JB 93-03

Entitled:

"A Dynamic Routing Algorithm in
Mixed Media Networks with
Integrated Voice and Data Traffic"

Authors:

(with Shihwei Chen)

1993 Conference on Information Sciences and Systems

The Johns Hopkins University
Baltimore, Maryland

March 1993

A Dynamic Routing Algorithm in Mixed Media Networks with Integrated Voice and Data Traffic *

John S. Baras † and Shihwei Chen
Institute for Systems Research
and Electrical Engineering Department
University of Maryland
College Park, MD 20742

Submitted to 1993 CISS as a **regular** paper

Abstract

In this paper, we consider mixed-media networks with multi-media traffic to find the optimal splitting ratios between satellite network and terrestrial network for data and voice traffic dynamically. We constantly monitor the traffic and measure the arrival rate and occupancy of every link. Based on these data, we optimize data traffic delay in a suitable time-frame under the constraints of voice traffic blocking probabilities of voice transmission links being less than specified values, which is up to system's design or users' requirements. A dynamic routing algorithm is presented.

1 Introduction

As networking technologies evolve at a steady and growing speed, the services available to users become more diversified. Multi-media information transmission is the goal of the next evolution of networks. Voice, data, images, graphics, etc., will be sent through networks at gigabits per second with high quality at users' demand. In addition to the traffic being heterogeneous, the transmission media can be hybrid. For example, the path of transmission can consist of a combination of ground and space routes. This kind of hybrid networks are the so-called mixed-media networks. Routing and flow control in such networks are important functions in achieving the goal of providing new services to users.

There have been several related works in the routing design of a mixed-media network. Huynh et al. [6] 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 sub-network. The first assumption makes their capacity assignment problem mathematically tractable and thus a closed-form solution was obtained using analytic procedures involving Lagrange multipliers. The second assumption eliminates the routing problem within the terrestrial sub-network and reduces the flow assignment problem to one of determining the optimal amount of traffic which goes through satellite links for each pair of source/destination nodes.

*Research supported by NASA grant NAGW-2777.

†Martin Marietta Chair in Systems Engineering.

More recently, Yuan and Baras focused on the control of multimedia [1] communication processors where the transmission of messages has a time constraint given by the user [15]. They approximated the delay distribution by that of a product-form network model. Using a DFT algorithm, they computed a time index from the delay distribution so that 99 percent of messages arrived before the index. From this index, they derived the "Gittin index" which guides the processor to switch the packets to the subnetwork with the minimum discounted cost (delay).

However, these papers considered only one uniform transmission and switching mode, i.e., only 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 the performance measure for voice traffic. In the telecommunication network, the blocking probability of the voice transmission is the major concern for phone-call establishments 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 [5] 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 [10].

A similar type of objective function is used by Viniotis-Ephremides. They use the theory of Markov decision processes and dynamic programming to obtain the optimal admission and routing strategy at a simple ISDN node [13]. The results are characterized by the same series of "switching curves". Results in the same vein have been obtained by Lambadaris-Narayan for a circuit-switched node [9]. However, these results are limited to a system with low degree of dimension, that is, a network of one or two queues (or a simple ISDN node) .

In our previous paper [3], we have obtained the optimal routing ratios for such mixed-media networks with integrated voice and data traffic using this type of objective function. However, the physical meaning of the weighting coefficients is not clear, except that a larger weight on delay means that data traffic is more of concern than voice traffics; smaller weight on delay means that data traffic is of less concern. For example, there is not much difference in the physical meanings between two systems with weighting coefficients of 7 and 10 respectively. Furthermore, since the delay could be as large as infinity while the upper limit for blocking probability is one and these two measures have different units, the linear combination of these two functions is likely to be biased. Therefore, we must choose the weights carefully. Nonetheless, we don't know any systematic procedures to find the weighting coefficients so far.

Due to these reasons, we feel that a different approach is necessary. Sometimes, the designer of the system may not want to optimize both data delay and voice blocking probability at the same time. The designer may just want to optimize the voice blocking probability or residual capacity without sacrificing the data traffic too much. Alternately the data traffic may have some real time constraints to meet. In these cases, it is more appropriate to optimize the performance objective for voice traffic under the constraints on data traffic link delays. Using similar arguments, we may alternately want to guarantee some link or trunk performance of voice traffic while optimizing the data traffic delay. In the former case, we must use a single objective (average network delay) and the constraints are that 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 network. In either case, we have a non-linear programming

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

problem with non-linear constraints [4].

However, the aforementioned previous work [4] dealt with a static traffic situation, which is hardly suitable for a real application. To be more practical, we must present a dynamic routing algorithm and modify the model to mimic the real system and to satisfy the needs of real-time control requirements of the system and to react to the dynamic nature of the real network system. Nonetheless, theses previous efforts paved a backbone structure to solve the problem of dynamic routing.

In this paper, we assume that there is a Central Control Center (CCC) which is responsible for the collection of network data from local switches (nodes) and the computation of the optimal ratios for the mixed-media network. CCC can communicate with local switches and send the result (ratios) to local switches. Local switches control the traffic based on these ratios so as to achieve the optimal performance at the particular time-frame. Local switches are also responsible for measuring the voice and data arrival rates and send them to CCC at the beginning of each time-frame.

For the completeness of the presentation, we present here the model used in our experiments which is the same as in the authors' previous work [3]. The algorithm is described following the problem formulation.

2 Model

A mixed-media network could comprise several sub-networks. However, to simplify the problem, we consider a communication network composed of two sub-networks: one is the ground subnet, the other the satellite subnet. A mixed-media network with more than two sub-networks will be considered in the future, though the problem would become harder. The nodes are the locations of interface message processors (IMPs) linked together by landlines. There are special nodes called satellite IMPs (SIMPs) 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 of the input traffic at SIMPs between ground and satellite subnets; (2) routing on 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 ground routing procedure, and link capacities, we want to minimize the average system delay of data traffic under the constraints that the voice blocking probabilities of the ground links and the satellite channel be less than the specified performance requirements. We can consider other performance measures to be optimized depending on the application. For example, we may want to maximize the residual capacities of trunk groups with (or without) the delay constraints. Or we may minimize capital cost which is to be defined in different cases. What we present here is a general algorithm which can be used in different applications.

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. In the following, we will derive the overall objective function.

2.1 Data delay in the ground

Suppose that we are given a data sub-network in the ground of N nodes linked by L ground links of capacities C_{dl} (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 arrival processes at different nodes, independence of service times at successive nodes, then we have the following expression for the average data delay in the ground [8],

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

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 sub-network, λ_{dl} = the traffic rate on link l, and $\frac{1}{\mu_d}$ = the average length of a packet.

2.2 Voice blocking in the ground

Suppose that we are given a telephone sub-network in the ground which may use the same transmission links and switching facilities of the data sub-network in the ground. This voice sub-network has N_1 nodes linked by L_1 trunks (links) of capacities $C_{vl}, l = 1, 2, ..., L_1$ in a specified topology. The capacity C_{vl} of link i can be divided into N_{vl} channels. A traffic rate matrix $[\Gamma_{ij}]$ 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 $M/M/N_{vl}/N_{vl}$ system and the average blocking probability P_b is [1]

$$P_b = \frac{1}{\Gamma}\sum_{l=1}^{L_1}\lambda_{vl}P_l, \qquad P_l = \frac{(\lambda_{vl}/\mu_v)^{N_{vl}}/N_{vl}!}{\sum_{n=0}^{N_{vl}}(\lambda_{vl}/\mu_v)^n/n!} \tag{2}$$

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

2.3 Satellite channel

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 Λ_v calls/min (excluding the retransmission due to collision). The overall transmission rate (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 $1/\mu_v$ minutes.
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.

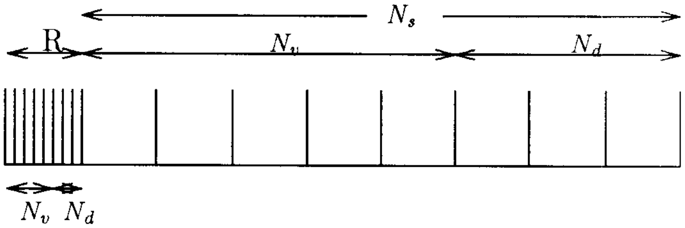

Figure 1: The satellite channel assignment

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

2.3.1 Data delay in the satellite channel

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. 1.

Define p_{suc} to be the probability that a data packet will be successfully transmitted via a data channel. Then p_{suc} is $\frac{\Lambda'_d}{m}Te^{-\frac{\Lambda'_d}{m}T}$ (see [12] page 430). Here we assume that 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 m (could be 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 random access data (RAD) protocol is [14]:

$$D_{RAD}(m) = [1.5 + S + (e^{\frac{\Lambda'_d}{m}T} - 1)t_{rx}]T \qquad (3)$$

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 $\Lambda_d = \Lambda'_d e^{-\Lambda'_d T}$ [12]. 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 has 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.

$$D_{DAD} = [\frac{1.5}{n} + 2S + \frac{\Pi'(1)}{\Lambda'_d \frac{T}{n} e^{-\frac{\Lambda'_d}{m}\frac{T}{n}}} + (e^{\frac{\Lambda'_d}{m}\frac{T}{n}} - 1)t_{rx}]T \tag{4}$$

where $\Pi'(1)$ is the average queue length [2], 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". Therefore, it is not suitable for real time traffic.

2.3.2 Voice blocking in the satellite channel

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

$$P_s = \frac{\Lambda_v}{\Gamma} P_v, \quad P_v = \frac{(\Lambda_v/t_v)^{N_v}/N_v!}{\sum_{k=0}^{N_v} (\Lambda_v/t_v)^k/k!} \tag{5}$$

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

$$\pi_v(n) = \frac{(\frac{\Lambda_v}{\mu_v})^n/n!}{\sum_{k=0}^{N_v} (\frac{\Lambda_v}{\mu_v})^k/k!} \tag{6}$$

2.3.3 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 are the same as in sections 2.3.1 and 2.3.2.

2.3.4 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 larger than call request and set-up time which includes propagation delay and random retransmission delay in the satellite channel. The average packet delay is

$$D_{MB} = \sum_{k=0}^{N_v} \pi_v(k) d_{data}(N_v - k) \tag{7}$$

where d_{data} is obtained from either Equation 3, or 4, 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.

2.3.5 Average delay in the satellite channel

The overall average delay D_s in the satellite channel is

$$D_s = \frac{\Lambda_d}{\gamma} D_d \tag{8}$$

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.

3 Optimization problem

We are given blocking probabilities of all links as constraints to minimize the average delay. The overall objective function which we want to minimize is thus given as

$$\begin{aligned} f(\lambda_{dl}, \lambda_{vl}, \Lambda_d, \Lambda_v) &= D_g + D_s \\ \text{subject to} \quad 0 \le g_{ij} \le 1, \quad 0 \le s_{ij} &\le 1, \quad P_l \le c_l, \ \forall l, \ P_v \le c_s \end{aligned} \tag{9}$$

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 satellite link, and c_s is the bound of it. As we pointed out earlier, the objective function and constraints can be changed depends on the applications.

4 The algorithm

The algorithm is an adaptation of the algorithm described in [7].

1. During the interval [**t**-T,**t**), the voice and data arrival rates of each link are measured by the local switches.
2. At time **t**, the measurements are sent to CCC.
3. At the CCC, we solve the optimization problem defined by Equation 9.
4. The results are sent to the appropriate local switches.
5. During [**t**,**t**+T), the switches send the incoming traffic to satellite channel accordingly.

4.1 Discussion

The measurements of the arrival rates in the previous interval are not the exact arrival rates for the present interval, but they are good estimates. We can apply Kalman's filtering techniques to smooth out the estimates [7].

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

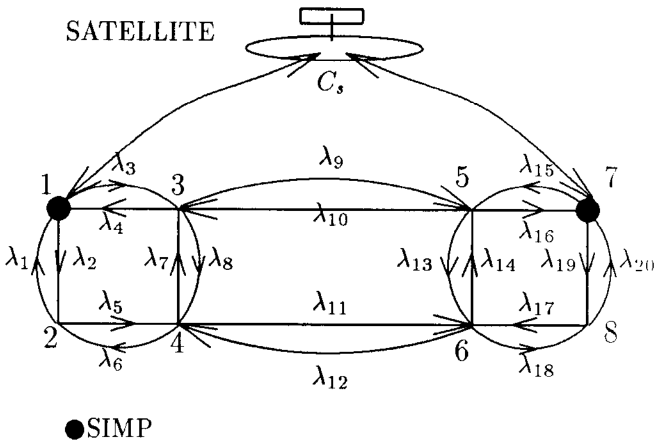

Figure 2: A mixed-media network

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.

5 Numerical example

The optimization problem can be solved by different algorithms. Among those are primal-dual methods [10, 11], and Sequential Quadratic Programming (SQP) [16]. We have used the FSQP (Feasible SQP) subroutines developed by J. Zhou and A. Tits at the University of Maryland.

[Example 1:] We consider a network topology taken from [6]. In this example, we consider a ground sub-network capable of transmitting and switching both voice and data traffic through the same IMPs and transmission links. This network has eight nodes (IMPs) and 20 links as shown in Fig. 2. In this network, there are two regions consisting of nodes {1, 2, 3, 4} and {5, 6, 7, 8} respectively. The regional SIMPs are located at nodes 1 and 7, which are also IMPs. 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.

For movable boundary, we assume the following parameters of the system: the number of voice channels in the satellite is 25 and the number of data channels is 10, ground capacity in each link is 5 channels/link.

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) $1 \rightarrow 2 \rightarrow 4 \rightarrow 6 \rightarrow 8$.

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 into links l_8 and l_9 equally. Using this ground sub-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

Table 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 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.			

2), 1/4 over path 3), and 1/2 over path 4).

The results that CCC can get are given in Tables 1 and 2.

6 Conclusion

We have proposed a dynamic optimal routing algorithm for the hybrid networks with both voice and data traffic using an optimization based approach with multiple criteria. With a suitable update time-frame, we are able to adjust to the dynamics of the networks and find optimal switching ratios accordingly.

Acknowledgment: We would like to thank Drs. J. Zhou and A. Tits for the use of the powerful FSQP package.

References

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

[2] 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.

[3] 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.

[4] S. Chen and J. S. Baras. "Tradeoff Curves in Mixed Media Networks with Integrated Voice and Data Traffic". In *(submitted to) Proc. GLOBECOM'93,* 1993.

[5] 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.

[6] 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.

[7] S. Kheradpir. "PARS: A predictive access-control and routing strategy for real-time control of telecommunication networks”. In *Proc. Network Management Control*, pages 389–413, 1990.

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

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

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

[11] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: Algorithms and complexity.* Prentice-Hall, 1982.

[12] M. Schwartz. *Telecommunication networks: Protocols, modeling and analysis.* Addison Wesley, 1987.

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

[14] 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.

[15] 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.

[16] J. Zhou and A. Tits. "User's guide for FSQP Version 2.4”. Technical Report TR-90-60rle, University of Maryland, 1992.