocol Independent MIB". In *Proc. second IEEE Network Control and Management Workshop*, 1993.

[62] J. Yuan. The Control of Multi-Media Communication Processors: For Messages with Time Constraints. Master's thesis, Dept. of Elect. Eng., U. of Maryland, 1988.

[63] W. I. Zangwill and C. B. Garcia. "Equilibrium Programming: The Path following Approach and Dynamics". *Math. Prog.*, vol. 21:pp. 262–289, 1981.

[64] J. Zhou and A. Tits. "User's Guide for FSQP Version 2.4". Technical Report TR-90-60r1, University of Maryland, 1992.