TECHNICAL RESEARCH REPORT

Service Integration in Next Generation VSAT Networks

by M. Hadjitheodosiou

CSHCN T.R. 97-29
(ISR T.R. 97-76)

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/

SERVICE INTEGRATION IN NEXT GENERATION VSAT NETWORKS

Michael H. Hadjitheodosiou

Center for Satellite & Hybrid Communication Networks
University of Maryland, College Park
MD 20742, USA

Tel:+301-405-7904 Fax:+301-314-8586
e-mail: michalis@isr.umd.edu

Abstract

Very Small Aperture Terminal (VSAT) satellite networks have so far been successful in the provision of specific communication services to geographically dispersed users. However, user demands are becoming more complex, and VSAT networks are expected to provide a much wider range of services (voice, data and multimedia). We investigate how this service integration could be achieved and show that performance improvements are possible if efficient multi-access protocols and speech compression with voice activity detection techniques are used. We also discuss the future role VSATs could play in the provision of access to the Integrated Broadband Communications Network to remote users. We discuss the possibility of using VSATs for ATM service provision. The need of careful consideration of the advantages and limitations of using VSAT networks for this type of service is discussed Finally, we highlight a method for dynamic bandwidth allocation in a broadband satellite network.

Keywords: Satellite networks; Very Small Aperture Terminal (VSAT); voice/data integration; multiple access; dynamic bandwidth allocation; ATM.

SERVICE INTEGRATION IN NEXT GENERATION VSAT NETWORKS

1. Introduction

Very Small Aperture Terminal (VSAT) networks (Fig. 1) have so far been successful in the provision of very specific communication services to geographically dispersed users. The combination of new, more powerful satellites, a number of recent technological innovations that resulted in the decrease of the station size and cost, and a global deregulation of the telecommunication industry, should make VSAT systems even more attractive service providers in the future, capable of supporting a wide range of two-way integrated telecommunication services [MARA95].

However, user demands are becoming more complex, and communication networks are expected to provide a much wider range of services on top of the original packet data service. A number of technical issues need to be resolved so that these new services become financially competitive. The development of an adaptive, dynamic protocol for this wide range of services that will result in the most efficient allocation of the space segment, one of the most expensive commodities in the service provision, is a critical requirement. Recent technology, such as the development of cheap, low-bit-rate vocoders with voice activity detection (VAD) could further improve the efficient use of the channel, and exciting new innovations, such as DirecPC [AROR96] offering direct access to the internet, need to be accommodated.

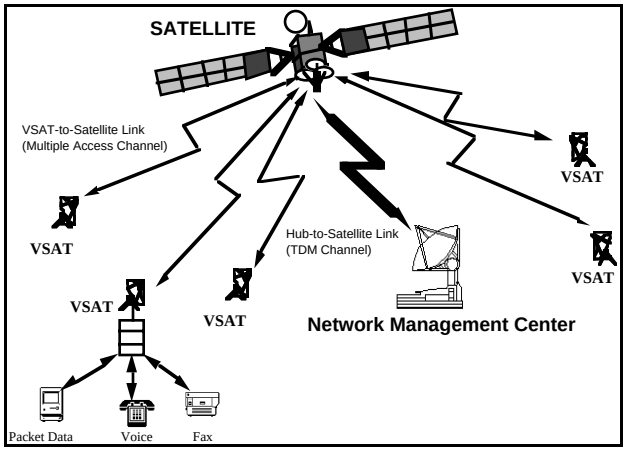

Fig.1 A Typical VSAT Network

In this paper, we look at the state-of-the-art in VSAT networking and focus on a discussion of how this service integration could take place and the performance improvements that could be achieved. We then extend our discussion on the possibility of using VSATs for the provision of access to the integrated broadband communications network to remote users. Finally, we highlight a method for dynamic bandwidth allocation in a broadband satellite network.

2. Traffic Source Models

2.1 Interactive Data Transmission

It is well known that generation of data from a single data source is well represented by a Poisson arrival process(continuous time) or by a geometric inter-arrival process(discrete time). A single packet is generated at each time. This could be either a fixed length packet or a variable length packet, with length represented by a certain distribution of fixed mean. (Neg. Exponential, Tailed Exponential, Geometric and Uniform are the most commonly used). Without loss of generality we can define such a distribution and use it as an example.

This traffic is of bursty nature, relatively short in length, and requires relatively small delay in transmission. Delay variance is not a major problem, but error free transmission is an important requirement. Examples of such traffic for a VSAT scenario are transaction/credit card verification, hotel/airline reservations and various short data message transmissions.

2.2 Voice Sources

Modelling of a voice source is much more complex, mainly because of the strong correlation among arrivals. We start by examining a rather simplified model.

Observations on the nature of speech show that a speech source creates a pattern of active talkspurts and silent gaps [BRAD68]. There are principal spurts and gaps related to the talking, pausing and listening patterns of a conversation. There are also “mini-gaps” and “mini-spurts” due to the short silent intervals that punctuate continuous speech.

Statistical analysis on a number of conversations shows that the “active” period covers only (approx.) 40% of the time, while 60% of the time consists of a mix of long and short silences. Therefore, by using a speech activity detector close to a speech source one can distinguish between active and silent parts in a conversation, and allow reuse of the channel when a silence is detected. (this is called Digital Speech Interpolation and has been extensively used in telecommunications). With a “slow”

detector one can distinguish between active periods and long silences(2 states) while with a "fast" detector 3 states can be observed: active, long silence, short silence.

The following assumptions apply:

- The arrival process of new voice calls and the distribution of their duration can be characterized by a Poisson process and by an exponential distribution, respectively.
- All spurts and gaps have exponentially distributed durations.

Talkspurts are referred to as ON states and silences as OFF, and they appear in turn.

The transition from ON to OFF occurs with probability γ, and the transition from OFF to ON occurs with probability σ. In a discrete time case, ON and OFF periods are geometrically distributed with the mean $1/\gamma$, $1/\sigma$ respectively.(Fig.2).

Packets are generated during the ON period according to a Bernoulli distribution with rate λ. No packets are generated in the OFF state.(The continuous time equivalent case can be represented by an exponential distribution using a Poisson process.)

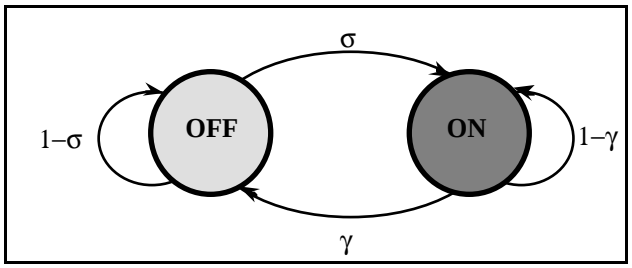

Fig. 2 Simplified 2-stage Markov Model for Speech

3. Voice/Data Integration

The link from the hub station to the VSATs is usually configured using conventional Time Division Multiplexing (TDM). The multiple access link from the VSATs to the hub however has been subjected to a greater degree of variation, and debate continues about the most suitable and efficient choice for particular traffic demands. The main difficulty in satellite access protocols is the long propagation time of the geostationary satellite link which can impose unacceptably long co-ordination times.

A number of multiple access protocols for this type of network have been proposed in recent years, ranging from contention/random access-based schemes such as variations of the ALOHA protocol [RAYC87] or other Collision Resolution Algorithms [HADJ95] to systems with no contention (e.g. static TDMA or Single Channel Per Carrier (SCPC) allocations), and also combinations of the two, such as Dynamic Reservation TDMA.

Different classes of traffic have different performance requirements. Stream-type traffic (voice calls, file transfer) usually require a collision-free allocated channel, while random access transmission could work well with small data messages and bursty traffic. It is therefore necessary to develop an access scheme that accommodates all traffic classes in an efficient way. One way of treating this problem is develop a combined random access / channel reservation protocol and try to make it adaptive to changing traffic mix.

In this case we consider a traffic scenario consisting of three types of service:

1. *Small Size Data Message Transmission:* Typically single packet messages. These could be updates of the value of a particular quantity (e.g. share price) or specific requests from a central database (e.g. number of items in stock).

2. *File Transfer Transmission:* Connections for file transfer or other relatively long data transmission (e.g. complete list of prices for items on sale).

3. *Voice Calls:* Typically business calls of short duration (e.g. a mean call time of 120 sec), using low-bit-rate coded voice for efficient use of channel capacity.

In this paper we will not focus so much on the selection or performance optimisation of a random access protocol (we assume that one has already been selected and optimized) but on techniques for dynamic allocation of capacity to different traffic types so that we maximize the efficient use of the (expensive) satellite channel.

3.1 Dynamic Channel Allocation

Most existing VSAT networks handling speech use dedicated channels which can be pre-defined or allocated on demand. One way to optimize the overall network performance would be to develop an efficient dynamic allocation scheme.

There are a number ways the satellite bandwidth can be allocated, and the service the network must provide determines which is the best choice. There are also other factors that affect the choice of allocation scheme, such as the complexity of implementation and the network architecture.

Some of the channel allocation options are listed below, in an order of increasing complexity:

- The bandwidth is divided into identical sub-channels, and a fixed number of these is allocated for random access and reservation requests while the remaining sub-

channels are available for stream traffic connections. The partitions are determined by the expected traffic load. A special case of this would be to define a single random access channel for reservations and make the rest available for connections.

• There are no predetermined reservation, voice or data message channels. Reservations take place over multiple channels, and the partition between the channels is dynamically determined according to changes in the traffic load. All channels are identical.

• Same as before, but message channels are not identical and do not have a predetermined bandwidth. The required bandwidth is reserved for each transmission.

This dynamic access is a reservation scheme that dynamically re-assigns reservation channels for optimum performance. The protocol operation is split in two parts:

• *Channel Reservation:* Users with long data messages or voice calls transmit their requests in this part of the protocol. A random access protocol operates and a collision resolution scheme is used to resolve possible collisions of the request packets. The choice of the access scheme could be an important factor in the protocol's overall performance.

• *Message/Call Transmission:* Successful requests enter a global queue (operating in a FCFS mode) for message or voice call channels, and when a suitable channel becomes available it is allocated for the duration of the transmission.

3.2 Protocol Operation

The satellite channel of capacity B is subdivided into N equal size channels, each having a capacity B/N . There are now N_r reservation, N_v voice channels and N_d data channels , such that $N_r + N_v + N_d = N$ (Fig.3). Note that the terms "*channel*" and "*sub-channel*" do not necessarily represent frequency allocation. Channel division may take place in either the time or frequency domain, or even using Spread Spectrum techniques.

3.2.1 Packet Data Message Service

A message reservation request contains information about the message's origin and destination ID, type (Short Data Message/File Transfer) and length. When a VSAT has a message to transmit, it monitors the status channel to find the identities of the current reservation channels and sends a reservation request over one of these. The Network Management Center (NMC), located at the Hub, assigns the N_m message channels to the VSATs whose reservation requests were successful. It also keeps track of how long each customer has been in service at each channel and how many requests are waiting to be served. Using this information, it can estimate what the backlog is at each channel at a specific time.

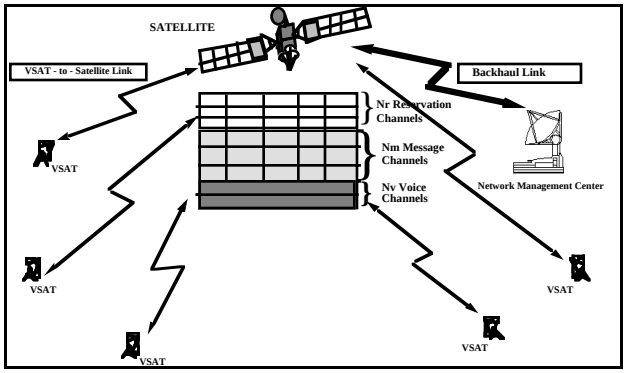

Fig. 3 VSAT Network Architecture

The VSAT then times out and waits for an Acknowledgment (ACK) from the NMC. If positive acknowledgment is not received within a pre-set period, the VSAT retransmits the request after a random waiting time. For slotted systems a guard time is used to compensate for the propagation delay difference for VSATs at different ends of a beam coverage.

When the Hub receives a reservation request, it assigns the message channel with the smallest backlog to it, calculates the time before backlog is cleared and returns ACK to the transmitting VSAT. The ACK contains the ID of the VSAT, the ID of the assigned message channel, and the holding time, d_h, if known. The VSAT can start transmission on the assigned channel at the allocated time.

3.2.2 Voice Calls

If a successful voice call request arrives at the NMC and there are voice channels available, the NMC will assign the call to one of these and return an acknowledgment to the call originator. The ACK contains the ID of the originator and the IDs of the respective assigned inbound and outbound channels. For two-way voice communications and a star network architecture, voice channels have to be assigned in pairs, one from the outbound channels and one from the inbound channels. However, the inbound is much less congested than the outbound so we can assume that for every free outbound channel there is always a free inbound channel. The situation is simpler in a mesh system, where after the call set-up is completed, VSATs can do directly to each other. A voice request arriving when all call channels are busy will get a busy signal and will be cleared from the system (switching system with pre-determined blocking probability). There is always a trade-off between the data delay constraints and the call blocking probability, and this trade-off will be examined next.

A very important performance parameter that has to be optimized is the channel partition $N_r : N_v : N_m$. The Hub must dynamically re-assign voice, data and reservation channels from the available channel pool so that the blocking probability of the voice traffic and the total average delay of the data traffic are minimized. It therefore has to estimate the voice message arrival rate Λ_v and the data message arrival rate Λ_d, by counting the number of successful requests for each type of message that have arrived in a given time interval. For a heavy traffic load, there should be more message channels so that the backlog is cleared, and fewer reservation channels, so that there is a limit on the system input. On the other hand, if traffic is light, more reservation channels should be allocated for optimum performance.

The analysis that follows will determine the optimal partition for a given set of Λ_v and Λ_d. The Hub can use such information to periodically select the optimum operation point for the system, and inform the VSATs of all channel re-assignments. In order to avoid any possible confusion, a waiting period, typically one round-trip propagation delay, is used before a re-assignment is activated.

3.3 Performance Analysis

In this analysis of the protocol's performance, the delay characteristics for different message arrival rates are obtained and the channel allocation ratio $\alpha^*=N_r/N_m$. is optimized to minimize end-to-end message delay. The optimum channel allocation ratio then defines the optimum system delay-throughput performance characteristic.

3.3.1 Model Assumptions

The following assumptions are made in the analysis of the protocol's performance:

1. Stations generate Poisson traffic with rates Λ_v calls/s and Λ_d messages/s.
2. There are N active VSATs and they randomly choose between the N_r reservation channels, for either a voice call or a data message transmission.
3. Each VSAT carries either one active message or one call at a particular time.
4. All reservation, message, and acknowledgment channels are error free (i.e. error rates are negligible).
5. Processing delays are negligible compared to the channel propagation delay.
6. Binary feedback information (Collision/No-collision) about the channel state can be transmitted to all users.
7. Both data and voice channels are assigned in a FCFS way. Voice channels operate in a "blocked calls cleared" mode.
8. Successful reservation requests arrive at the Hub in a Poisson manner, with rate $\Lambda_v+\Lambda_d$. (Λ_v,Λ_d are both i.i.d. Poisson R.V.'s).
9. Arriving data messages consist of a reservation part of length τ and an information part of length β. For simplicity in the analysis, the reservation part is assumed to consist of one packet, with length equal to the channel slot size in slotted systems.
10. There is no restriction on the distribution of the voice call duration.

3.3.2 Reservation Channel Delay

This is the time from the generation of a new reservation request to the time it arrives successfully at the NMC. The mean Reservation Delay, D_r, is the mean random access delay over the satellite channel (which depends on the multiple access /collision resolution scheme used).

3.3.3 Message Channel Delay

Successful reservation requests enter a global queue, identical to the arrival process of the reservation requests at the Hub, except for a time shift of d_p slots. The message channel delay is therefore given by

$$D_m = d_B + d_g + d_T + d_p$$

where :

d_B = Queueing delay for the message in the global queue,

d_p = Channel propagation delay,

d_g = Guard time to compensate for differences in propagation delay,

d_T = Transmission delay of message on satellite channel.

In this protocol, the NMC assigns the channel with the minimum backlog to an arriving reservation request. This minimum-wait queueing system is equivalent to the standard *M/G/s* queue. There are exact expressions for the expected waiting time in an *M/G/s* system however a very good approximate expression is given in [BOXM79], with a maximum error of 3%. Using this, d_B is given by

$$d_B = \frac{1}{2}\left[1+\frac{1}{3}\left(\frac{\delta}{\beta}\right)^2\right]\frac{W_M}{N_D(s)+A(s)[1-N_D(s)]}$$

where:

$$N_D(s)=\sum_{i=0}^{5}A_i(s)\rho^i,$$

$$A(S)=\left(\frac{s+1}{s-1}\right)\left[\frac{1+\frac{1}{3}\frac{\delta^2}{\beta}}{1+\frac{1}{2s+1}\left(\frac{\delta}{\beta}\right)^{s+1}}-1\right]$$

$N_D(s)$ can be recursively computed using the method discussed in [LI 84], and W_M is the waiting time of an *M/M/s* queue:

$$W_M = \beta \frac{\dfrac{1}{(1-\rho)^2}\dfrac{1}{s}\dfrac{(s\rho)^s}{s!}}{\left[\displaystyle\sum_{k=0}^{s-1}\frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!}\frac{1}{(1-\rho)}\right]}$$

where: ρ = traffic intensity per channel = $\lambda\beta / s$
β = the average service time of the M/G/s system

The overall mean message delay is then given by

$$D = D_r + D_m$$

where D_m and D_r were defined earlier.

3.3.4 Voice Channel Blocking Probability

Once a voice channel is assigned, it remains occupied until either side of the conversation hangs up, and information about this status change arrives at the Hub. In this way, a total of $2d_p$ seconds is wasted for each voice channel re-assignment. Therefore, any successful voice request will hold the channel for an average time of

$$D_h = (2d_p + d_g + d_v)$$

where d_v is the average call duration and the d_g the guard time discussed earlier.

This is an M/G/s/s, s-server loss system, in which a blocked call is cleared from the queue. In such a system the blocking probability depends only on the mean of the service time [SCHW85] and is given by the Erlang-B formula:

$$P_B = \frac{(s\rho_v)^s/s!}{\displaystyle\sum_{k=0}^{s}(s\rho_v)^k/k!}$$

where

$$\rho_v = \frac{\Lambda_v}{s}[\, d_v + 2d_p + d_g \,]$$

The behaviour of the total average message delay D and the blocking probability P_B can be studied as a function of the total input traffic and the channel partition ratios.

3.3.5 Protocol Stability

For the system to remain stable the following two conditions must be satisfied:

(i). The random access protocol operating in the reservation channels must be stable.
(ii). The global queue operating in the message channels, must remain stable. Therefore, the channel utilization must always be less than unity, i.e.,

$$\frac{\lambda}{N_m}\left(d_g + d_T\right) \le 1$$

3.3.6 Numerical Results & Discussion

Using the delay expressions defined earlier, the performance characteristics for the network are computed. The behaviour of the average data message delay and the voice call blocking probability are studied as a function of the total input traffic and the channel partition ratios.

We can thus estimate the maximum call arrival rate the system can handle for a mean call handling time and a specific number of voice channels. By plotting the end-to-end mean data message delay for various data arrival rates and a particular voice call blocking probability the optimal channel allocation ratio a = (No of Reservation Channels / No of Message Channels) for a particular traffic load can be determined (Fig.4).

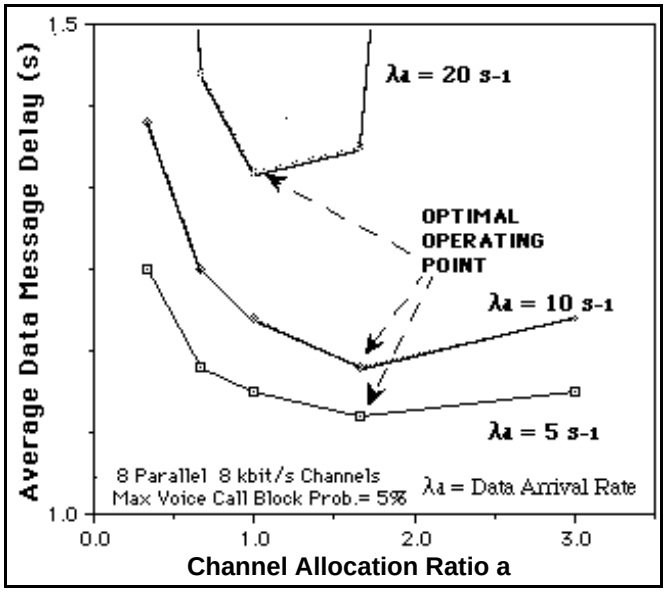

Fig. 4 Optimal Operating Point for Channel Allocation

There is clearly a need to optimize the channel allocation (i.e. number of channels allocated to voice calls or random access reservations) to suit the traffic mix. If this mix is known in advance we can adjust the allocation of reservation and voice channels accordingly. If however the traffic mix changes considerably, it would be beneficial if we could adjust this allocation at regular time intervals. By defining a maximum blocking probability for voice calls, we can estimate the data

message performance for various loads, and determine the channel allocation ratio that provides "optimum" performance. The Network Management Center can then periodically update the channel allocation based on the traffic load and inform the stations accordingly.

3.4 Compression & Multiplexing Techniques

We can further maximize the efficient use of the satellite channel by employing sophisticated coding schemes in the case of multimedia transmission. These usually require a considerable processing time (tens of milliseconds) and need a specific internal frame time for packet arrangement management. It is obvious that the Quality of Service required would determine the required minimum bit rate.

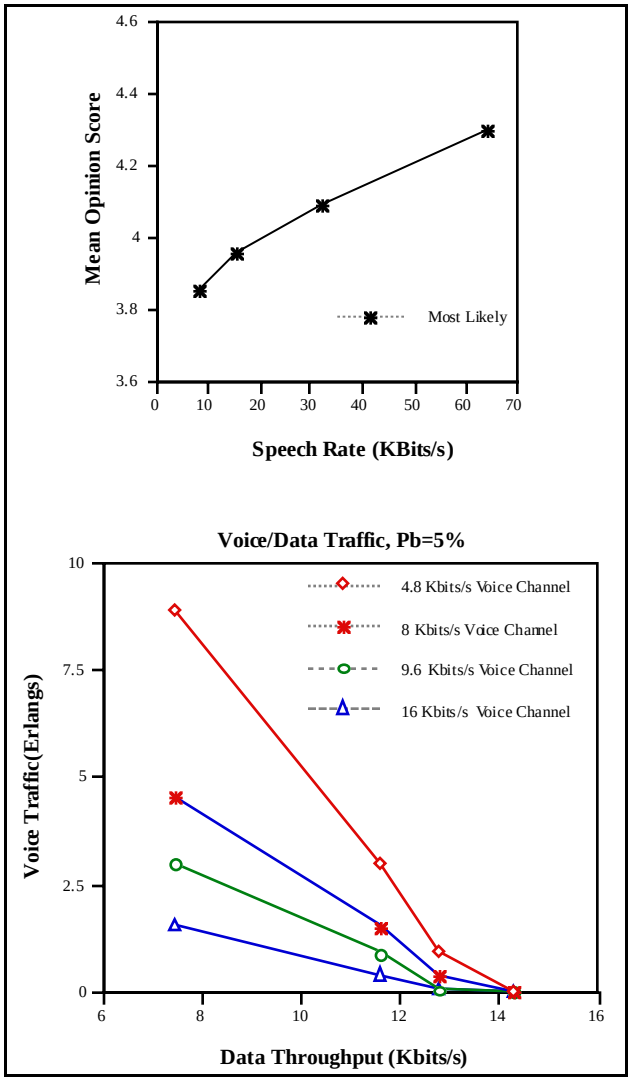

Fig. 5 Mean Opinion Score Degradation and possible Data Throughput increase with Increase of Speech Compression Rate

For speech, one such coding algorithm is the *code-excited linear predictive* (CELP) coding [SHRO85] which has been shown to produce good quality speech at bit rates below 16 kbit/s. The basic CELP algorithm applies vector quantisation of the excitation signal to achieve efficiency in coding, a technique described as *fractional bits per sample coding* . However, the high complexity of CELP and its relatively low robustness to transmission errors means that the basic algorithm has to be extensively revised to meet the constraints of a VSAT application. Due to the high propagation delay the speech signals experience, the speech codec implementation should also reserve processing capacity for echo cancellation [KOND93].

It is also possible to use a speech coder with multi-rate capability with a VSAT system. In such a system, if only packetized speech needs to be transmitted, a higher rate codec can be switched on to provide better quality speech. The arrival of data messages at the VSAT could trigger a switch of the speech to a lower rate codec that will allow the multiplexing of data packets on the same channel, using the bandwidth that becomes available from the higher speech compression, at the expense of lower speech quality.

Fig. 5 shows some experimental results of the degradation of the speech quality, expressed in Mean Opinion Score (MOS) values (5: Excellent, 1: Extremely Bad), as we increase the speech compression rate. Contrasting these with the possible increase in data (or additional voice call) throughput possible if we use a higher compression rate we can see that we can accommodate a significantly higher throughput for only a 5 or 10% degradation of the speech quality perceived by listeners. Therefore we can dynamically use a multirate codec to avoid congestion in cases where traffic load is higher, while we can revert to better quality service when we have less traffic.

3.5 Using Speech Silences for Data Transmission

Having investigated the problem of sharing a common satellite channel amongst a large number of VSATs, we next turn our attention on maximizing the efficient use of this resource. We look at the possibility of integration of voice and data on a second level, over the same channel.

Since a speech source creates a pattern of active talkspurts and silent gaps by using a speech activity detector close to a speech source one can distinguish between active and silent parts in a conversation, and allow re-use of the channel when a silence is detected. With a "slow" detector one can distinguish between active periods and long silences (2 states) while with a

"fast" detector 3 states can be observed: active, long silence, short silence.

If we have a low bit rate vocoder with voice activity detection [GRUB81], assuming this takes 4 to 5 30 ms speech frames for a silence detection to take place, we can take advantage of the silence intervals that are longer than 2 frames to transmit data packets from the same source. This will help us make a more efficient use of the channel, and will reduce the transmission delay of long data files. These can be broken into smaller fixed size packets, given a sequence number and transmitted during these silences. The waiting time before the data transmission is completed, although should ideally be as low as possible, is not a critical limitation in this type of file transmission.

The threshold level needs to be set to a very low value to avoid missing large portions of speech at the beginning of a talkspurt and at low speech levels. This of course makes the system susceptible to high noise levels, which is one of the problems that need to be taken into account. In order to eliminate the possibility of cutting out speech mid-bursts a further condition is applied, by adding a hangover stage to the VAD output.

The operation of the VAD is based upon these basic assumptions:

- Speech is a non-stationary signal. Its spectral shape usually changes after short periods of time, typically 20-30 ms.
- Background noise is usually stationary during much longer time periods and it changes very slowly with time.
- The speech signal level is usually higher than the background noise level, otherwise speech is unintelligible.

Based on these assumptions, a VAD algorithm can be developed that can detect silence gaps and distinguish background noise (with or without speech). Assuming that in most VSAT systems the background noise is relatively low, a simple fixed energy threshold can be used to detect the silence regions (unlike mobile systems, where there is a high and variable noise environment that needs to be compensated by a more adaptive algorithm). Various implementations of VAD systems for CELP coding can be found in the literature [KOND93].

3.6 Performance Improvements from Voice/Data Integration

We assume a VSAT network using a reservation multiple access scheme to allocate channels to incoming voice calls. We isolate a typical channel and focus on optimizing the efficiency of our system by taking advantage of the silent intervals of speech. We initially assume that once a voice call has been successful in obtaining a channel it will be the only call on the channel and it can use it for the duration of the conservation, but, during the detected silent intervals, data packets from messages waiting for transmission from the same VSAT terminal can be send through this channel.

An important factor affecting the speech quality is the variance in delay the speech packets experience. In the case of a satellite network there is no complicated routing process that can introduce a big variation in the delay packets experience, however, because the channel propagation delay is so high, problems might arise even if there is a small variance in the delay. The introduction of a hangover has the additional advantage of mitigating the variable delay impairment by increasing the mean duration while reducing the rate of talkspurts. In general, fairly large delays can be accommodated provided a sufficiently long hangover is used [GRUB81].

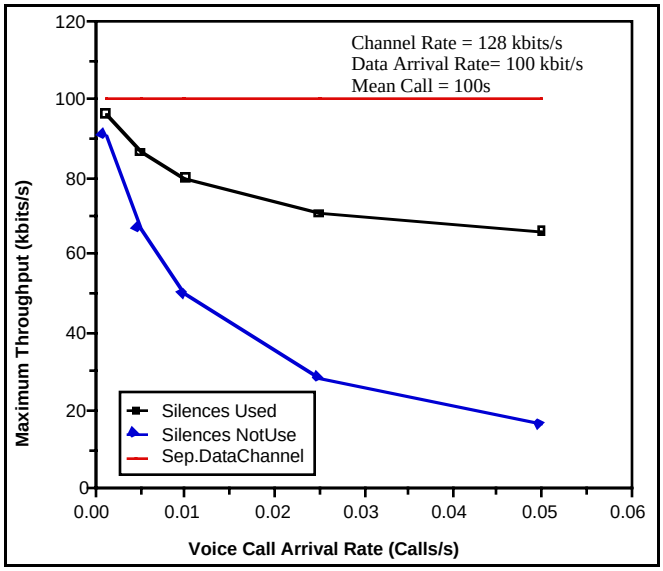

Fig.6 Maximum Throughput and Performance Improvements for different voice loads when data is transmitted during long silences

Fig. 6 shows a comparison between the maximum possible throughput for data messages for three cases:

- No voice activity detection, arriving data messages have to wait for a voice call to be completed before using the channel.
- Data is transmitted during the longer silence intervals
- There is always a second channel available for data transmission when channel is busy with voice call.

Details of the analysis can be found in the APPENDIX. Clearly the first and third cases represent the extreme boundaries, and for a practical system with VAD the performance is somewhere between the two. Significant improvements are possible however if VAD is used effectively.

4. Provision of ATM Service via satellite

Communication satellites can be used to provide multimedia services over a large area, without the need of excessive investment in the early phase of ATM network deployment, especially to areas where the terrestrial network infrastructure is not very well developed. They can offer complementary service to terrestrial networks, especially for widely dispersed users, and the broadcast nature of satellites makes them an ideal choice for point-to-multipoint transmissions. New users can be accommodated swiftly by simply installing new earth stations at customer premises therefore network enlargement is not a significant planning problem.

Satellite networks can play two main roles:

- the interconnection of a few geographically distributed broadband networks, usually called "broadband islands" [HADJ94].
- the provision of a network interface to a large number of thin-route users.

For the first scenario large earth station gateways would be required to accommodate the high and usually continuous traffic rates that would be expected. For the second scenario, VSAT-type networks could be used. The 155 and 622 Mbit/s transmission rates conventionally associated with ATM are well above the maximum rates possible with today's VSAT technology. However, in practice, most individual users will usually require significantly lower traffic rates, especially if there are only a few data or voice terminals located at a remote location. This large number of users with bursty traffic will need a cost-efficient way to occasionally access the ATM network.

A number of high data rate satellite systems have recently been proposed, planning to offer Fixed Satellite Services to large numbers of globally dispersed users. Since using a TDMA-type multiple access enables the support of multiple users and a variety of high-speed communication services and a significant number of these systems will probably employ a TDMA multi-access scheme for their uplink. However, as satellite capacity is a very expensive commodity, and as, unlike traditional satellite systems currently in operation, the services that these new high data rate satellite systems will be required to offer are so varied in requirements, it would be necessary to develop a close to optimal bandwidth allocation policy that maximizes the utilization of the satellite capacity and provides certain guarantees on performance, based on the nature of the traffic mix we have. This should ideally be able to offer services to users in an integrated manner, support different classes of service and adapt to changing traffic loads.

4.1 System Model

4.1.1 Network Architecture

A satellite network consisting of a satellite in GEO orbit, S gateway stations and a Network Management Center (NMC) is considered. Each station carries traffic from a variety of traffic sources such as voice calls, data messages or video. A framed TDMA multiple access scheme is used for the uplink and a TDM multiplexing operates in the downlink.

4.1.2 Frame Structure

Fig. 7 shows the frame structure we consider for the Time Division Multiple Access (TDMA) scheme operating on the uplink. Transmission is organized in fixed size, slotted frames. The slot(s) in the first part of the frame is dedicated for transmission of information on the station status , while all other N slots are used for information transmission. Movable boundaries exist between the different types of service (voice, video and data) and the objective is to optimize the boundary positions on a frame-by-frame basis.

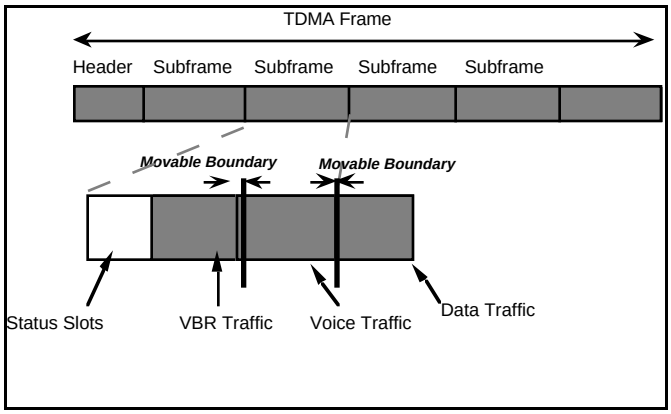

Fig. 7 TDMA Frame Structure

For simplicity assume that the slot size is equal to the smallest single-packet and that there are (N+1) slots in each frame.

If the voice rate is R_a bits/s then the length of the voice packet is $L_a = R_a T$ bits. If R_c and b are the channel bit rate and the number of bits the channel can carry in the duration of a slot, then

$$N+1=\left\lfloor \frac{R_c T}{L_a} \right\rfloor = \left\lfloor \frac{R_c}{R_a} \right\rfloor$$

$$b=\frac{R_c T}{N+1} bits$$

According to the requirements for the High Data Rate ground terminal in the ACTS system, the TDMA frame duration is T= 32 msec, and the TDMA time slot is equal to 32 microseconds. (this is equivalent to N=1000 slots).

There is clearly a need for a close-to-optimal bandwidth allocation policy that maximizes the utilization of the satellite capacity but also provides certain guarantees on performance, based on the nature of the traffic mix we have. This should ideally be able to offer services to users in an integrated manner, support different classes of service and adapt to changing traffic loads. Apart from voice and data, other services such as internet connection or VBR video, are of considerable interest to potential service providers and could probably be a major factor in enabling the deployment of these systems in the future.

4.2 Optimal Capacity Allocation

The concept of optimal capacity allocation in communication networks (and satellite systems in particular) is not new and has been extensively studied in a variety of forms [AEIN77][CONT96] in the past. Optimal capacity allocation studies for the particular case of integrated voice/data at a packet switched TDM system have also been done [WU 92].

The objective of the optimal capacity allocation analysis is to find a stationary policy which assigns the available slots to the various traffic types, based upon the activity characteristics of video and voice traffic as well as the data packet backlog at the gateway buffers, by minimizing a cost function at each frame. A policy is called *stationary* if the action chosen at a frame only depends on the state at the frame. The cost function is defined as the weighted sum of the distortion rate of video traffic, the packet dropping probability of voice traffic, and the unassigned data packets in the gateway buffers.

The optimization policy is based on movable boundary(ies) among the slots assigned to voice and data users (or VBR Traffic in the general case)(Fig. 7). The system operates in a frame format and the channel capacity (bandwidth) is allocated at each frame, which depends upon the activity characteristics of the real-time traffic as well as the number of data packets in the buffer of gateways. The performance measures evaluated are the packet dropping probability and the throughput of the video and voice traffic, and the delay, packet loss rate, and throughput of data traffic.

A Markov Decision Process (MDP) could be employed to solve the optimal capacity allocation problem by minimizing a proposed cost function, defined in this case as the weighted sum of the packet dropping probability of voice traffic, and the unaccessed rate of data traffic at each frame time [YANG94]. The weighting factors can be adjusted according to the requirements in real operation and the traffic mix. A value-iteration algorithm is applied to this MDP. The size of a frame can be fixed or variable. Basically, the transmission assumes an Asynchronous Transfer Mode (ATM) system with a frame structure, in which the smallest packet size could be the fixed size ATM cell.

Since voice cannot tolerate large random delay, voice packets are discarded without retransmission. However, the packet dropping probability of voice traffic can be still within a tolerable range. The data users transmit (and retransmit, if necessary) their packets over allocated slots within a frame. If there are not enough slots for data packets, these are stored in buffer for transmission in the coming frames. A group of active terminals is integrated over a broadband satellite channel, which has N + 1 slots per frame. The first slot called "status slot" is used for bandwidth allocation information and other information. The other N slots are allocated for the information transmission of these active terminals. The interesting problem raised here is how to allocate the bandwidth for the group of terminals while meeting their performance requirements.

Earlier work in [GHAF93], [YANG94] focused on the development of an optimal capacity allocation policy for a similar system. Our next objective is to revisit the problem of dynamic capacity allocation for a TDMA satellite system, and develop an implementable, near-optimal allocation algorithm. This should take into account the propagation delay of the satellite link. Our aim is to derive a near-optimal version of the dynamic bandwidth allocation method, that operates on a single frame-by-frame basis, and then try to extend this operation over M multiple frames. Using analysis and simulation, we plan to compare the performance advantages of the near-optimal allocation to a system that does not use a bandwidth allocation optimization, and weigh this against the processing delay and additional complexity cost introduced by this process.

5. Summary & Further Work

We have focused on recent work on the development of adaptive, dynamic schemes that will result in the most efficient allocation of the space segment. In order to achieve the highest possible utilization of the satellite bandwidth, it is not enough to develop an efficient resource allocation scheme. Various compression and multiplexing techniques can also be employed to ensure the most economical bandwidth use. The development

of low-bit-rate vocoders, enabling a large number of voice users to share the limited bandwidth, and the use of Voice Activity Detection to take advantage of idle intervals in conversations, could be incorporated and thus further improve the efficient use of the channel.

Finally, as VSAT networks need to be made compatible with the developing terrestrial Integrated Broadband Communication Network and to extend to areas such as network interconnection and provision of ATM-based services, we have addressed some of the problems that need to be resolved and provided some suggestions of the important role VSATs could play in the ATM era. We have seen that the link capacity and the window size could have severe implications on the system's performance, and there are advantages to be gained by a more efficient sharing of the satellite resource.

Work is currently under way on a near-optimal bandwidth allocation policy for a TDMA satellite system that maximizes the utilization of the satellite capacity and provides certain guarantees on performance, based on the nature of the traffic mix we have.

REFERENCES

[AEIN77] Aein, J., Kosovych,O. "Satellite Capacity Allocation" Proc. of the IEEE, **Vol. 65** (3), March 1977.

[AROR96] Arora,V., Suphasindu,N., Baras, J., Dillon, D., Asymmetric Internet Access over Satellite-Terrestrial Networks, *Proc. 16th AIAA Int. Comm. Satellite Systems Conf.*, Washinton,DC, 1996.

[BOXM79] Boxma, O.J., Cohen, J.W. , Fuffels, N., "Approximations of the Mean Waiting Time in an M/G/S Queueing System", *Oper. Research*, **V27**(6), Nov-Dec 1979, pp.1115-1127.

[BRAD68] Brady, P.T., "A statistical analysis of On-Off patterns in 16 conversations", *Bell Systems Tech.Journal*, Jan. 1968, pp. 73-91.

[CONT96] Conti, M., Gregori, E. "Analysis of Bandwidth Allocation Schemes for transmission of VBR video traffic ona FODA Satellite Network", IEE Proc. Commun. Vol. I43 (1), February 1996.

[FISC79] Fischer, M.J., "Data Performance In A System Where Data Packets Are Transmitted During Voice Silent Periods-Single Channel Case.", *IEEE Trans. Comms.*, **COM-27**(9),Sept.1979, pp. 1371-1375.

[GHAF93] Ghaffari, B., Geranitis, E. "Voice, Data And Video Integration for Multi-Access in Brodaband Satellite Networks", CSHCN TR 93-15/ ISR TR 93-15, University of Maryland, College Park, 1993.

[GRUB81] Gruber, J.G., "Delay Related Issues in Integrated Voice and Data Networks", *IEEE Trans. Comms*, **COM-29**(6), June 1981, pp. 786-800.

[HADJ94] Hadjitheodosiou, M.H *et al.*, "Broadband Island Interconnection via Satellite- Performance Analysis for the CATALYST Project", *Int. Jrnl. Sat. Comms.*, **V12**(3), May 1994, pp. 223-238.

[HADJ95] Hadjitheodosiou, M.H., Coakley, F.P., "Adaptive Multiple Access Protocols for VSATs Providing Voice/Data Services and ATM Interconnection", *Proc. 10th Int.Conf. Digital Satel. Comm. 95 (ICDSC-10)*, Brighton, May 95.

[KOND93] Kondoz, A.M., Evans, B.G., " A High Quality Voice Coder with Integrated Echo Canceller and Voice Activity Detector for VSAT Systems", *Proc. 3rd Eur. Conf. Sat. Comms.*, 1993, pp.196-200.

[LI 84] Li, V.O.K., Yan, T.Y., "Adaptive Mobile Access Protocol (AMAP) for the Message Service of a Land Mobile Satellite Experiment (MSAT-X)", *IEEE Transactions on Vehicular Technology*, **VT-33**(3), August 1984, pp. 237-243.

[MARA95] Maral, G., "VSAT Networks", *Wiley Series in Comms.& Distributed Systems*, John Wiley & Sons, 1995.

[RAYC87] Raychaudhuri, D., Joseph, K., "Ku-Band Satellite Networks using VSATs-Part1:Multi-access Protocols", *Int. Jrnl. Sat. Comms.*, 1987, pp. 195-212.

[SCHW85] Schwartz,M., "Telecommunication Networks: Protocols, Modelling and Analysis", *Addison-Wesley*, ISBN 0-201-16423-X, 1987.

[SHRO85] Shroeder, M.R., Atal, B.S., "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", *Proc. ICASP-85*, pp. 1649-1652.

[WEIN78] Weinstein, C.J., "Fractional Speech Loss and Talker Activity Model for TASI and for Packet Switched Speech", *IEEE Trans. Comms.*, **COM-26**(8), August 1978, pp. 1253-1257.

[WU 92]Wu, G., Mark, J.W., "Capacity Allocation for Integrated Voice/Data Transmission at a Packet Switched TDM", *IEEE Trans. Comms*., **COM-40** (6), June 1992.

[YANG94] Yang, W.B., Geraniotis, E., "Dynamic Bandwidth Allocation in Broadband Satellite Networks" Proc. 14th ISCSC, AIAA, San Diego, 1994.

APPENDIX

Performance when data is transmitted during silent intervals (Single Channel Case)

A single channel case is considered, in which the voice calls have priority over the data packets, in that data packets are transmitted only when there are no voice calls present in the system, or the voice conversation is in a long silent period. We assume that the arrival process of voice calls and data packets are independent Poisson processes, with parameters λ_1 and λ_2, respectively. We assume the voice call and data packet lengths are exponentially distributed with means μ_1^{-1} and μ_2^{-1}, respectively. No queue is allowed for voice calls(arriving calls are cleared if a voice call is present) and an infinite queue is allowed for data packets.

We also assume that a voice call begins with a talkspurt, and the length of the talkspurts is exponentially distributed with mean α^{-1}. At the end of a talkspurt there could be one of two types of silence, type 1 - (short) , with probability Π and type 2 - (long) with probability $(1-\Pi)$. The silence lengths are also exponentially distributed , with means β_1^{-1} and β_2^{-1}, respectively. All these random variables are assumed to be statistically independent and the voice conversation can end in either activity or silence. Data packets can be transmitted at all times when a voice call is not present and when there is a voice call in type 2-(long) silence intervals. Voice calls always have priority over data packets, and new calls arriving pre-empt any data packets being transmitted. The pre-empted packet is placed at the head of the data queue, and packets are selected in a FCFS mode when the channel becomes available for data transmission.

Let Q_V and Q_D be the steady state number of voice calls and data packets in the system. Furthermore, in steady state define V to be the random variable representing the status of the voice call. We then have:

$$V = \begin{cases} 0 - NoCall \\ 1 - Talkspurt \\ 2 - Silence(Short) \\ 3 - Silence(Long) \end{cases}$$

Define $P_{i,j} \equiv \Pr\{V = i, Q_D = j\}$, for i = 0,1,2,3 and j = 0,1,... .
The steady state equations for $P_{i,j}$ become, for j = 1,2, ... :

$$\begin{aligned} (\lambda_1+\lambda_2)P_{0,0} &= \mu_1 P_{1,0} + \mu_1 P_{2,0} + \mu_1 P_{3,0} + \mu_2 P_{0,1} \\ (\lambda_1+\lambda_2+\mu_2)P_{0,j} &= \mu_1 P_{1,j} + \mu_1 P_{2,j} + \mu_1 P_{3,j} + \mu_2 P_{0,j+1} + \lambda_2 P_{0,j-1} \end{aligned} \qquad (\alpha.1)$$

$$\begin{aligned} (\lambda_2+\alpha+\mu_1)P_{1,0} &= \lambda_1 P_{0,0} + \beta_1 P_{2,0} + \beta_2 P_{3,0} \\ (\lambda_2+\alpha+\mu_1)P_{1,j} &= \lambda_1 P_{0,j} + \beta_1 P_{2,j} + \beta_2 P_{3,j} + \lambda_2 P_{1,j-1} \end{aligned} \qquad (\alpha.2)$$

$$\begin{aligned} (\lambda_2+\beta_1+\mu_1)P_{2,0} &= \alpha\Pi P_{1,0} \\ (\lambda_2+\beta_1+\mu_1)P_{2,j} &= \alpha\Pi P_{1,j} + \lambda_2 P_{2,j-1} \end{aligned} \qquad (\alpha.3)$$

$$\begin{aligned} (\lambda_2+\beta_2+\mu_1)P_{3,0} &= \alpha(1-\Pi)P_{1,0} + \mu_2 P_{3,1} \\ (\lambda_2+\beta_2+\mu_1+\mu_2)P_{3,j} &= \alpha(1-\Pi)P_{1,j} + \lambda_2 P_{3,j-1} + \mu_2 P_{3,j+1} \end{aligned} \qquad (\alpha.4)$$

For $|z| \le 1$, define $$P_i(z) = \sum_{j=0}^{\infty} P_{i,j} z^j \qquad (\alpha.5)$$

Then, from equations $(\alpha.1)$-$(\alpha.4)$ we have a system in matrix form

$$A(z)P(z) = B(z) \qquad (\alpha.6)$$

with

$$A(z)=\begin{bmatrix} a_0(z) & -\mu_1 z & -\mu_1 z & -\mu_1 z \\ -\lambda_1 & a_1(z) & -\beta_1 & -\beta_2 \\ 0 & -\alpha\Pi & a_2(z) & 0 \\ 0 & -\alpha(1-\Pi)z & 0 & a_3(z) \end{bmatrix}$$

$$P(z)=\begin{bmatrix} P_0(z) \\ P_1(z) \\ P_2(z) \\ P_3(z) \end{bmatrix}, B(z)=\mu_2(z-1)\begin{bmatrix} P_{0,0} \\ 0 \\ 0 \\ P_{3,0} \end{bmatrix}$$

and

$$\begin{aligned} a_0(z) &= -\lambda_2 z^2 + (\lambda_1+\lambda_2+\mu_2)z - \mu_2 \\ a_1(z) &= -\lambda_2 z + \lambda_2 + \alpha + \mu_1 \\ a_2(z) &= -\lambda_2 z + \lambda_2 + \beta_1 + \mu_1 \\ a_3(z) &= -\lambda_2 z^2 + (\lambda_2+\beta_2+\mu_1+\mu_2)z - \mu_2 \end{aligned}$$

Therefore, once $P_1(z)$ is determined, $P_0(z)$,$P_2(z)$ and $P_3(z)$ can also be calculated.
Using Cramer's Rule in the matrix above, we have

$$P_1(z)=\frac{\det[A_1(z)]}{\det[A(z)]} \qquad (\alpha.7)$$

$$\det[A_1(z)]=\mu_2(z-1)a_2(z)\left[P_{3,0}(\beta_2 a_0(z)+\lambda_1\mu_1 z)+\lambda_1 a_3(z)P_{0,0}\right] \qquad (\alpha.8)$$

and

$$\begin{aligned} \det[A(z)] &= a_0(z)a_1(z)a_2(z)a_3(z) - \alpha\Pi a_3(z)(\beta_1 a_0(z)+\lambda_1\mu_1 z) \\ &-\alpha(1-\Pi)z a_2(z)(\beta_2 a_0(z)+\lambda_1\mu_1 z) - \lambda_1\mu_1 a_2(z)a_3(z)z \end{aligned} \qquad (\alpha.9)$$

Therefore, if we calculate $P_{0,0}$ and $P_{3,0}$, the problem is solved.
Applying L'Hospital's Rule in (α.7), we have

$$\frac{P_1\det[A'(1)]}{\lambda_1\mu_2(\beta_2+\mu_1)(\beta_1+\mu_1)}=P_{0,0}+P_{3,0} \qquad (\alpha.10)$$

where

$$P_1=\Pr\{V=1\}=\frac{\rho_1}{1+\rho_1}\left(1+\frac{\alpha\Pi}{(\beta_1+\mu_1)}+\frac{\alpha(1-\Pi)}{(\beta_2+\mu_1)}\right)^{-1} \qquad (\alpha.11)$$

It can be shown that there exists a unique z_0, in (0,1) such that det[A(z_0)]=0. Since the numerator of (α.7) must also be zero at z_0, we have

$$P_{3,0}(\beta_2 a_0(z_0)+\lambda_1\mu_1 z_0)+\lambda_1 a_3(z_0)P_{0,0}=0 \qquad (\alpha.12)$$

and combining (α.10) and (α.11)

$$P_{3,0}=\frac{P_1 a_3(z_0)\det[A'(1)]}{\mu_2(\beta_2+\mu_1)(\beta_1+\mu_1)(\lambda_1 a_3(z_0)-\beta_2 a_0(z_0)-\lambda_1\mu_1 z_0)} \qquad (\alpha.13)$$

This determines $P_{3,0}$ and $P_{0,0}$ can be found from (α.10). Thus $P_1(z)$ is specified, and $P_0(z)$, $P_2(z)$, $P_3(z)$ can also be found now.

In order to investigate the system's performance we need to determine the expected number of data packets in the system, which can be definedas:

$$E(Q_D) = P_0^{'}(1) + P_1^{'}(1) + P_2^{'}(1) + P_3^{'}(1) \qquad (\alpha.14)$$

Unfortunately, it is not possible to derive rigorous expressions for the parameters of interest, but it is simple(but tedious) to obtain numerical values by solving the problems for specific values of the parameters, and thus establish the effect of they have on the system performance.

Applying Little's theorem, the prime quantity of interest, the waiting time of data packets in the system, is

$$E(W_D) = \frac{E(Q_D)}{\lambda_2} \qquad (\alpha.15)$$

We also need to determine the maximum throughput of data packets. Clearly, for a stable system, $\rho_2 = \frac{\lambda_2}{\mu_2}$ must be less than the proportion of time the channel is available for transmission of data packets, which is given by $P_0 + P_3 = \Pr\{V = 0\} + \Pr\{V = 3\}$.

The existence condition for the random variable Q_D is :

$$\rho_2 \langle P_0 + P_3 = \frac{1}{1+\rho_1} + \frac{\rho_1}{1+\rho_1} \frac{\frac{\alpha(1-\Pi)}{\beta_2 + \mu_1}}{1 + \frac{\alpha\Pi}{\beta_1 + \mu_1} + \frac{\alpha(1-\Pi)}{\beta_2 + \mu_1}} \qquad (\alpha.16)$$

and the maximum throughput is then

$$S_{max} = (P_0 + P_3)\mu_2 \qquad (\alpha.17)$$

It will be of particular interest to compare the waiting time for the data with the case where data does not use the long silences in conversations, which can be derived from the analysis in [WEIN79] as

$$E(W_{D_0}) = \frac{\frac{(1+\rho_1)^2}{\mu_2} + \frac{\rho_1}{\mu_1}}{(1+\rho_1)(1-\rho_1\rho_2-\rho_2)} \qquad (\alpha.18)$$

The maximum throughput achieved in this case is given by/

$$S_{max} = \mu_2 /(1+ \rho_1) \qquad (\alpha.19)$$

Finally, if we suppose that a separate channel is always available for data packets, instead of using the channel that the voice conversation takes place on, then the data performance up to transmission can be modelled by the basic M/M/1 queue (or M/M/s for s available channels). The waiting time will then be simply

$$E(W_{D_q}) = \frac{1}{\mu_2 - \lambda_2} \qquad (\alpha.20)$$

and the maximum throughput

$$S_{max} = \mu_2 . \qquad (\alpha.21)$$