AD A039330

(12) NW

Eighth Semiannual Technical Report
March 1977
For the Project
INTEGRATED DOD VOICE & DATA NETWORKS
AND GROUND PACKET RADIO TECHNOLOGY
VOLUME 1 - PART 2
INTEGRATED DOD VOICE & DATA NETWORKS

network analysis corporation

100 1493

DDC
RECEIVED
MAY 12 1977
C

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

AD No.
DDC FILE COPY

nac

(12)

Eighth Semiannual Technical Report
March 1977

For the Project
INTEGRATED DOD VOICE AND DATA NETWORKS
AND GROUND PACKET RADIO TECHNOLOGY

VOLUME 1 PART 2

INTEGRATED DOD VOICE AND DATA NETWORKS

DDC
RECEIVED
MAY 12 1977
C

Principal Investigator: Howard Frank
Co-principal Investigator: Israel Gitman

Contractor
NETWORK ANALYSIS CORPORATION
Beechwood, Old Tappan Road
Glen Cove, New York 11542
(516) 671-9580

ARPA Order No. 2286
Contract No. DAHC 15-73-C-0135

Effective Date: 13 October 1972
Expiration Date: 30 June 1977

Sponsored by
Advanced Research Projects Agency
Department of Defense

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

See 1473

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

EIGHTH SEMIANNUAL TECHNICAL REPORT

TABLE OF CONTENTS

VOLUME 1

INTEGRATED DOD VOICE AND DATA NETWORKS

CHAPTER 1 A CLASSIFICATION OF ROUTING STRATEGIES FOR TELECOMMUNICATIONS

CHAPTER 2 ANALYSIS OF INTEGRATED SWITCHING LINKS

CHAPTER 3 DESIGN OF INTEGRATED SWITCHING NETWORKS

CHAPTER 4 NETWORK MODELS FOR PACKETIZED SPEECH; and

CHAPTER 5 A CIRCUIT SWITCH NODE MODEL

VOLUME 2

COST TRENDS FOR LARGE VOLUME PACKET NETWORKS

CHAPTER 6 LARGE SCALE PACKET SWITCHED NETWORK DESIGN TRADEOFFS

VOLUME 3

TOPOLOGICAL GATEWAY PLACEMENT STRATEGIES

CHAPTER 7 TOPOLOGICAL DESIGN OF GATEWAYS FOR PACKET SWITCHED INTER-NETWORK COMMUNICATION

VOLUME 4

GROUND PACKET RADIO TECHNOLOGY

CHAPTER 8 MARKOV CHAIN INITIALIZATION MODELS FOR PACKET RADIO NETWORKS

CHAPTER 9 MARKOV CHAIN INITIALIZATION MODELS WITH FIFO LABEL QUEUE MANAGEMENT AT THE STATION

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

CHAPTER 4

NETWORK MODELS FOR PACKETIZED SPEECH

CHAPTER 4

TABLE OF CONTENTS

		PAGE
4.1	OVERVIEW	4.1
4.1.1	Purpose	4.1
4.1.2	Scope	4.2
4.1.3	Summary of Results	4.3
4.2	TRAFFIC MODELS	4.5
4.2.1	Introduction	4.5
4.2.2	Speech Events	4.5
4.2.3	Speaker Models	4.10
4.2.4	Parameters	4.29
4.2.5	Comparison with Empirical Data	4.35
4.2.6	Call Origination Model	4.41
4.3	PERFORMANCE CRITERIA	4.52
4.3.1	Results of Subjective Studies	4.53
4.3.2	Smoothness Criteria	4.56
4.4	LINK MODEL	4.71
4.4.1	Introduction	4.71
4.4.2	Model Description	4.71
4.4.3	Delay Analysis	4.80
4.4.4	Solution Strategy	4.82
4.4.5	Link Performance Variables	4.91
4.4.6	More General Arrival Scheme	4.94

CHAPTER 4

TABLE OF CONTENTS (Cont'd)

		PAGE
4.5	SINGLE LINK BEHAVIOR	4.97
4.5.1	Properties of the Delay Distribution	4.97
4.5.2	Output Distribution	4.105
4.5.3	Delay Mean and Variance	4.111
4.5.4	Optimal Packet Length	4.120
4.5.5	Finite Buffer Case	4.128
4.5.6	Effect of Speech Models	4.131
4.5.7	Transient Behavior	4.135
4.6	APPROXIMATIONS	4.144
4.6.1	M|D|1 Approximation	4.144
4.6.2	M|M|1 Approximation	4.154
4.6.3	G|G|1 Heavy Traffic Approximation	4.157
4.6.4	Empirical Approximation	4.159
4.6.5	Comparison	4.160
4.6.6	Approximation Conclusions.	4.163
4.7	TANDEM LINK MODEL.	4.164
4.7.1	Introduction	4.164
4.7.2	Model Assumptions and Notations.	4.164
4.7.3	Methodology	4.170
4.7.4	Line Utilization	4.170
4.7.5	Queue Operation	4.173
4.7.6	Queue Delay Distribution	4.176
4.7.7	Approximations	4.178
4.7.8	Extension to Network of Queues	4.180
4.8	CONCLUSION	4.181
REFERENCES		4.184

CHAPTER 4
FIGURES

PAGE

FIGURE 1: THE PACKET STREAM FOR AN ACTUAL SPEECH WAVE.....4.9

FIGURE 2: SIX-STATE MODEL FOR SPEECH PATTERN..... 4.11

FIGURE 3: REPRESENTATIVE CONVERSATIONAL SEQUENCE OF EVENTS..... 4.12

FIGURE 4: FOUR-STATE MARKOV CHAIN MODEL..... 4.15

FIGURE 5: EXAMPLE OF A FOUR-STATE MODEL EVENT SEQUENCE..... 4.16

FIGURE 6: THREE-STATE MODEL WITH TRANSITION PROBABILITIES..... 4.18

FIGURE 7: EXAMPLE OF A THREE-STATE MODEL EVENT SEQUENCE..... 4.19

FIGURE 8: TWO-STATE MODEL..... 4.21

FIGURE 9: EXAMPLE OF A TWO-STATE MODEL EVENT SEQUENCE..... 4.22

FIGURE 10: POSSIBLE TRANSITIONS FOR THE CHAIN Z..... 4.24

FIGURE 11: FOUR-STATE LONG/SHORT SPEECH MODEL..... 4.28

FIGURE 12: BRADY'S EMPIRICAL SPEECH DATA..... 4.36

FIGURE 13: COMPARISON OF EMPIRICAL TALKSPURT LENGTH DATA WITH VARIOUS GEOMETRIC MODELS..... 4.37

FIGURE 14: COMPARISON OF EMPIRICAL TALKSPURT LENGTH DATA WITH IMPROVED GEOMETRIC MODEL..... 4.40

CHAPTER 4

FIGURES (Cont'd)

PAGE

FIGURE 15: CALL ORIGINATION MODEL INTERFACED TO A TWO-STATE SPEECH MODEL.....4.44

FIGURE 16: CALL ORIGINATION MODEL INTERFACED TO A THREE-STATE SPEECH MODEL.....4.45

FIGURE 17: BEHAVIOR OF CALL ORIGINATION MODEL.....4.46

FIGURE 18: UNLIMITED WAITING PROTOCOL.....4.57

FIGURE 19: LIMITED WAITING PROTOCOL.....4.63

FIGURE 20: END BUFFERING.....4.68

FIGURE 21: NETWORK OF QUEUES.....4.72

FIGURE 22: PROBABILISTIC QUEUE ARRIVAL SEQUENCE.....4.77

FIGURE 23: ITERATION SCHEMATIC.....4.90

FIGURE 24: TYPICAL TOTAL DELAY DISTRIBUTIONS.....4.100

FIGURE 25: EXPECTED DELAY AS A FUNCTION OF ρ.....4.102

FIGURE 26: COMPARISON BETWEEN MODEL DELAY DISTRIBUTION AND EXPONENTIAL APPROXIMATION.....4.103

FIGURE 27: CONVEXITY OF TOTAL DELAY VS. PACKET LENGTH.....4.123

FIGURE 28: REPRESENTATIVE OPTIMIZATION CURVE FOR THE PACKET LENGTH.....4.124

FIGURE 29: COMPARISON BETWEEN THE DELAY DISTRIBUTION FOR INFINITE AND FINITE BUFFER FACILITY.....4.129

CHAPTER 4

FIGURE (Cont'd)

PAGE

FIGURE 30: COMPARISON BETWEEN THE DELAY DISTRIBUTION FOR A 2-STATE SPEECH MODEL AND A 3-STATE SPEECH MODEL.....4.134

FIGURE 31: TRANSIENT BEHAVIOR.....4.136

FIGURE 32: TRANSIENT BLOCKING RATE FOR A FOUR BUFFER SYSTEM.....4.138

FIGURE 33a: TRANSIENT DELAY DISTRIBUTION.....4.139

FIGURE 33b: TRANSIENT DELAY DISTRIBUTION.....4.140

FIGURE 34: INSTANTANEOUS UTILIZATION.....4.141

FIGURE 35: TRANSIENT DEPARTURE PROBABILITIES.....4.143

FIGURE 36: EXPONENTIAL APPROXIMATION TO GEOMETRIC.....4.149

FIGURE 37: COMPARISON OF APPROXIMATIONS.....4.161

FIGURE 38: COMPARISON OF APPROXIMATIONS.....4.162

FIGURE 39: FULL DUPLEX TANDEM LINKS.....4.165

FIGURE 40: SIMPLEX TANDEM LINKS.....4.168

FIGURE 41: FLOW GRAPH OF TANDEM LINK TRAFFIC.....4.172

FIGURE 42: PACKET ARRIVALS.....4.175

CHAPTER 4

TABLES

		PAGE
TABLE 1:	DELAY DISTRIBUTION	4.84
TABLE 2:	MEAN AND 95^{th} PERCENTILE AS A FUNCTION OF ρ	4.101
TABLE 3:	EXAMPLE OF OUTPUT PROCESS	4.108
TABLE 4:	LONG PERIOD OUTPUT PROCESS	4.109
TABLE 5:	EFFECT OF FINITE BUFFER SIZE ON OUTPUT PROCESS	4.110
TABLE 6:	CONSTANT μ, ρ, h	4.112
TABLE 7:	PARAMETER VARIATION FOR FUNCTIONAL DEPENDENCIES	4.113
TABLE 8:	EXPERIMENTAL VALUES FOR f_2 AND f_3	4.115
TABLE 9:	EXPERIMENTAL VALUES FOR f_3	4.116
TABLE 10:	FIT OF FUNCTION FORMS OF f_3	4.117
TABLE 11:	COMPARISON OF E(D) WITH VALUE PREDICTED USING FORM 5	4.118
TABLE 12:	HIGH UTILIZATION BEHAVIOR OF EMPIRICAL FUNCTION COMPARED TO M|M|1	4.119
TABLE 13:	PACKET LENGTH OPTIMIZATION	4.125
TABLE 14:	INVARIANCE OF OPTIMAL PACKET LENGTH WHEN L AND A ARE HELD CONSTANT	4.126

CHAPTER 4

TABLES (Cont'd)

PAGE

TABLE 15: "MAP" VARIATIONS FOR L, A, AND Ø EFFECTS ON P_{opt}.....4.127

TABLE 16: EFFECT OF FINITE BUFFER.....4.130

TABLE 17: FRACTION OF PACKETS NOT BLOCKED.....4.132

TABLE 18: COMPARISON OF OPTIMAL PACKET SPEECH LENGTH DETERMINATIONS.....4.155

TABLE 19: OPTIMAL PACKET SPEECH LENGTH AND AVERAGE DELAY FROM FORMULA.....4.156

CHAPTER 4

NETWORK MODELS FOR PACKETIZED SPEECH

4.1 OVERVIEW

4.1.1 Purpose

Packet-switched networks have established themselves as an attractive option in a wide variety of data transmission environments. Circuit-switched networks have been the media for transmitting human speech on a real-time conversational basis. Continuing research efforts have developed a wide variety of methods and devices for digitizing and encoding the analog speech signal. Each digitizing technique has its own cost/performance tradeoffs.

Network transmission adds corruptive effects (noise, delay, echo) to the speech signal. Subjective studies have been conducted to assess the impact of these effects. However, no techniques are available to design packet-switched networks for transmission of speech. The focus of this research is to develop such design techniques so that the cost/performance tradeoffs for the network can be exposed. With these techniques as a tool, system optimizations can be performed with the characteristics of the terminals, the network topology and the protocols all considered jointly. Furthermore, assessments can be made of the relative merits of packet-switching, circuit-switching and hybrid alternatives for handling speech traffic, and ultimately for mixed data and speech traffic.

4.1.2 Scope

In this chapter, we restrict our attention to the modeling, analysis and design of packet-switched networks carrying only real-time speech packets. We thus ignore the needed presence of network control traffic. We further ignore any switch design considerations and assume ideal switch behavior. The related switch issues will be addressed in a later report.

The next section (4.2) contains a discussion of packetized speech traffic models followed by a section (4.3) discussing performance criteria. These traffic models and performance criteria are significant because these are the environmental conditions which provide the primary reasons why packetized speech network design differs from the data case. In general, speech traffic has a more regular and predictable arrival pattern than data and so, intuitively, one would expect the network design to be able to capitalize on this by achieving higher facility utilization with speech than is possible with data. On the other hand, network performance criteria for speech will be more stringent and, in particular, require a regularity or consistency that is not required for data transmission.

Section 4.4 describes a mathematical model for a single link carrying packetized speech data. This is, of course, the simplest possible network, but has sufficient complexity to reveal many of the issues that will be present in more general links and networks. It is thus a good starting point for our investigation and also of some interest in its own right. A fully connected network with only direct routing would consist only of links of this type. Multi-link hopping will generally be less viable with speech traffic because of tight delay constraints.

Section 4.5 contains the results of a wide-ranging series of experiments on the single link model including optimization of packet length. Section 4.6 presents some closed form approximations to the single link case, adapted from standard references on queueing, and compares their behavior with the results of our model. The seventh section (4.7) develops a model of a more general tandem link situation. Section 4.8 is a summary of the conclusions and direction for further research. The numerous references used in the study are listed next.

Future research will be directed toward developing models for more general network situations and evolving methodologies for generating network designs and accurately evaluating their expected performance.

4.1.3 Summary of Results

The major accomplishment reported on in this chapter is the formulation and study of a mathematical model of a single link communications channel carrying packetized voice. The major results are:

1. A unified treatment of speaker behavior models.

2. An analysis of the relationship between a variety of protocols and performance criteria.

3. A computational scheme is developed for obtaining the steady state delay distribution.

4. The delay distribution is shown to be approximately exponential; the single parameter characterizing the distribution is obtained, as a function of system parameters.

5. The delay dependencies on packet size are obtained. A closed form expression for "optimal" packet size is derived for an approximation to the delay model. It is shown that a network serving low bandwidth terminals (e.g., vocoders) requires very small packets, while a network for high bandwidth terminals (e.g., PCM) operates best with somewhat longer packets.

6. A close approximation to the single link steady state output process is presented.

7. The effects of finite buffers is investigated. It is shown that a small number of buffers suffices to sustain excellent performance even at high utilization.

8. The transient behavior of the system is analyzed. *Transient performance degradation* is shown to be of limited duration.

9. The model is extended via a formulation of a tandem link model.

4.2 TRAFFIC MODELS

4.2.1 Introduction

The statistical analysis of speech patterns has attracted attention for the past half century. As a product of such study, several models for telephonic speech patterns have emerged. In general, to obtain a better fit to empirically measured data, the models must grow correspondingly in sophistication and complexity.

The major realization of the investigation has been that a Markov process, with appropriate number of states, describes the speech mechanics well. In this section we describe in an organized fashion various models which can be used to study speech behavior in a statistical sense. Some of the models have been studied in the literature; others have been developed to fill in the tradeoff gaps between model complexity and correspondence with the empirical data. For each model we give a short discussion of its relative merits and weaknesses. These issues are also discussed in [JAFFE, 1964], [BRADY, 1967], [BRADY, 1969], and [BRADY, 1970]. A good starting reference is [NORWINE, 1938]. [BRADY, 1969] represents the most comprehensive model and treatment.

The accuracy of a model is its ability to predict the length of the 10 speech events described in the next subsection. The more sophisticated models (e.g., 4.2.3.1 below) can accurately predict the distribution of all of these events. The less sophisticated models (e.g., 4.2.3.3) can predict the distribution of only a few events, particularly the talkspurt length and pause length. However, it is these events which are of most interest for our traffic models for network statistics. These simpler models yield more tractable analytical formulations.

4.2.2 Speech Events

The following ten events are relevant to speech patterns:

1. Talkspurt.

2. Pause.

3. Double talk.

4. Mutual silence.

5. Alternation silence.

6. Pause in isolation.

7. Solitary talkspurt.

8. Interruption.

9. Speech after interruption.

10. Speech before interruption.

Consider two speakers, A and B. The events are defined as follows:

(1 and 2) Talkspurt, Pause ([Brady, 1969]) - The technique of obtaining on-off speech patterns is summarized as follows. A flip-flop is set any time speech (full-wave rectified and unfiltered) from speaker A crosses a threshold. This flip-flop is examined and cleared every 5 milliseconds, with the output being a 1 if the threshold was crossed, 0 otherwise. The resuling string of 1's (spurts) and 0's (gaps) is examined for short spurts; all spurts $\leq$ 15 msec. are erased. After this is done,

all gaps $\leq$ 200 msec are filled in to account for momentary interruptions, such as those due to stop consonants. The resulting on-off pattern consists, by definition used here, of talkspurts and pauses. An identical procedure is used for speaker B.
Note: This definition of talkspurt is not universal, other investigators, [JAFFE, 1964], [NORWINE, 1938], [EMLING, 1963], use alternate definitions.

(3) Double talk - A time when speech is present from both A and B.

(4) Mutual silence - A time when silence is present from both A and B.

(5) Alternation silence - The period of mutual silence between the end of one speaker's talkspurt and the beginning of the other's. Event 5 is a subset of 4. If a speaker alternation results from an interruption so that there is no mutual silence period, then an alternation silence has not occurred. (There are no negative alternation silences.)

(6) Pause in isolation - A pause in which the other speaker is silent throughout the pause. Event 6 is a subset of both 2 and 4.

(7) Solitary talkspurt - A talkspurt which occurs entirely within the other speaker's silence. Event 7 is a subset of 1.

(8) Interruption - If A interrupts B, the time at which A's talkspurt begins determines the start of an interruption. The interruption terminates at the end of A's talkspurt, unless B stops and then interrupts A, in which case A's interruption terminates upon B's counter interruption.

(9) Speech after interruption - If A interrupts B, the remainder of B's talkspurt is entered here, unless A pauses and then again interrupts the same B talkspurt. The first "speech after interruption" would terminate upon A's reinterruption, and a second speech after interruption would begin.

(10) Speech before interruption - If A interrupts B, B's speech interval up to the interruption is entered here. If A then pauses at time t_1 and reinterrupts at time t_2, (assuming B continues talking), a new B speech before interruption (t_2-t_1) is entered. If A continues talking and B pauses and then counter interrupts, the length of B's pause is entered as A's speech before interruption.

We diverge slightly from the approach employed in the literature, in that we consider a packetized talkspurt; namely, instead of considering a continuous Markov process we set up a discrete Markov chain. We furthermore assume the discrete chain is homogeneous (time invariant). We presume the existence of a device or software algorithm which can test a speech packet for an energy threshold. Our definition of talkspurt is as follows:

Definition: A contiguous sequence of non-empty packets from a single talker constitutes a talkspurt. A packet is considered to be non-empty if it exceeds the energy threshold. Thus, any pause duration less than 1 packet's time length will most likely be swallowed up in the discrete packetization and the talkspurt will be considered not to have been interrupted by a pause. Pauses of up to 2 packets in length could be swallowed up, depending on the time phasing of the pause. (See Figure 1.)

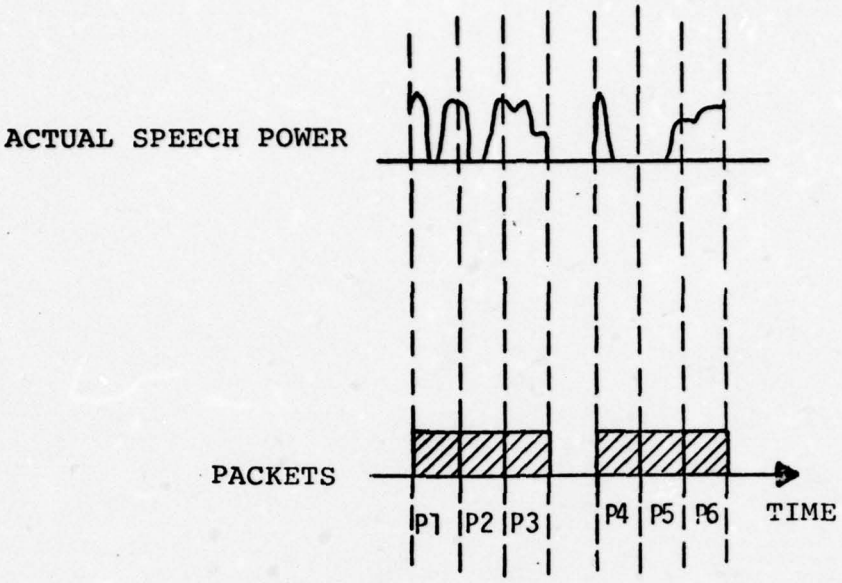


FIGURE 1: THE PACKET STREAM FOR AN ACTUAL SPEECH WAVE

4.2.3 Speaker Models

We assume a time-slotted environment in which the speech takes place. That is, a clock divides the time axis into segments, during each of which a speaker generates an empty or non-empty packet, depending on whether the threshold energy was exceeded during that time period. These time segments will be called frames. Note that this specifically excludes voice-actuated synchronization. Thus an isolated frame length utterance - a rare event - embedded in a silent period will most likely generate two packets. In the transmitted speech signal we assume that empty packets have been discarded and any references in the sequel to a packet mean a non-empty packet. We make the additional simplifying assumption that only one member of the speaker-listener pair changes his speech activity state during a frame.

4.2.3.1 Six-State Markov Chain Model - Brady Model

The Markov chain state transition diagram is depicted in Figure 2. A possible sequence of events is depicted in Figure 3. Observe that a talkspurt for A is made up of any concatenation of states 1, 2 and 3.

Let $P_{AB} = (p_{ij})$ be the state transition matrix, where p_{ij} is the probability of being in state i for a frame that immediately follows a frame in which the state was j.

Let $P^{(n)} = (P_1^{(n)}, P_2^{(n)}, \ldots, P_6^{(n)})$ be the vector of unconditional probabilities of being in state 1, 2, 3, ..., 6 at frame n. $P^{(0)}$ therefore will represent an initial condition.

Then $P^{(n)} = P^{(0)} P_{AB}^n$ is the vector of unconditional state probabilities at frame n.

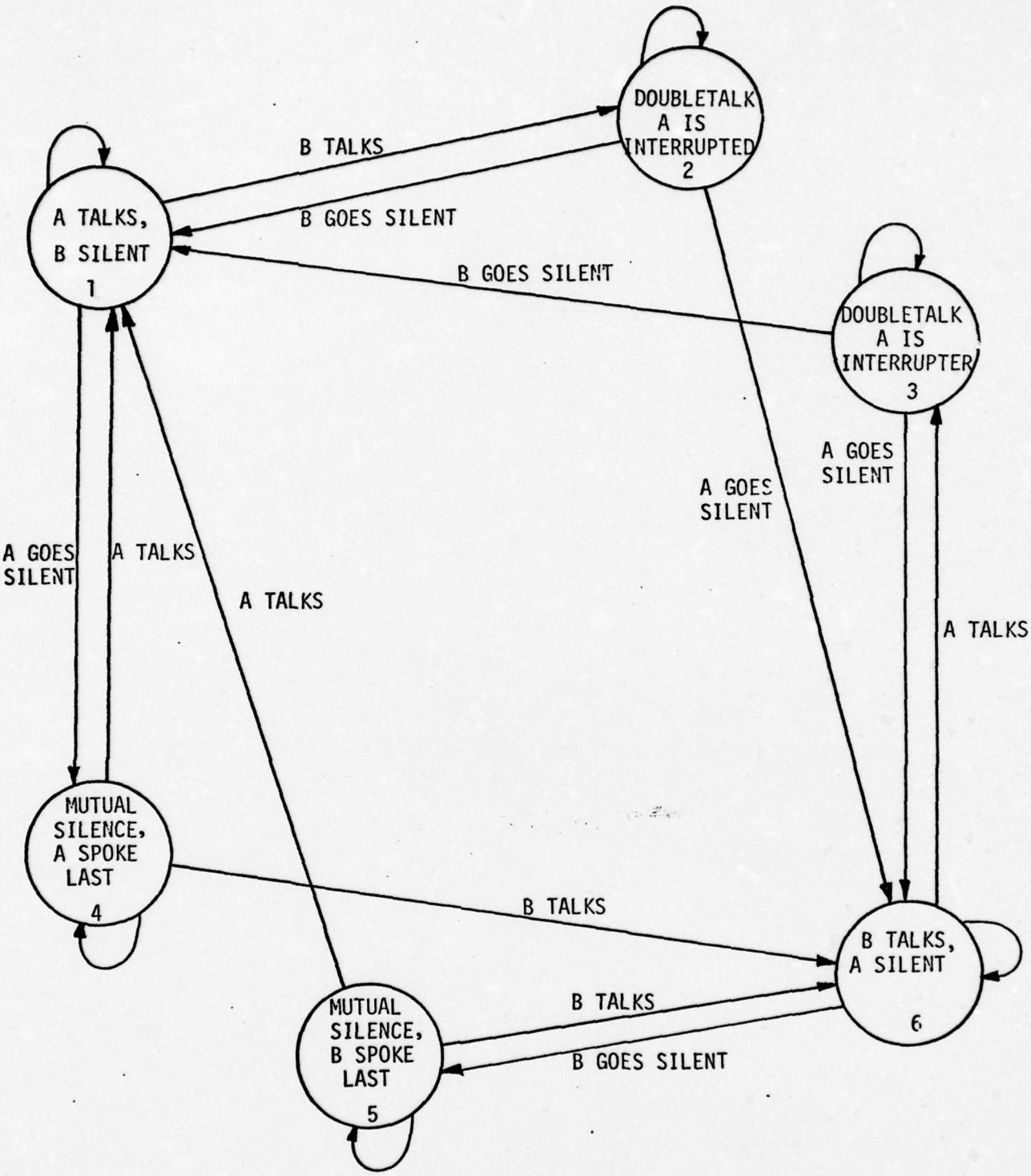
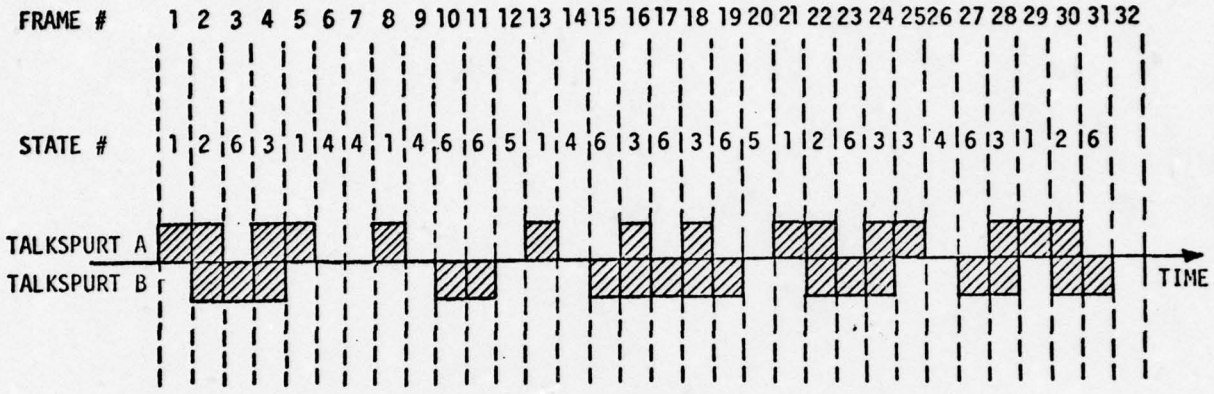


FIGURE 2: SIX-STATE MODEL FOR SPEECH PATTERN



Shaded areas indicate a non-empty packet issued by the speaker

FIGURE 3: REPRESENTATIVE CONVERSATIONAL SEQUENCE OF EVENTS

From the well-known properties of homogeneous Markov chains, we have the fact that the steady state ($n\to\infty$) distribution of the number of consecutive frames spent in a particular state is geometric. The probability that speaker A supplies a packet at frame n is $P_1^{(n)}+P_2^{(n)}+P_3^{(n)}$, the sum of the first three components of $P^{(n)}$. Note that neither the talkspurt length nor the silence length is distributed geometrically despite the fact that they are a combination of events each of whose length is geometrically distributed.

We now transfer our attention from the behavior of a single talker and examine the implications of this model for a conglomerate of talkers.

Define an associated chain, $X_j^{(n)}$, for speaker j such that $X_j^{(n)}$ is 1, if speaker j at the n^{th} frame is in states 1, 2 or 3; 0, otherwise.

We are interested in

$$Z^{(n)} = \sum_{j=1}^{m} X_j^{(n)} \qquad (1)$$

where m is the number of speakers actively engaged in conversation who are accessing the same switch. $Z^{(n)}$ is thus the number of packet arrivals to the switch in the n^{th} frame.

Let $Z_{\ell k} = \text{Prob}(Z^{(i)}=k \mid Z^{(i-1)}=\ell)$. $Z_{\ell k}$ is independent of i because of the homogeneous assumption.

Such a transition event for the Z chain occurs (see Figure 10) if

k-s	of the X_j's go from 0 to 1
s	of the X_j's go from 1 to 1
ℓ-s	of the X_j's go from 1 to 0
m-ℓ-k+s	of the X_j's go from 0 to 0

where the valid range of s is

$$\max(0, k+\ell-m) \leq s \leq \min(k, \ell).$$

This valid range on s can be determined from the simple requirement that the number of transitions of each type be positive.

Consider an example. A particular X_j could go from 1 to 0 if the corresponding chain goes from states

1 to 4
or
2 to 6
or
3 to 6.

Thus, we require

y of the X_j's to go from 1 to 4

w of the X_j's to go from 2 to 6

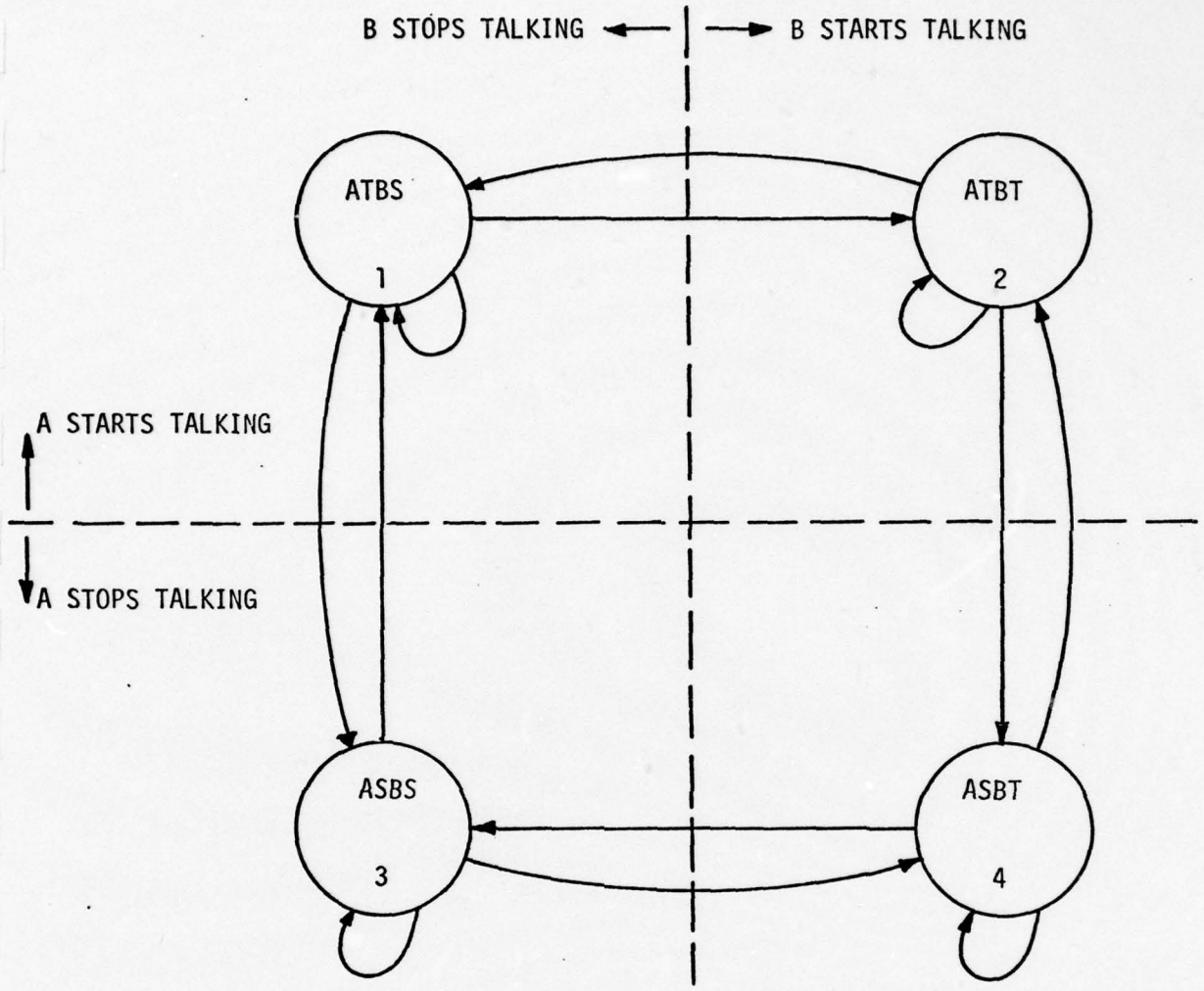
ℓ-s-y-w of the X_j's to go from 3 to 6.

The determination of the transition probabilities and unconditional probabilities for the Z chain is clearly complex.

While this six-state model is of interest because of its excellent predictive ability for the ten events, there exists simpler models which, although they cannot accurately describe certain events in the dynamics of a conversation, still yield accurate talkspurt and pause lengths.

4.2.3.2 Four-State Markov Chain Model

By collapsing states 2 and 3 into a single state and similarly collapsing 4 and 5 into a single state, we obtain the four-state chain depicted in Figure 4. Figure 5 shows one of the possible events.



NOTATION: XTYS ≡ X TALKS, Y SILENT

FIGURE 4: FOUR-STATE MARKOV CHAIN MODEL

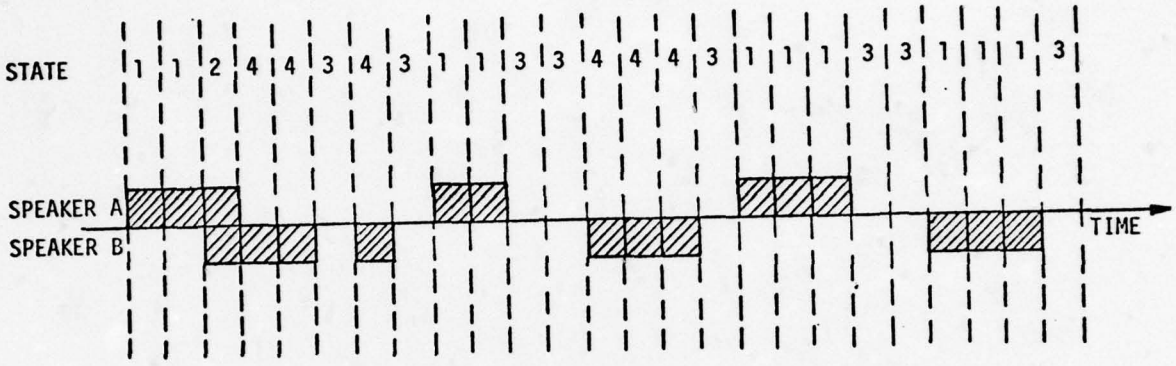


FIGURE 5: EXAMPLE OF FOUR-STATE MODEL EVENT SEQUENCE

Observe that now a talkspurt from A consists of the concatenation of states 1 and 2 only. This model predicts talkspurt length and pause length well, but does not provide an accurate fit to empirical data for doubletalk.

As in the previous section, we can obtain the probability that A supplies a packet in frame n as $P_1^{(n)}+P_2^{(n)}$. We can also formally obtain $Z^{(n)}$.

Using this model, neither the talkspurt length nor the silent period are distributed strictly geometrically. The model accuracy is diminished from the six-state case by being no longer able to distinguish how a mutual silence or doubletalk state was arrived at. It has been claimed that an eight-state model does not predict empirical behavior any better than the six-state model; i.e., it makes little difference how a speaker has arrived in a state where he alone is speaking. However, the transition from mutual silence or doubletalk is significantly influenced by who "had the floor" last. This influence is not accounted for in the four-state model.

4.2.3.3 Three-State Markov Chain Model

By ignoring the possibility of doubletalk altogether, namely eliminating state 2 in the previous model, we obtain a further simplified model which can predict talkspurt length and silence length distribution, but cannot even represent doubletalk. (See Figures 6 and 7.) Note that the talkspurt is characterized by a single state while the silence period of a particular talker is characterized by two states. With this model, it turns out that the talkspurt length is geometrically distributed, but the silent period length is not. This happens to be also true of the empirical data. The distribution of the silent period length can be computed from the state transition probabilities given in Figure 6 as follows:

P_{AB} =

	1	2	3
1	1-p	p	o
2	x	1-x-y	y
3	0	r	1-r

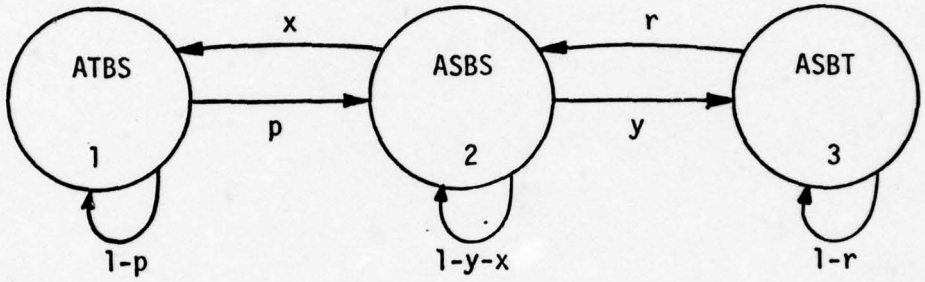


FIGURE 6: THREE-STATE MODEL WITH TRANSITION PROBABILITIES

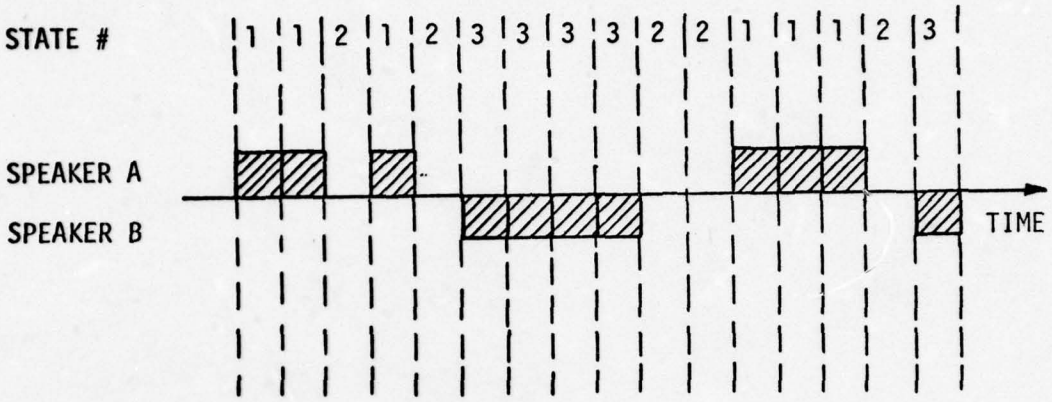


FIGURE 7: EXAMPLE OF A THREE-STATE MODEL EVENT SEQUENCE

Silent Period Length (In Frames or Packets)	Possible State Sequences	Probability
1	121	px
2	1221	$p(1-x-y)x$
3	12221 12321	$p(1-x-y)^2x+pyrx$
4	122221 122321 123221 123321	
⋮		
etc.		

Now, however, the probability that user A generates a packet in frame n is simply $P_1^{(n)}$. We can define $Z^{(n)}$ for a group of speakers as in the previous subsections and obtain its unconditional probability distribution.

4.2.3.4 Two-State Markov Chain Model

As a final simplification we can assume an elementary two-state chain. This can still model the talkspurt length fairly reasonably, but does not model the silence length well. This chain is obtained by eliminating the possibility for mutual silence, namely state 2, in the previous chain. Thus speaker A and B alternate their turns speaking with no pauses or response time silences. (See Figures 8 and 9.) Such a model was analyzed by [JAFFE, 1964]. The talkspurt length and the silence length are both geometrically distributed (only the former is realistic). In view of its simplicity this model has been investigated further and applied to our link model analysis. The other models could also be used with a slight increase in computational effort (see Section 4.5.6).

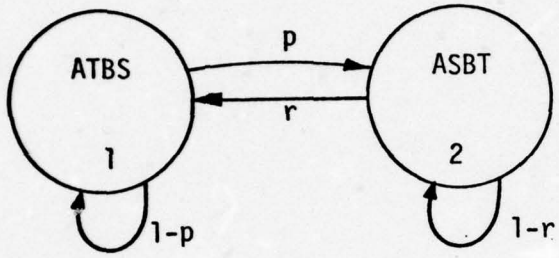


FIGURE 8: TWO-STATE MODEL

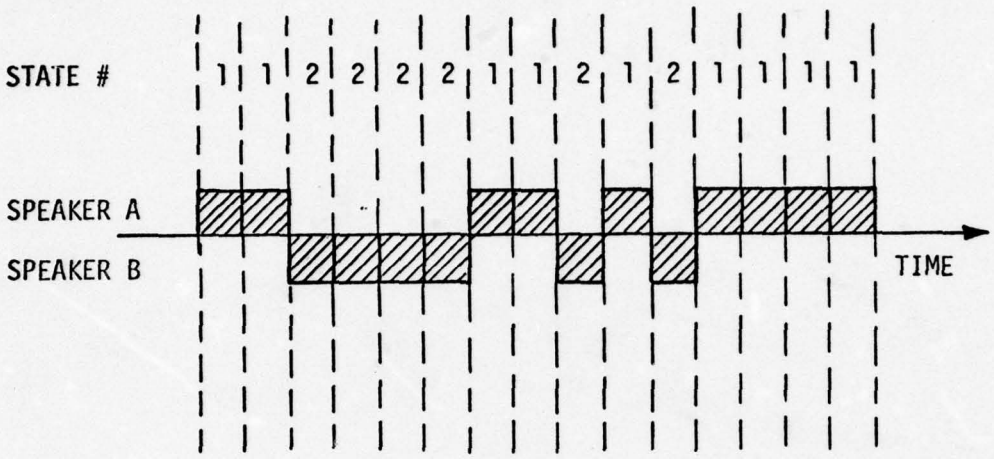


FIGURE 9: EXAMPLE OF A TWO-STATE MODEL EVENT SEQUENCE

The probability that speaker A inputs a packet at frame n is $P_1^{(n)}=P_1^{(0)}p_{11}^{(n)}+P_2^{(0)}p_{21}^{(n)}$, where $p_{ab}^{(n)}$ is the ab^{th} entry of P_{AB}^n, the n^{th} power of the state transition matrix.

To get the conglomerate picture, we proceed as before by letting $Z^{(i)}$ be the number of packets generated by a group of speakers in frame i. Clearly,

$$Z^{(i)} = \sum_j X_j^{(i)} \tag{2}$$

where $X_j^{(i)} = 1$ if terminal j supplies a packet in slot i; 0 otherwise.

The transition probabilities for $Z^{(i)}$ are as follows (see Figure 10):

$$Z_{\ell k} = \text{Prob}\ (Z^{(i)}=k \mid Z^{(i-1)}=\ell)$$

$$= \sum_{s=\max(0,k+\ell-m)}^{\min(k,\ell)} \binom{\ell}{s}\binom{m-\ell}{k-s}\, r^{k-s}(1-r)^{m-\ell-k+s}\, p^{\ell-s}(1-p)^{s} \tag{3}$$

since we can have

k-s	transitions 0→1
s	transitions 1→1
ℓ-s	transitions 1→0
m-ℓ-k+s	transitions 0→0.

We have assumed that the $X_j^{(i)}$'s are independent of j. This would not be true if the collection of m speakers contains pairs who are mutually conversing. Thus, we assume that speakers connected to the network through the same packet switch will not enter into the line transmission queues and need not be considered here. We note, however, that the number of packets generated at slots i and i-1 are not independent, but positively correlated. We have also assumed that each speaker in the collection has the same speech characteristics, i.e., the same state transition matrix.

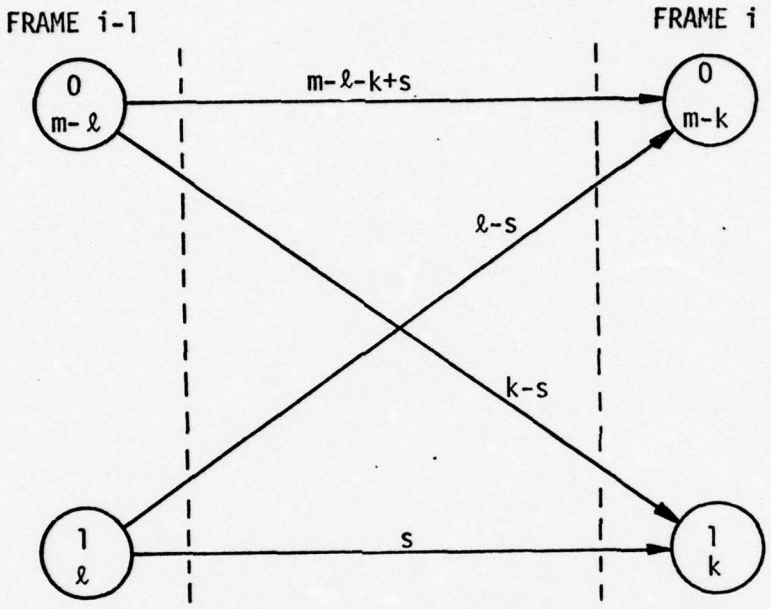


FIGURE 10: POSSIBLE TRANSITIONS FOR THE CHAIN Z

Let

$$\mathbb{P} = (Z_{\ell k}). \tag{4}$$

$\mathbb{P}^n$ is the n^{th} step transition probability matrix, where the elements of $\mathbb{P}^n$ are denoted by $Z_{\ell k}^{(n)}$.

Let

$$\mathbb{P}^{(n)} = (\mathbb{P}_0^{(n)}, \mathbb{P}_1^{(n)}, \ldots, \mathbb{P}_m^{(n)}) \tag{5}$$

be the vector whose elements are unconditional probabilities of finding the Z-chain in state 0, 1, ..., m, after n steps.

Then

$$\mathbb{P}^{(n)} = \mathbb{P}^{(0)}\ \mathbb{P}^n \tag{6}$$

Let $\tilde{\Pi} = \lim_{n\to\infty} \mathbb{P}^{(n)}$ be the vector of steady state probabilities of finding the Z-chain in state 0, 1, ..., m.

The above exact formulation does not lead to a tractable solution. We thus take an alternate approach. Instead of finding the unconditional distribution of $Z^{(i)} = \Sigma X_j^{(i)}$, we find the unconditional distribution of the $X_j^{(i)}$'s and then determine $Z^{(i)}$ from these.

The transition matrix for a typical $X_j^{(i)}$ is

$$\begin{pmatrix} 1-r & r \\ p & 1-p \end{pmatrix}$$

from which

$$P_{AB}^n = \begin{pmatrix} p_{11}^{(n)} & p_{12}^{(n)} \\ p_{21}^{(n)} & p_{22}^{(n)} \end{pmatrix} = \begin{pmatrix} \frac{p}{r+p} + r\,\frac{(1-r-p)^n}{r+p}, & \frac{r}{r+p} - r\,\frac{(1-r-p)^n}{r+p} \\ \frac{p}{r+p} - p\,\frac{(1-r-p)^n}{r+p}, & \frac{r}{r+p} + p\,\frac{(1-r-p)^n}{r+p} \end{pmatrix}. \tag{7}$$

Thus

$$\left.\begin{aligned} P_1^{(n)} &= P_1^{(0)}\ p_{11}^{(n)} + P_2^{(0)}\ p_{21}^{(n)} \\ P_2^{(n)} &= P_1^{(0)}\ p_{12}^{(n)} + P_2^{(0)}\ p_{22}^{(n)}. \end{aligned}\right\} \tag{8}$$

Now

$$\text{Prob(there are k packets inputted at the } n^{th} \text{ frame)} =$$

$$\text{Prob(k of the } X_j^{(n)} \text{ are in state 1)} = \binom{m}{k}[P_1^{(n)}]^k[P_2^{(n)}]^{m-k}$$

$$= \text{Prob}(Z^{(n)}=k). \tag{9}$$

The steady state X-chain transition probabilities are:

$$\lim_{n\to\infty} P_{AB}^n = \begin{pmatrix} \frac{p}{r+p} & \frac{r}{r+p} \\ \\ \frac{p}{r+p} & \frac{r}{r+p} \end{pmatrix}. \tag{10}$$

and the steady state probability of the X_j-chain being in state 1 is

$$P_1 = P_1^{(0)}\ \frac{p}{r+p} + P_2^{(0)}\ \frac{p}{r+p} = \frac{p}{r+p}, \text{ since } P_1^{(0)} + P_2^{(0)} = 1.$$

Similarly,

$$P_2 = \frac{r}{r+p} \tag{11}$$

from which we obtain

$$\lim_{n\to\infty} P_k^{(n)} = \lim_{n\to\infty} \text{Prob(k packets from m speakers at frame n)}$$

$$= \binom{m}{k}\ (\frac{p}{r+p})^k(\frac{r}{r+p})^{m-k}. \tag{12}$$

This last expression gives the k^{th} element in the vector $\tilde{\Pi}$, which is the steady state probabilities for the states of the Z-chain.

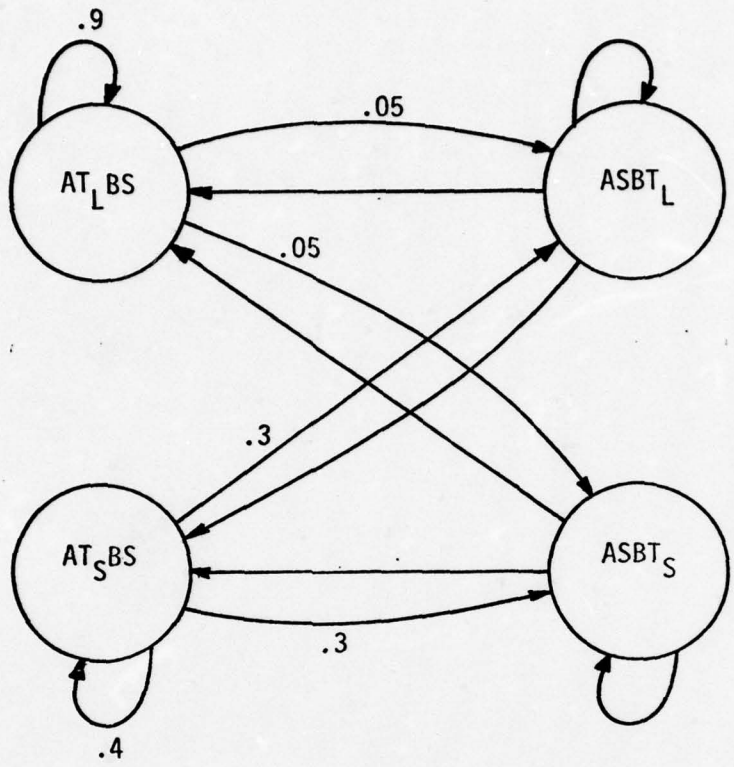
Thus, we observe that the steady state arrival process is a Binomial process. This is clearly due to the fact that each speaker's behavior is an independent Bernoulli trial, with probability $\frac{p}{p+r}$ of supplying a non-empty packet. This result can be used to easily compute the steady state traffic statistics; also it will be employed in subsequent queueing models.

The steady state traffic can be used in studying long term network behavior but the n^{th}-step unconditional probabilities would be needed for any investigation of network transient behavior.

4.2.3.5 Other Models

For the purpose of studying network behavior, the most significant aspect of the traffic models for a speaker is the resulting distribution of the length of the stream consecutive packets (i.e., talkspurt length). In comparing the implications of our previous models with empirical data (see Section 4.2.5), it was noticed that although the empirical distribution talkspurt length is close to the geometric that our models predict; the most significant discrepancy is an under prediction of the frequency of short talkspurts.

An intuitive conjecture is that a speaker operates in at least two modes, one where he is the "controlling" speaker, and the other where he is merely issuing short utterances to "reinforce" the speech he is receiving. Thus, we can extend any one of the previous models to include two distinct states for "A talking, B silent." In Figure 11, we augment the two-state chain to a four-state chain by introducing two such new states - one for each speaker. One state has a high probability for speech to be continued into the next frame (.9 in Figure 11) and the other state has a much lower continued speech probability (.4 in the figure). Thus, each active speaker can be in a "long burst" or "short burst" mode.



NOTATION: SUBSCRIPT L IS "LONG BURST"
SUBSCRIPT S IS "SHORT BURST"

FIGURE 11: FOUR-STATE LONG/SHORT SPEECH MODEL

4.2.3.6 Application of the Speech Model to the Network Model

The network queuing model to be developed requires knowing the probability that the population of off-hook terminals accessing a specific switch supplies a given number of packets per frame. Because of the complexity involved in obtaining closed form solutions for the distribution of $Z^{(n)}$ for other than the two-state or three-state Markov chain cases, we have based our subsequent development on the random variables $X_j^{(n)}$ themselves. Furthermore, the individual probability of speaker j supplying a packet in frame n, $\text{Prob}(X_j^{(n)} = 1)$, is computed numerically by matrix multiplication, rather than in closed form. The details of this approach will be treated later when applied to the queuing model.

In our application we assume a conversation of infinite length (see Section 4.2.6); with this assumption we see that the two-state model is the most conservative in the sense that it predicts the existence of more non-empty packets than, say, the three-state model. Since the pause length is in general longer than that predicted by the two-state model, we can in reality capitalize on such additional inactivity periods of the speakers to support additional speakers. Thus, for a given network or line capacity, an analysis based on the two-state model and the infinite length conversation overestimates the actual load and is a conservative evaluation of performance.

4.2.4 Parameters

A fundamental assumption we can make for the speech models, is that both members of a speaker-listener pair are symmetric. That is to say, there is the same tendency for A to talk (go silent, interrupt, resume speech) as there is for B. This means that the transition matrices discussed above have special structures.

We now re-examine the two-state and three-state models, with this additional consideration.

For the two-state chain, the transition matrix is

$$P_{AB} = \begin{pmatrix} 1-r & r \\ p & 1-p \end{pmatrix}. \tag{13}$$

But if we assume the same probability that A continues talking once he is talking, as that B continues talking once he is talking, we have r = p and

$$P_{AB} = \begin{pmatrix} 1-r & r \\ r & 1-r \end{pmatrix}. \tag{14}$$

In this case a single parameter, r, characterizes the distribution. Thus

$$\frac{r}{r+p} = \frac{p}{r+p} = \frac{1}{2}, \tag{15}$$

so that

$$\lim_{n\to\infty} \text{Prob}(Z^{(n)} = k) = \binom{m}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{m-k} = \binom{m}{k} \left(\frac{1}{2}\right)^m. \tag{16}$$

This is quite intuitive, since it says that half of the time a particular member of the population is talking, and half he is listening. Thus, on the average, half of the population accessing a switch is active per frame, while the other half is idle. This is so, independently of the value of r.

What is the meaning of r, then? It represents the "aggressiveness" of the idle party to "grasp the floor". It affects the distribution of the length of the talkspurt;

$$\text{Prob (talkspurt length = w)} = (1-r)^{w-1} r, \quad w = 1,2,\ldots \tag{17}$$

$$E(\text{talkspurt length}) = \frac{1}{r}. \tag{18}$$

However, it does not influence the steady state distribution of the number of the packets supplied by a population of users at a given frame. We can say that the statistical fluctuations of one speaker cancels out with the statistical fluctuations of another speaker, and the conglomerate number of packets supplied is stable.

This is, however, strictly true only in the steady state (see $P_1^{(n)}$ and $P_2^{(n)}$ in Section 4.2.3.4).

To further illustrate the role of the parameter r, we now derive the frame-to-frame correlation for the two-state model. Clearly

$$\text{Prob}(X_j^{(n-1)}X_j^{(n)} = 1 \mid X_j^{(n-1)} = 1) = \text{Prob}(X_j^{(n-1)} = 1, X_j^{(n)} = 1 \mid X_j^{(n-1)} = 1)$$

$$= 1-r. \tag{19}$$

Thus

$$E(X_j^{(n-1)}X_j^{(n)}) = (1-r)\ P_1^{(n-1)}. \tag{20}$$

Also

$$E(X_j^{(n)}) = P_1^{(n)} E\left[(X_j^{(n)})^2\right] = P_1^{(n)} \tag{21}$$

$$E(X^{(n-1)}) = P_1^{(n-1)}\ \ E\left[(X_j^{(n-1)})^2\right] = P_1^{(n-1)}. \tag{22}$$

Thus

$$\rho_{n,n-1} = \frac{(1-r)\ P_1^{(n-1)} - P_1^{(n-1)}\ P_1^{(n)}}{\sqrt{(P_1^{(n)} - [P_1^{(n)}]^2)(P_1^{(n-1)} - [P_1^{(n-1)}]^2)}}. \tag{23}$$

If we let $n \to \infty$ and define $\rho = \lim_{n\to\infty} \rho_{n,n-1}$,

$$\rho = \frac{(1-r)\ P_1 - P_1^2}{P_1 - P_1^2} = \frac{(1-r) - P_1}{1 - P_1}. \tag{24}$$

Expressing P_1 in terms of the transitional probabilities, namely $P_1 = p/(r+p)$, we get

$$\rho = 1-r-p. \tag{25}$$

Thus if r+p = 1, the frame-to-frame correlation is zero. With the speaker symmetry assumption of p=r, we conclude

$$\rho = 0 \text{ if and only if } r=p=\frac{1}{2}. \tag{26}$$

We now see the effect of the parameter r on frame-to-frame correlation; if $r=\frac{1}{2}$ we are in a completely random situation analogous to coin flipping with no frame-to-frame correlation. As $r \to 0$, the talkspurt length approaches infinity, and $\rho \to 1$ as should be the case. For a typical transition matrix for speech, r=.1 yielding $\rho=.8$.

One should not conclude that since $\lim_{n\to\infty} P_1^{(n)} = P_1$ and $\lim_{n\to\infty} P_2^{(n)} = P_2$, the frame-to-frame correlation dies off in the steady state. On the contrary, each $P_i^{(n)}$ is computed directly from $P_1^{(n-1)}$ and $P_2^{(n-1)}$, thus it is clearly dependent on the previous frame. The above analysis has shown that the parameter of the two-state model with speaker symmetry influences the frame-to-frame correlation and the spurt length. Actually, the *steady state* queuing delay will be shown to be independent of this frame-to-frame correlation, since corresponding to a (high) correlation of continued speech for a specific talkspurt, there is a (high) correlation for a continued silence once the speaker goes silent; the steady state accounts for the weighted probability of either event. Also as shown above, the distribution of the number of packets supplied per frame is independent of r. (See Section 4.5.6)

The symmetry assumption for the three-state model implies that p=r and x=y, so that the transition matrix is

$$\begin{pmatrix} 1-p & p & 0 \\ x & 1-2x & x \\ 0 & p & 1-p \end{pmatrix}.$$

With this simplification, it is straightforward to compute the steady state distribution of $Z^{(n)}$. The steady state unconditional probabilities for the three-state chain are:

$$\left.\begin{aligned} P_1 &= \frac{x}{2x+p} \\ P_2 &= \frac{p}{2x+p} \\ P_3 &= \frac{x}{2x+p} \end{aligned}\right\} \qquad (27)$$

Note that $P_1 = P_3$, as expected.

Thus

$$\lim_{n\to\infty} \text{Prob}(Z^{(n)}=k) = \binom{m}{k} \left(\frac{x}{2x+p}\right)^k \left(\frac{x+p}{2x+p}\right)^{m-k} \qquad (28)$$

It then follows that

$$\lim_{n\to\infty} E(Z^{(n)}) = m\frac{x}{2x+p} \; . \qquad (29)$$

This expression is needed to compute the steady state link utilization when the input traffic is taken as the three-state *speech model*.

Comparing the expected number of packets supplied by m users under the three-state model to that of a two-state model we see that

$$m\frac{1}{2} > m\frac{x}{2x+p} \qquad (30)$$

for all choices of the parameters x and p, substantiating the conser-servative nature of a two-state model traffic assumption.

The symmetry of the four-state chain implies that

$$P_{24} = P_{21}$$
$$P_{42} = P_{12}$$
$$P_{13} = P_{43}$$
$$P_{31} = P_{34}.$$

Thus the transition matrix is

$$\begin{pmatrix} 1-A-B & A & B & 0 \\ D & 1-2D & 0 & D \\ C & 0 & 1-2C & C \\ 0 & A & B & 1-A-B \end{pmatrix}.$$

The steady state probabilities are easily obtained and are

$$\left.\begin{aligned} P_1 &= \frac{DC}{2DC+AC+BD} \\ P_2 &= \frac{AC}{2DC+AC+BD} \\ P_3 &= \frac{BD}{2DC+AC+BD} \\ P_4 &= \frac{DC}{2DC+AC+BD} = P_1. \end{aligned}\right\} \qquad (31)$$

Observe that $P_1=P_4$ as expected from the symmetry condition. On the other hand $P_2 \neq P_3$, since we may anticipate a stronger (or weaker) tendency to break away from a double talk situation than from a mutual silence situation.

As before

$$\lim_{n\to\infty} \text{Prob}(Z^{(n)}=k) = \binom{m}{k}[P_1+P_2]^k[P_3+P_4]^{m-k} \qquad (32)$$

$$= \binom{m}{k}\left[\frac{C(D+A)}{2DC+AC+BD}\right]^k\left[\frac{D(B+C)}{2DC+AC+BD}\right]^{m-k}.$$

The interpretation of each term in this expression can be related to the Markov chain transitional structure but we omit further explanation here. Again

$$\lim_{n\to\infty} E(Z^{(n)}) = m\,\frac{C(D+A)}{2DC+AC+BD}. \qquad (33)$$

Similar expressions can be obtained for the six-state model and for the long/short models.

4.2.5 Comparison with Empirical Data

The good fit of the exponential random variable to the empirical time-length distribution of the talkspurts is well documented in the literature [BRADY, 1969]. Let Z_c be a continuous talkspurt length random variable and let Z_p be the talkspurt length in packets. If h is the fixed time length of a packet (i.e., frame width), then $Z_p=j$ if and only if $(j-1)h < Z_c \leq jh$. It is easy to show that if Z_c is distributed exponentially with parameter α, then Z_p is distributed geometrically with parameter $p=1-e^{-\alpha h}$.

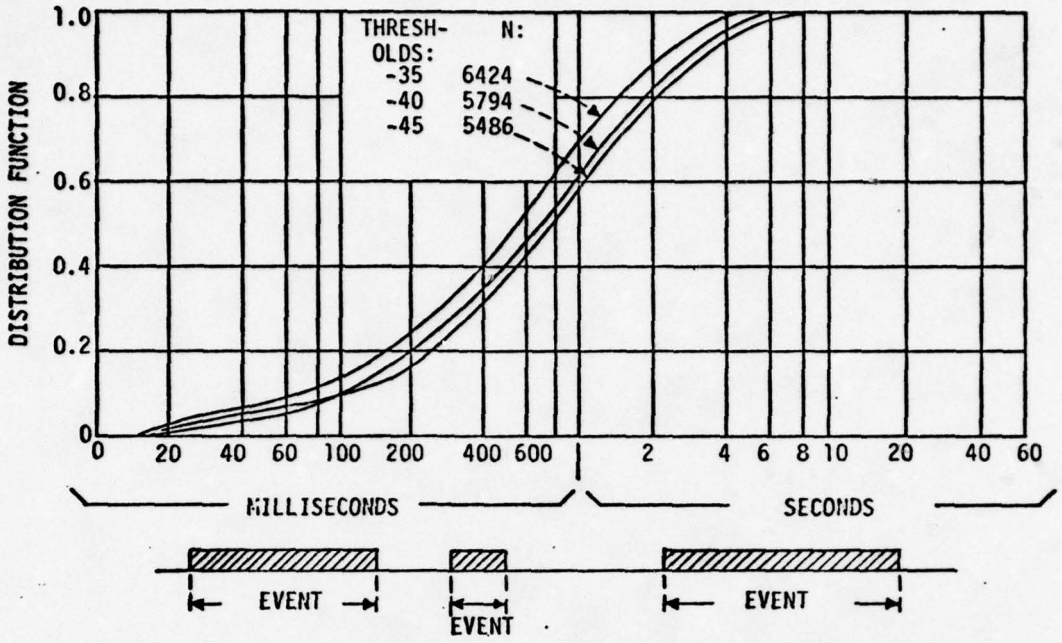
The purpose of this section is to compare mathematical models implying geometric distributions and the empirical data. We do this with the [BRADY, 1967] data.

First a word of caution on the validity of the approach. Our talkspurts consist of a sequence of packets each of which is generated if the speech energy is above a certain threshold during a frame. We have already indicated that this would imply possible inclusion of a short pause within a packet. If the experiments were perfect (no noise, instantaneous and unambiguous detection of silence and speech) the two definitions could not be reconciled. However, it was pointed out in the definition of talkspurt that the experiments were not perfect, in the sense that a "talkspurt" of length less than 15 ms. was discarded, and any "pause" less than 200 ms. was not counted as a pause, but as speech. Thus we see that the two definitions are practically equivalent, provided that the h defined above is less than or equal to 100 ms. With only a minor disparity, we may even let h be as high as 200 ms.

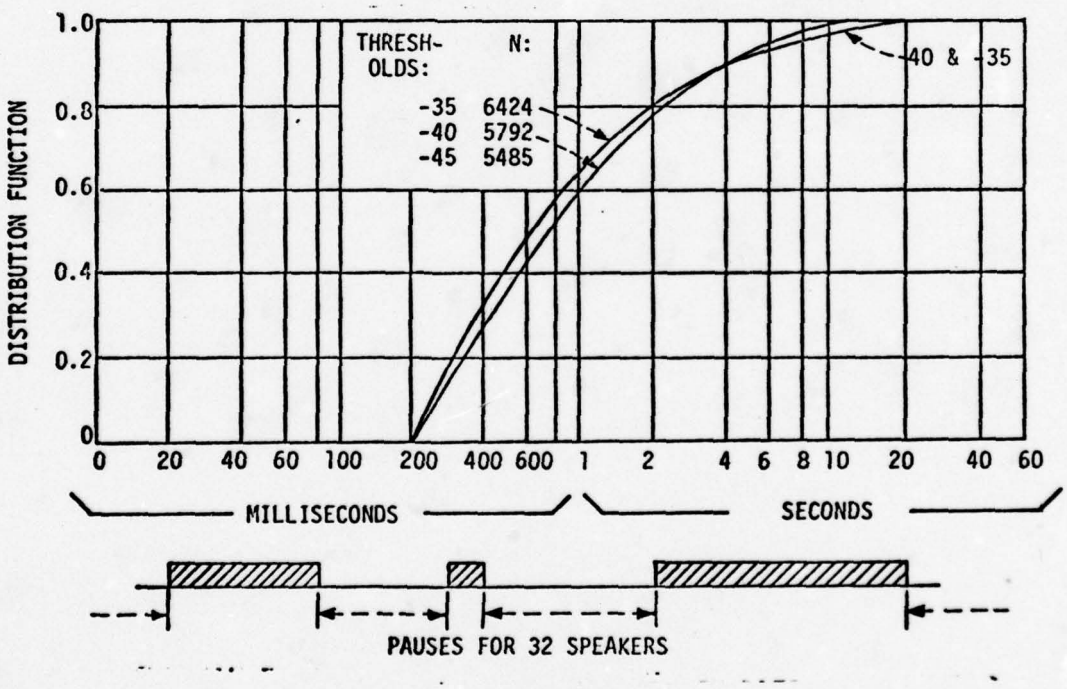
Figure 12, taken directly from [BRADY, 1967], depicts the empirical distribution of the length of a talkspurt and a pause (we restrict our attention to one threshold value). Figure 13 shows a redrawing of the talkspurt length distribution with the following change:

> The x-axis has been converted from talkspurt length to number of packets by the relation:

$$\text{Number of packets} = \left[\frac{\text{talkspurt length}}{h}\right] \qquad (34)$$



TALKSPURTS FOR 32 SPEAKERS



PAUSES FOR 32 SPEAKERS

FIGURE 12: BRADY'S EMPIRICAL SPEECH DATA

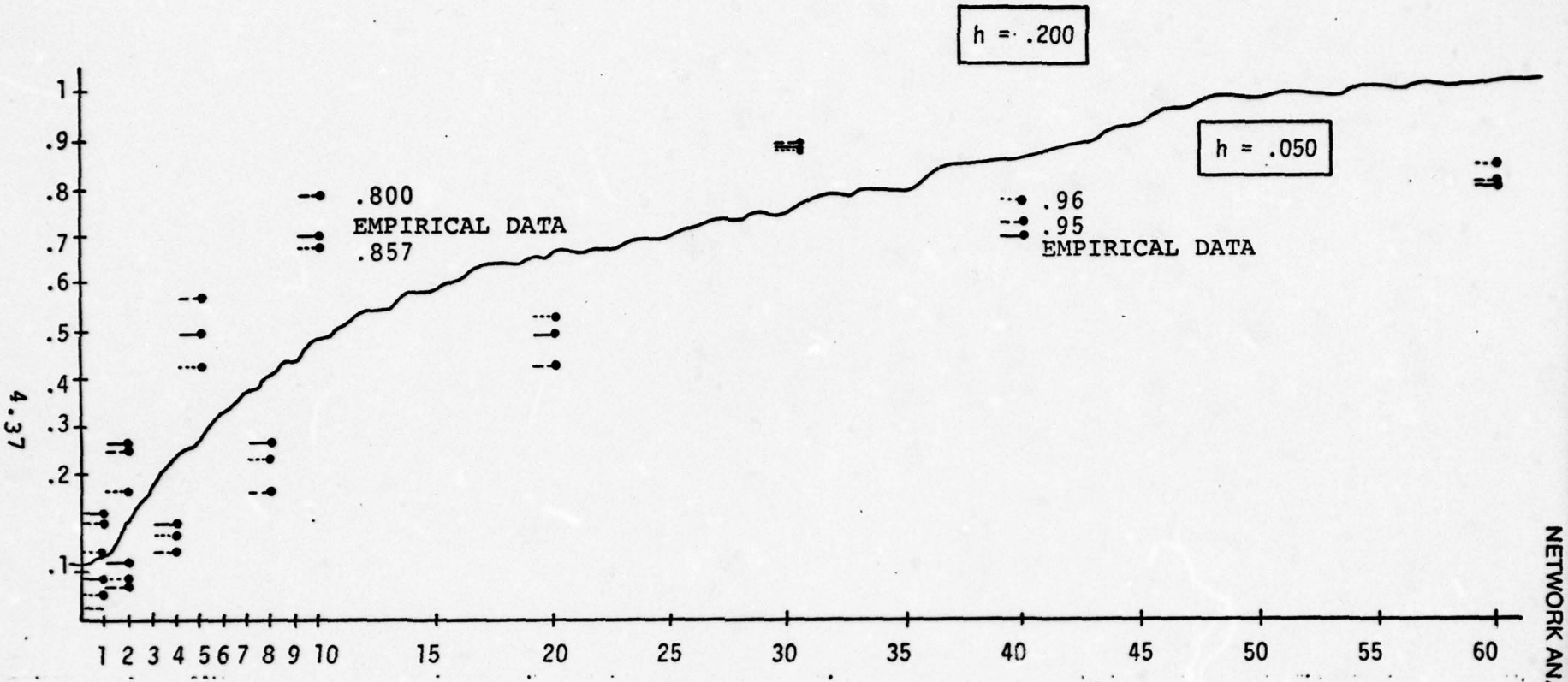


FIGURE 13: COMPARISON OF EMPIRICAL TALKSPURT LENGTH DATA WITH VARIOUS GEOMETRIC MODELS

NETWORK ANALYSIS CORPORATION

where

$$[x] = \inf(n),\ n \text{ integer.}$$
$$x \leq n$$

The empirical data is compared with two geometric distributions (this is done for two values of packet length). We note that the fit is reasonable, but not perfect. The major deficiency is an understatement of the short talkspurt lengths and an overstatement of the long talkspurt lengths. Appropriate selection of the model parameter, e.g., making the mean talkspurt length equal to the empirical mean, results in an accurate prediction of the long talkspurts, but underestimate of the short ones.

To improve the fit we can resort to the expanded model of Section 4.2.3.5 where we have two states for "A talks, B silent." We now compare the talkspurt length distribution accuracy with that predicted by the simpler two-state model. It can be shown that

$$\text{Prob}(Z_p=n) = \alpha(1-p_{10})^{n-1}p_{10} + \beta(1-p_{20})^{n-1}p_{20} \qquad (35)$$

where

α = Prob(issuing a short burst)
β = Prob(issuing a long burst)
p_{10} = Prob(a silent frame follows a long talkspurt frame)
p_{20} = Prob(a silent frame follows a short talkspurt frame).

Naturally, we require $\alpha+\beta=1$. Letting Y_1 and Y_2 be geometric random variables with parameters p_{10} and p_{20}, respectively, we get

$$E(Z_p) = \alpha E(Y_1) + \beta E(Y_2) = \frac{\alpha}{p_{10}} + \frac{\beta}{p_{20}} \qquad (36)$$

and

$$G_{Z_P}(m) = \alpha G_{Y_1}(m) + \beta G_{Y_2}(m) = \alpha(1-q_1^m) + \beta(1-q_2^m) \qquad (37)$$

where G_R is the cumulative distribution of random variable R and $q_i = 1-p_{io}$. Figure 14 has a comparison of the actual data with the model just derived. A major improvement is obtained particularly at the two endpoints of the interval.

The parameter values for the plots of Figure 14 are:

$$h = .050 \begin{cases} \alpha = .8 & \beta = .2 \\ q_1 = .93 & q_2 = .99 \end{cases} \qquad h = .200 \begin{cases} \alpha = .8 & \beta = .2 \\ q_1 = .80 & q_2 = .90 \end{cases}$$

So a speaker issues a "short" speech burst 80% of the time, and "long" bursts the remaining 20%. A more sophisticated selection of parameters may yield an even tighter fit.

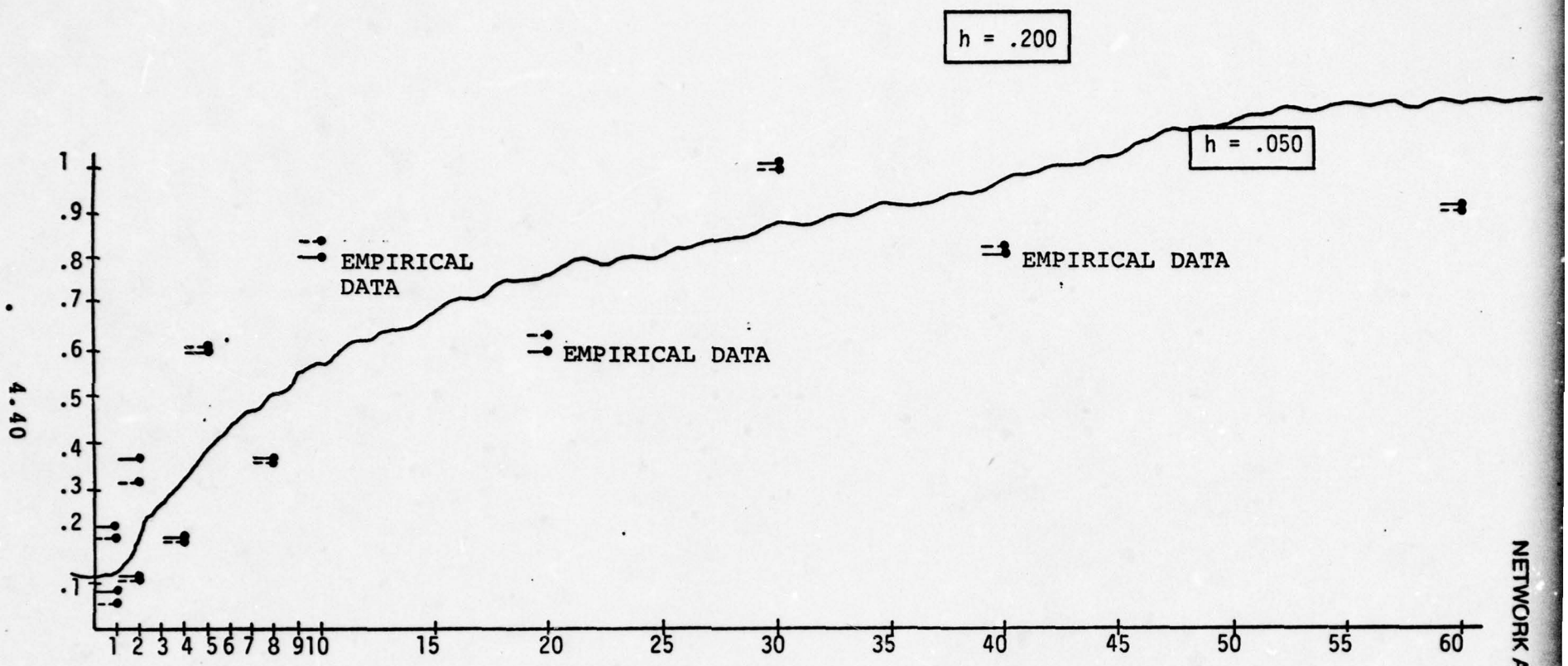


FIGURE 14: COMPARISON OF EMPIRICAL TALKSPURT LENGTH DATA WITH IMPROVED GEOMETRIC MODEL

NETWORK ANALYSIS CORPORATION

4.2.6 Call Origination Model

Thus far we have considered the behavior of a population of m terminals, where m is fixed. In this section we address the issue of the call origination process.

Let M be the maximum number of speakers that can access the packet switch under consideration; let $m^{(n)}$ be the number of off-hook terminals (i.e., active terminals) at frame n.

In standard telephone traffic theory it is assumed that there is an infinite number of traffic sources ($M=\infty$) and that traffic is offered via an infinite number of access lines; it is, however, already recognized that this assumption is not fulfilled in practice; the redeeming feature is that the number of access lines is usually much larger than the number of trunks available [SIEMENS, 1974]. We can assume that M is a relatively small number, (a large commercially obtainable channel of 1.5 MBS would support only thirty 50 KBS PCM terminals). This finite nature of M will be used in Section 4.4, when deriving the queuing model for the single link.

Two issues need to be addressed:

1. The statistical behavior of $m^{(n)}$.

2. The length (holding time) of a typical call.

While specific answers can be given only with exact data from the community of users for which the network is intended, mathematical models which capture the flavor of the phenomenon (rather than the exact numbers) are easily constructed.

4.2.6.1 Standard Models

In standard telephone theory [SIEMENS, 1974], [COOPER, 1972]:

1a. The probability $P_j(t)$ that there are j terminals active at time $t \geq 0$ has the Poisson distribution

$$P_j(t) = \frac{[\lambda tp(t)]^j}{j!} e^{-\lambda tp(t)} \quad j=0,1,2... \tag{38}$$

with time dependent mean $\lambda tp(t)$ where

$$p(t) = 1-H(t) + \int_o^t \frac{x}{t} dH(x) \tag{39}$$

where H(x) is the holding time distribution and λ is arrival rate.

1b. In the steady state

$$P_j = \lim_{t\to\infty} P_j(t) = \frac{a^j}{j!} e^{-a} \quad j=0,1,2 \tag{40}$$

where $a=\lambda \int_0^\infty x dH(x)$ (41)

2. The holding time is exponentially distributed; i.e., the probability H(T) that an existing call will continue to exist throughout and beyond the time T following the instance considered is

$$H(T) = e^{-\frac{T}{t_h}} \tag{42}$$

where t_h is the mean holding time for the servers.

4.2.6.2 Models for Present Investigation

The model to be employed in our investigation has the advantages of:

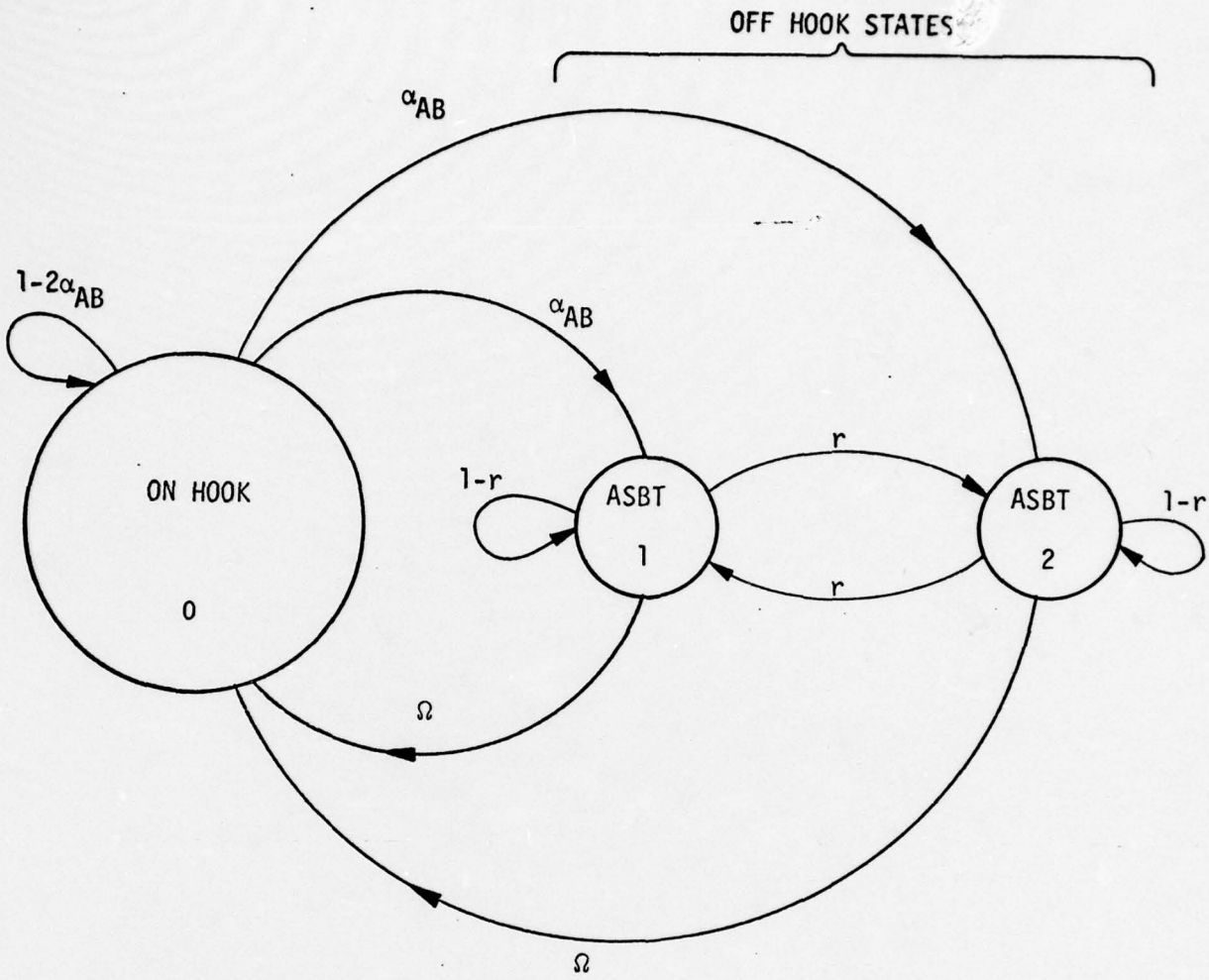
1. Retaining the major traits of the standard telephone model;

2. Interfacing easily with the speaker models of Section 4.2 and with the queuing model of Section 4.4.

We will model the speech originating process by a more complex Markov chain which drives the imbedded speaker models. Figures 15 and 16 illustrate the technique for the two-state and three-state speech models respectively. Let the new chain be denoted by $Y_i^{(n)}$. Note that the holding time (i.e., the time the $Y_i^{(n)}$ is in a state greater or equal to 1) is geometrically distributed in each case, since $Y_i^{(n)}$ remains in state 0 with probability $1-s\alpha$, where s is the number of states in the speech chain and α is the probability of a transition from an on-hook state to each of the off-hook states. This is consistent with the memoryless holding time standard assumption.

It is clear (see item 2, Section 4.2.6.1) that by appropriately selecting the parameter Ω we can make the conversation statistically longer or shorter; also we can control the number of transitions to the off-hook state, namely the number of calls the pair A-B is likely to make in a time interval t.

To answer the question on the number of users that are active, we can collapse all the off-hook states into one state, as in Figure 17. Note that the call termination probability is Ω, not 2Ω, as might be expected at first, since given that we are in state 2 or 3 of the imbedded speaker model chain, the termination probability is indeed Ω. As before, with Q's (q's) replacing P's (p's) in our notation:



$$\begin{pmatrix} 1-2\alpha_{AB} & \alpha_{AB} & \alpha_{AB} \\ \Omega & 1-r-\Omega & r \\ \Omega & r & 1-r-\Omega \end{pmatrix}$$

α, Ω VERY SMALL

FIGURE 15: CALL ORIGINATION MODEL INTERFACED TO A TWO-STATE SPEECH MODEL

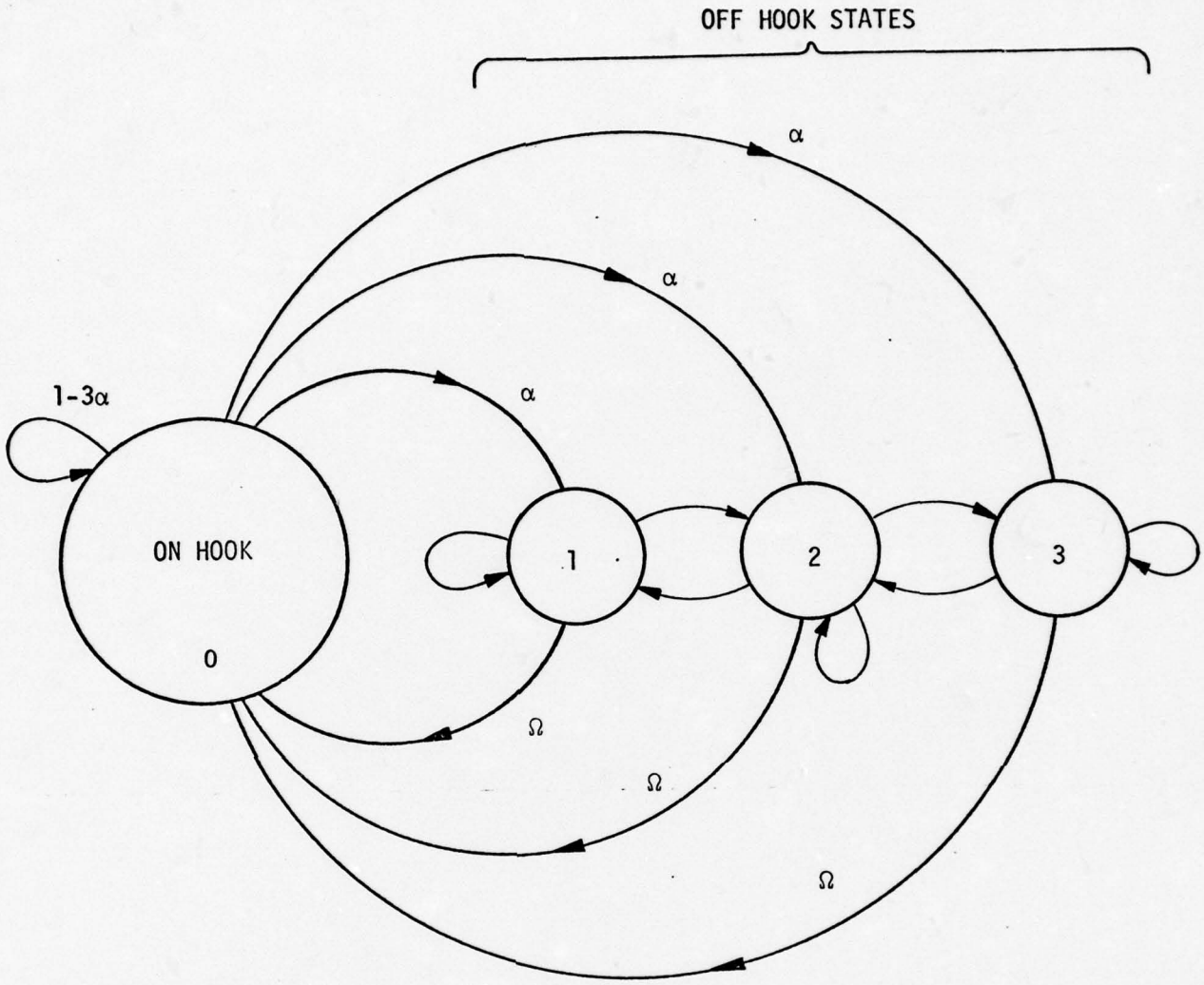
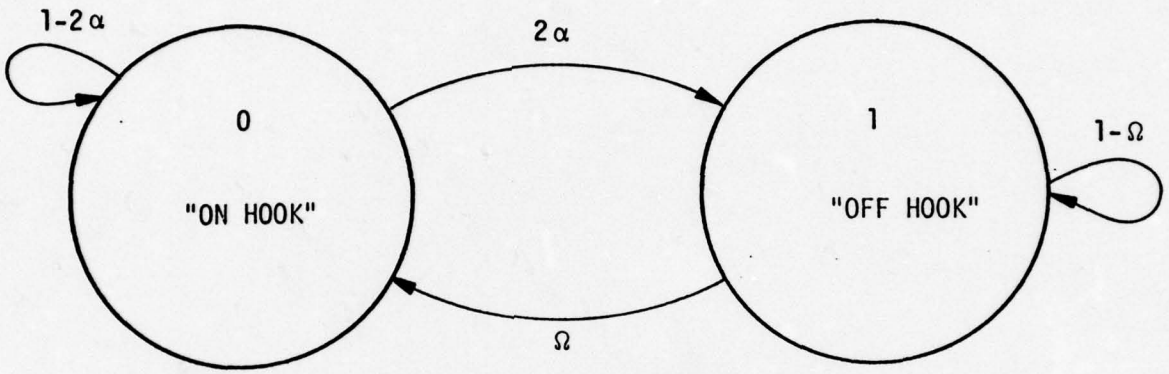


FIGURE 16: CALL ORIGINATION MODEL INTERFACED TO A THREE-STATE SPEECH MODEL



$$Q = \begin{pmatrix} 1-2\alpha & 2\alpha \\ \Omega & 1-\Omega \end{pmatrix}$$

FIGURE 17: BEHAVIOR OF CALL ORIGINATION MODEL

$$Q^n = \begin{pmatrix} q_{01}^{(n)} & q_{01}^{(n)} \\ q_{10}^{(n)} & q_{11}^{(n)} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\Omega}{\Omega+2\alpha} + 2\alpha \frac{(1-\Omega-2\alpha)^n}{\Omega+2\alpha} , & \frac{2\alpha}{\Omega+2\alpha} - 2\alpha \frac{(1-2\Omega-2\alpha)^n}{\Omega+2\alpha} \\ \frac{\Omega}{\Omega+2\alpha} - \Omega \frac{(1-\Omega-2\alpha)^n}{\Omega+2\alpha} , & \frac{2\alpha}{\Omega+\alpha} + \Omega \frac{(1-\Omega-2\alpha)^n}{\Omega+2\alpha} \end{pmatrix}. \quad (43)$$

Thus

$$\left.\begin{aligned} Q_0^{(n)} &= Q_0^{(0)} q_{00}^{(n)} + Q_1^{(0)} q_1^{(n)} \\ Q_1^{(n)} &= Q_0^{(0)} q_{01}^{(n)} + Q_1^{(0)} q_{11}^{(n)} \end{aligned}\right\} \quad (44)$$

and

$$\left.\begin{aligned} Q_0 &= \frac{\Omega}{2\alpha+\Omega} \\ Q_1 &= \frac{2\alpha}{2\alpha+\Omega} . \end{aligned}\right\} \quad (45)$$

We have

$$m^{(n)} = \sum_{i=1}^{M} Y_i^{(n)} \quad (46)$$

and

$$\text{Prob}(m^{(n)}=k) = \binom{M}{k} [Q_1^{(n)}]^k [Q_0^{(n)}]^{M-k} \quad (47)$$

where $Q_1^{(n)}$ and $Q_0^{(n)}$ are given above. This expression corresponds to item (1a) of Section 4.2.6.1 and represents the instantaneous call origination distribution. The steady state distribution is

$$\text{Prob}(m^{(\infty)}=k) = \binom{M}{k}\left(\frac{2\alpha}{2\alpha+\Omega}\right)^k\left(\frac{\Omega}{2\alpha+\Omega}\right)^{M-k} \tag{48}$$

Note that as $M\to\infty$, if the average number of off hook customers is kept at a constant a, the distribution of active users approaches

$$\text{Prob}(m^{(n)}=k) = \frac{a^k}{k!}e^{-a} \quad k=0,1,2... \tag{49}$$

exactly equivalent to item (1b) of Section 2.6.1! We see that our call origination model is consistent with the standard telephony models.

For an s-state speech model,

$$\left.\begin{aligned} Q_0 &= \frac{\Omega}{s\alpha+\Omega} \\ Q_1 &= \frac{s\alpha}{s\alpha+\Omega} \end{aligned}\right\} \tag{50}$$

$$\text{Prob}(m^{(\infty)}=k) = \binom{M}{k}\left(\frac{s\alpha}{s\alpha+\Omega}\right)^k\left(\frac{\Omega}{s\alpha+\Omega}\right)^{M-k} \tag{51}$$

Note, however, that Q_0 and Q_1 should not depend explicity on the speech model used; in particular the product $s\alpha$ should be independent of s.

General statistics on the population's call habits, along with the steady state probabilities Q_0 and Q_1, can be employed to determine empirically the values of α and Ω. In fact, if t_h is the average length of a call (in number of frames), it follows that

$$\Omega = \frac{1}{t_h} . \tag{52}$$

If the fraction of time a given voice terminal is being used over a time interval (say, a working day) is a known empirical value F,

$$\alpha = \frac{F}{1-F}\frac{1}{st_h} . \tag{53}$$

(Note that, as required, the product $s\alpha$ does not depend on the speech model.)

The formula breaks down if either F=1 and/or t_h=0 (some states are absorbing); however, on other grounds, F=0 implies α=0; F=1 implies α=1.

For typical situations, a frame is around 100 ms, a conversation is around 10 minutes and an individual may make 12 phone calls a day (8 hour shift); this yields F=.25, t_h=6000. For s=2, Ω=.000166, α=.000027. Note that in this case there is a higher tendency for a call to stop than for a new call to start; if F were increased to .75 we would get Ω=.000166, α=.000249 - the situation has reversed. Under the earlier situation, F=.25, the expected number of active users is .25M out of M potential callers.

If the call origination section is shut off by setting Ω=0 and $\alpha \neq 0$, conversations of infinite length are produced. This yields conservative results since we then have the worst case traffic - every potential speaker is active, i.e.,

$$\left.\begin{array}{l} \text{Prob}(m^{(\infty)} = M) = 0 \\ \text{Prob}\ (m^{(\infty)} = M) = 1. \end{array}\right\} \qquad (54)$$

These probabilities, however, cannot be obtained from the steady state probabilities, since state 1 is absorbing when Ω=0. Whether the call origination model is used or not (i.e., Ω=0, for worst case analysis), the major part of this report - except Section 4.5.7 - uses steady state values for all unconditional probabilities; this is not a requirement of the models to be developed, but is done for descriptive and notational ease.

Various approaches can now be taken to compute the probabilities to be employed in Section 4.4; we outline two methods, using the two-state speaker chain as an example.

The chain of Figure 15 has steady state probabilities

$$P_0 = \frac{\Omega}{\Omega+2\alpha_{AB}}$$

$$P_1 = \frac{\alpha_{AB}}{\Omega+2\alpha_{AB}} \qquad (55)$$

$$P_2 = \frac{\alpha_{AB}}{\Omega+2\alpha_{AB}}$$

Thus the steady state probability that the pair A-B is conversing is $P_1 + P_2$; the probability that A supplies a packet is P_1; the probability that B supplies a packet is P_2. One could then set up a model using P_1 as the steady state probability that packets are generated by A. This approach suffers from the fact that P_1 is the "average probability" that speaker A furnishes a packet; it includes the long period of inactivity (speaker on-hook) of A, giving a valid number of packets supplied only over a long period of time rather than on a short one. The above is thus poor for peak hour or worst case analysis. To remedy this, we can force a pair to sustain an infinitely long conversation by taking, as indicated, $\Omega=0$. Then, from the above steady state probabilities,

$$P_0 = 0$$

$$P_1 = 1/2 = P \qquad (56)$$

$$P_2 = 1/2 = Q = 1-P.$$

For a speaker model having more than two states, these probabilities will in general, not equal 1/2; we indicate this by using a symbol, P, rather than a specific value.

Another way of obtaining the same values for a peak period or worst case analysis is to consider the steady state probability that a speaker supplies a packet, *given that the speaker is off-hook.* We thus need

$$P = \frac{P_1}{P_1 + P_2}$$

$$Q = \frac{P_2}{P_1 + P_2} \qquad (57)$$

which both equal 1/2, as before.

In the sequel, the worst case assumption is retained, since we are interested in evaluating the potential delay when every subscriber is using the system (this may occur in reality, depending on the environment). One may view this approach as a particularization of the call origination model above (=0); or, alternatively one can interpret the technique as an attempt to determine the delay given that m *fixed speakers* (whether in fact m=M or not) are using the packet switch for an extensive period of time (10 minutes suffices to make the second interpretation acceptable). In any event, the by-product of this assumption is a fixed number m of users accessing the packet switch; this will simplify the modeling effort. A limited investigation for a fluctuating m is considered in Section 4.4.5.

4.3 PERFORMANCE CRITERIA

Well documented subjective testing and measurements have established ranges of transmission network corruptive effects on speech where these effects are either:

1. Not perceptible.

2. Perceptible but tolerable.

3. Not tolerable.

Estimates for the boundaries between these regions are available although further investigation into combinations of effects is needed. The "not perceptible" range is easy to deal with. The "not tolerable" range is slightly deceptive in that it will usually not result in a disruption of the conversation unless sustained for an unacceptable period of time. Thus, transient behavior of the network may be significant. A special complication arises in the middle range, "perceptible but tolerable". In this range speakers may alter their speech habits to compensate for network behavior. Thus, in general, there exists a feedback loop between the speech model parameters and the network behavior. Fortunately, it is only in the "not tolerable" range that speaker behavior leads to unstable degradation (shouting, heavy double talking, no information transfer, etc.). In the middle range the speakers eventually begin to act so as to produce more efficient information transfer (fewer interruptions, etc.); thus speaker models based on ideal network performance will tend to be conservative in estimating the degree of corruptive effects.

4.3.1 Results of Subjective Studies

Packetized speech belongs to the category of real-time data traffic. Consistent with this classification, it has stringent delivery requirements with respect to time, but somewhat tolerant requirements with respect to loss or error. Generally speaking, the delivery requirements can be divided into two categories:

1. Due to the psychological effects induced by delay, the end-to-end *average* network delivery time must be small.

2. Due to the psychological effects induced by "glitching" (gaps due to delay fluctuation, noise, buffer overflow losses and other protocol discardings, misaddressings), the end-to-end *variation* of the delivery time, including losses, must be small.

In other words, the human listener in a conversation has limited tolerance to both the average delay and the fluctuation of delay. It is thus apparent that the network designer must control not only the first moment (mean) of the delay, but also the second moment (variance). In Section 4.3.2 we establish this latter requirement formally.

4.3.1.1 End-to-End Delay

The overall end-to-end delay can be written as

$$\hat{D}(t) = V + h + d(t) + B, \qquad (58)$$

where V is the delay due to the speech analog-to-digital conversion h is the delay due to the packetizing period (equivalent to the packet time length), d(t) is the network delay at time t and B is the

waiting time at the receiver end (as in [COHEN, 1976]). The precise value of V depends on the terminal technology and is usually small compared to the other terms; we will ignore it in the sequel. Therefore, if D=h+d, D represents the total end-to-end delay before the application of any receive end buffering. The overall delay, $\hat{D}$, should not exceed 600 milliseconds, the value of delay that has been shown to be "commercially acceptable" by [KLEMMER, 1963], [KLEMMER, 1967], [FORGIE, 1975] and others. When the delay reaches 1200 milliseconds adverse psycological factors impede normal telephonic conversation, as shown in the above-mentioned papers, and in [KRAUSS, 1966], [BRADY, 1970], among others. A delay between 600 ms. and 1200 ms. is conditionally acceptable for a short portion of the conversation, when the occurrences of such delays are rare and far apart. In other words, there is a well established range of acceptable delay, and temporary degradation is admissible, as long as such degradation occurs with low probability and short duration. Particular applications may require more stringent constraints.

4.3.1.2 Glitching

Studies have been conducted where speech is temporally segmented and temporally interrupted at constant (deterministic) rates, [HUGGINS, 1976], or where speech is manipulated according to some random process [NSC, 1976]. The following results have been shown.

1. In interrupted speech (equivalent to loss or discard of packets)

    a. Intelligibility decreases to very low values (10%) as the packet size approaches .25 seconds.

b. Intelligibility increases to 80% as packet size approaches .019 seconds.

2. In segmented speech (equivalent to waiting for late packets)

a. For fixed active speech segment length, intelligibility increases as the silence period decreases.

b. For fixed silence length, intelligibility decreases as the active speech segment length decreases.

Curing suggestions, such as short packets and interleaving, have been offered by [HUGGINS, 1976] and [FORGIE, 1976]. The case where such alterations occur at random rather than regular instances has been studied by Forgie at Lincoln Labs. Official documentation does not exist at this point, but is expected.

Tapes played at [NSC, 76] seem to indicate:

1. Glitch rate over .5% is unacceptable.

2. Waiting for late packets is a bad policy, even if one plays back such old packets at higher speed.

Due to the high redundancy of the speech signal, speech loss as high as 50% can be sustained with marginal degradation, if such loss occurs for very small (e.g., 19 ms.) segments. See [HUGGINS, 1976]. This concept might be employed to control total network traffic in case of congestion. Thus, the acceptable packet loss rate is a function of packet size. Under certain speech encoding techniques, such as vocoding, the packets themselves may be composed of self-contained speech elements, called parcels, whose selective discarding could be used as a traffic throttle. With

PCM, for example, if an eight level quantification is used, each eight bit character can be considered a parcel.

4.3.2 Smoothness Criteria

The reconstructed continuous speech delivered to a listener by a packet-switched network contains gaps due to the statistical fluctuation of the network load and the consequent fluctuation of the network delay and loss performance. The gap structure perceived by the listener will not only be a function of these fluctuations but also a function of the network policy or protocol at the receiver end in dealing with these gaps. In this section we show the importance of obtaining the delay distribution, or at least the second moment of delay. For data traffic the mean delay has usually been the only design criterion.

4.3.2.1 Waiting for Late Packets

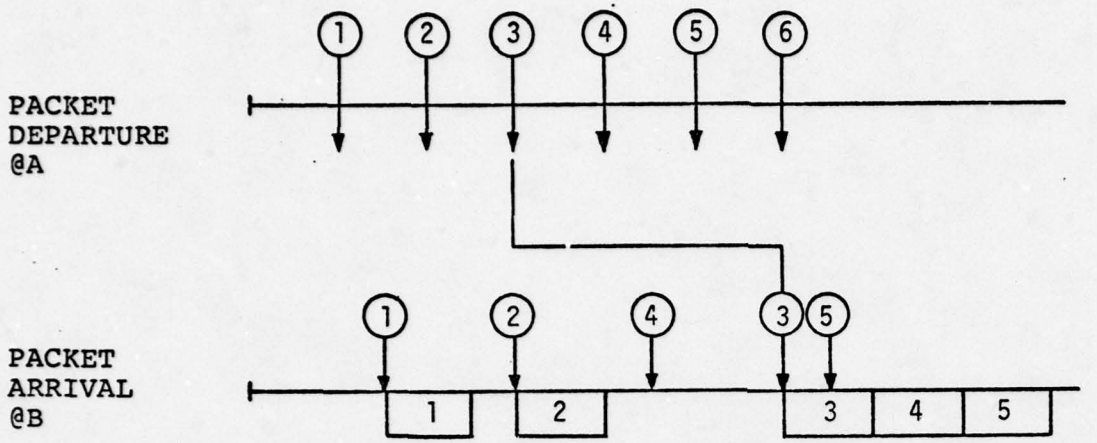
Consider an A talkspurt, i.e., a segment of continuous active speech between a speaker-listener pair A-B with a consecutive stream of non-empty packets issued by speaker A. Assume that at frame i, i=1,2,..., a packet issued by A experiences delay D_i, where the D_i are identically (but not independently) distributed for all i. Then for a fixed packet time length h, a_i=ih is the time when A issues packet i, and $b_i = D_i + ih$ is the time when B receives packet i. Note that, as defined here, h is identical to the packetizing delay in our previous delay formula. Let f_i be the temporal fluctuation between packets i-1 and i, as received by B. More precisely (see Figure 18),

$$f_i = D_i - D_{i-1} + \min(0, f_{i-1}), \quad i=1,2,..., \tag{59}$$

with the initial conditions $D_o = 0$, $f_o = 0$.

Then f_i represents the a posteriori lateness of the i^{th} packet relative to the lateness of the $(i-1)^{st}$ packet.

Since $0 \leq D_i \leq \infty$, the range of f_i is $-\infty \leq f_i \leq \infty$.



D_0	D_1	D_2	D_3	D_4	D_5
0	1	2	6	1	3

$f_1 = 1-0+0=1 \quad Z_1=0 \quad G_1=1$

$f_2 = 2-1+0=1 \quad Z_2=0 \quad G_2=1$

$f_3 = 6-2+0=4 \quad Z_3=0 \quad G_3=4$

$f_4 = 1-6+0=-5 \quad Z_4=5 \quad G_4=0$

$f_5 = 3-1-5=-3 \quad Z_5=-3 \quad G_5=0$

FIGURE 18: UNLIMITED WAITING PROTOCOL

NETWORK ANALYSIS CORPORATION

If $f_i \leq 0$, corresponding to the i^{th} packet arrival not later than the earliest time that an in-order delivery to B can be made, no temporal disruption results, provided that adequate buffering facilities exist at the receiver end.

For this protocol, the gaps in the reconstructed speech are simply

$$G_i = \max(f_i, 0) = \begin{cases} f_i & \text{if } f_i > 0 \\ 0 & \text{if } f_i \leq 0. \end{cases} \tag{60}$$

That is, gaps are introduced only when f_i is positive.

Clearly, the network should be designed such that the distribution of G_i is acceptable from a performance point of view. In particular we may require

$$\text{Prob}(G_i > K_\alpha) \leq \alpha, \tag{61}$$

with $K_\alpha > 0$.

We now simplify our notation by introducing

$$Z_i = \min(0, f_i) = \begin{cases} f_i & \text{if } f_i \leq 0 \\ 0 & \text{if } f_i > 0. \end{cases} \tag{62}$$

Thus G_i and Z_i are the respective positive and negative parts of f_i so that

$$f_i = G_i + Z_i. \tag{63}$$

Restating our earlier definition of f_i,

$$f_i = D_i - D_{i-1} + Z_{i-1} \tag{64}$$

and the last two equations yield

$$G_i = (D_i - Z_i) - (D_{i-1} - Z_{i-1}). \tag{65}$$

Note that the Z_i are *not* identically distributed (otherwise we could conclude that $E(G_i)=0$ which is obviously untrue, in general). For a strictly positive K_α we have

$$\text{Prob}(G_i > K_\alpha) = \text{Prob}(f_i > K_\alpha) = \text{Prob}(D_i - D_{i-1} > K_\alpha - Z_{i-1}). \tag{66}$$

Since Z_{i-1} is always non-negative, we enlarge the right-hand-side event by eliminating it from the event expression and obtain the inequality

$$\text{Prob}(G_i > K_\alpha) \leq \text{Prob}(D_i - D_{i-1} > K_\alpha). \tag{67}$$

But we further have

$$\text{Prob}(D_i - D_{i-1} > K_\alpha) \leq \text{Prob}(|D_i - D_{i-1}| > K_\alpha) \tag{68}$$

and by the Chebyshev inequality

$$\text{Prob}(|D_i - D_{i-1}| > K_\alpha) \leq \frac{V(D_i - D_{i-1})}{K_\alpha^2} \tag{69}$$

where V is the variance function. Since the D_i are identically distributed we have

$$V(D_i - D_{i-1}) = 2\sigma_D^2 (1-\rho_D) \tag{70}$$

where σ_D^2 is the variance of the common distribution of the D_i and ρ_D is the first order correlation coefficient of the delays; ρ_D

is bounded between -1 and 1, but likely to be positive (between 0 and 1) for reasonable network behavior. That is, consecutive packets in a talkspurt are likely to experience similar delays. Putting together all of the above equalities and inequalities we arrive finally at

$$\text{Prob}(G_i > K_\alpha) \leq \frac{2\sigma_D^2(1-\rho_D)}{K_\alpha^2}. \tag{71}$$

Thus for any performance criteria where we wish to control the tail of the gap distribution, given a K_α, we can impose constraints on σ_D^2 so that the probability of gaps exceeding K_α can have arbitrarily small probability. Note that K_α itself can be made arbitrarily small - certainly sufficient in an engineering sense - but that it cannot be made strictly zero.

We point out that the bound derived above for controlling the gap distribution tail will not necessarily be tight in the sense that the constraint implied on σ_D^2 may be more than necessary to achieve the control. Going further along this line, if ρ_D is not conveniently obtained, since $|\rho_D| \leq 1$, we have

$$\frac{2\sigma_D^2(1-\rho_D)}{K_\alpha^2} \leq \frac{4\sigma_D^2}{K_\alpha^2}. \tag{72}$$

Thus we can insure the original constraint, namely

$$\text{Prob}(G_i > K_\alpha) \leq \alpha \tag{73}$$

by satisfying

$$\sigma_D^2 \leq \frac{\alpha K_\alpha^2}{4}. \tag{74}$$

Note that as α decreases - the amount of area allowed in the tail - σ_D^2 must decrease. Also as K_α decreases - the cut off point for the tail - the needed σ_D^2 decreases.

4.3.2.2 Limited Waiting for Late Packets

Under the protocol in the last subsection we showed that the gaps in the reconstructed speech are

$$G_i = \max(0, f_i). \tag{75}$$

If the designer of the network could control the variance of the delay then the tail of the gap distribution would also be controlled. The variance, however, cannot be completely controlled nor, in general, reduced to any arbitrary value. Therefore an alternate protocol must be sought to prevent arbitrarily long gaps.

Waiting indefinitely for late packets at the receiver end (say, by buffering subsequent packets) implies not only a long gap in the reconstructed speech until such late packets arrive, but also temporal distortion of the consecutive spoken material from that point on. Also, such a protocol implies no recovery from lost or misaddressed packets. To avoid these complications, a protocol can declare missing, and subsequently ignore, a packet whose lateness exceeds a certain preset limit S. We will call such an action a discard. Assume that $(J-1)h \leq S \leq Jh$, where J is a positive integer. By substituting a period of silence of length Jh whenever a packet does not arrive in time, the temporal distortion of the overall speech string can be bounded by a predetermined value. Note that J=1, for example, implies a temporal distortion of at most S, and a gap of at most h, before the protocol is reapplied to the next packet. In this case a single packet's lateness can contribute at most h to a gap or S to the temporal distortion, but not both. Maximal temporal distortion occurs when the packet arrives infinitesimally before the discard decision, in which case the packet is still delivered. Further improvement can be obtained if the silent period following rejection is aborted upon the arrival of the next packet, but we have not yet investigated this alternative.

A new gap structure is implied by this protocol. We can, as before, define a relative lateness function for the i^{th} packet, whose positive part represents the gap caused by the i^{th} packet. The situation is described by the following set of equations:

$$f_o = 0 \qquad \tilde{D}_0 = 0 \tag{76}$$

$$f_i = D_i - \tilde{D}_{i-1} + \min(0, f_{i-1}) \tag{77}$$

$$\tilde{D}_i = \begin{cases} D_i & f_i \leq S \\ \tilde{D}_{i-1} - \min(0, f_{i-1}) & f_i > S \end{cases} = \begin{cases} D_i & f_i \leq S \\ D_i - f_i & f_i > S \end{cases} \tag{78}$$

$$g_i = \begin{cases} 0 & f_i \leq 0 \\ f_i & 0 < f_i \leq S \\ Jh & S < f_i . \end{cases} \tag{79}$$

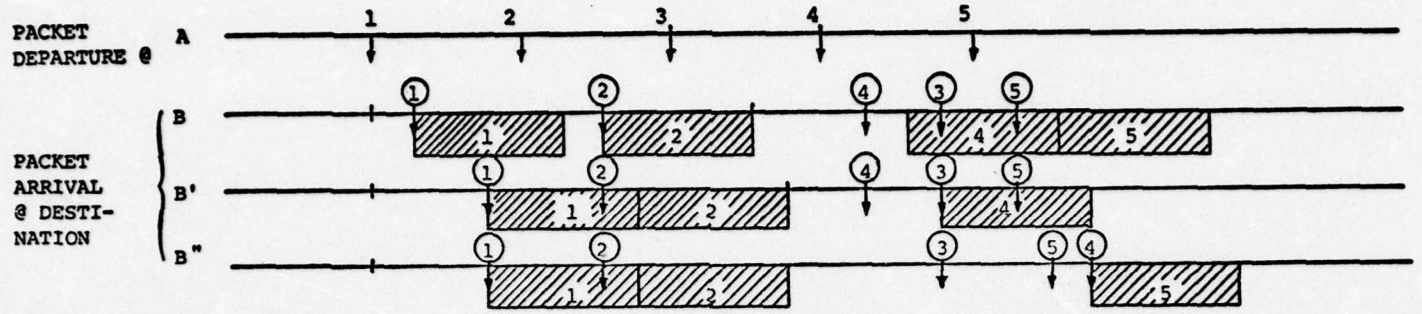
Note that the length of time the protocol waits for a late packet depends not on the value of delay, but on the lateness relative to delivery need. Figure 19 has three samples of this protocol.

If $S=\infty$, the previous protocol and equations are obtained. Under the new protocol there is an explicit tradeoff between the packet discard rate and the temporal distortion; as S increases more temporal distortion occurs but fewer packets are discarded.

An overall gap may be made up of several of the subgaps g_i. Define G_i, only for those packets, i, which are actually delivered to the listener, as the gap between the last delivered packet and packet i. Then

$$G_i = \sum_{k=j+1}^{i} g_k \tag{80}$$

where j is the frame number of the last delivered packet, prior to packet i. Let N_i be the number of consecutively discarded packets immediately preceding frame i, given that packet i is delivered.



B					
D_1	D_2	D_3	D_4	D_5	
1	2	7	1	1	

B'					
D_1	D_2	D_3	D_4	D_5	
3	2	7	1	1	

B"					
D_1	D_2	D_3	D_4	D_5	D_6
3	2	7	7*	1	1

PROTOCOL PARAMETERS: h=4, s=3, J=1

*(IF D_4=5 WE ACCEPT PACKET 4)

$f_1=D_1-\tilde{D}_0+\min(0,f_0)=1$
$g_1=1$
$\tilde{D}_1=D_1=1$

$f_2=D_2-\tilde{D}_1+\min(0,f_1)=1$
$g_2=1$
$\tilde{D}_2=D_2=2$

$f_3=D_3-D_2+\min(0,f_2)=5$
$g_3=4$ ($\Rightarrow$ discard packet 3)
$\tilde{D}_3=2$

$f_4=D_4-\tilde{D}_3+\min(0,f_3)=-1$
$g_4=0$
$\tilde{D}_4=1$

$f_1=D_1-\tilde{D}_0+\min(0,f_0)=3$
$g_1=3$
$\tilde{D}_1=D_1=3$

$f_2=D_2-\tilde{D}_1+\min(0,f_1)=-1$
$g_2=0$
$\tilde{D}_2=D_2=2$

$f_3=D_3-\tilde{D}_2+\min(0,f_2)=4$
$g_3=4$ ($\Rightarrow$ discard packet 3)
$\tilde{D}_3=3$

$f_4=D_4-\tilde{D}_3+\min(0,f_3)=-2$

$f_1=D_1-\tilde{D}_0+\min(0,f_0)=3$
$\tilde{D}_1=D_1=3$

$f_2=D_2-\tilde{D}_1+\min(0,f_0)=-1$
$\tilde{D}_2=2$

$f_3=D_3-D_2+\min(0,f_2)=4$
($\Rightarrow$ discard packet 3)
$\tilde{D}_3=3$

$f_4=D_4-\tilde{D}_3+\min(0,f_3)=4$
($\Rightarrow$ discard packet 4)
$\tilde{D}_4=3$

FIGURE 19: LIMITED WAITING PROTOCOL

NETWORK ANALYSIS CORPORATION

Therefore, after minor manipulation

$$G_i = N_i Jh + g_i. \qquad (81)$$

This equation points out clearly the tradeoff between the packet discard rate and the temporal distortion. The gaps under the present protocol are made up of two factors:

1. N_i: number of discarded packets immediately prior to frame i.

2. g_i: gap in waiting for delivery of packet i.

The effect of S on the gap structure is as follows:

1. If S is large:

   a. N_i is close to zero, with high probability.

   b. g_i is potentially large.

   Therefore the major contributor to the gaps is the waiting time for a delivered packet (similar to the protocol of the previous section).

2. If S is small:

   a. N_i is potentially large.

   b. g_i is close to zero, with high probability.

   Therefore the major contributor to the gap is the silence caused by the discarded packets.

An optimal value of S must exist, as a function of the performance requirements. As before, we want to bound the tail of the gap distribution G_i:

$$\text{Prob}(G_i \geq K_\alpha) \leq \alpha. \tag{82}$$

But

$$\text{Prob}(G_i \geq K_\alpha) = \text{Prob}(N_i Jh + g_i \geq K_\alpha)$$

$$\leq \text{Prob}(N_i Jh \geq K_\alpha/2) + \text{Prob}(g_i \geq K_\alpha/2) \tag{83}$$

Letting $K=K_\alpha/2$, we can bound each component on the right-hand-side separately to get our needed result.

From the result of the previous subsection and the fact that the i^{th} packet is delivered we have

$$\text{Prob}(g_i > K) \leq \begin{cases} \dfrac{4\sigma_D^2}{K^2} & \text{if } K \leq S \\ 0 & \text{if } K > S. \end{cases} \tag{84}$$

Since we need only consider the case when N_i is a positive integer, and since $S \leq Jh$, and K is positive,

$$\text{Prob}(N_i Jh > K) \leq \begin{cases} 0 & \text{if } K \leq S \\ \text{Prob}(N_i Jh > S) & \text{if } K > S \end{cases} \tag{85}$$

But we have

$$\text{Prob}(N_i Jh > S) \leq \text{Prob}(f_{j+1} > S) \leq \frac{4\sigma_D^2}{S^2} \tag{86}$$

for any positive S, where j is again the frame number of the last packet delivered prior to i.

Combining the above results we have

$$\text{Prob}(G_i \geq 2K) \leq \text{Prob}(g_i \geq K) + \text{Prob}(N_i Jh > K) \leq \begin{cases} \frac{4\sigma_D^2}{K^2} & \text{if } K \leq S \\ \frac{4\sigma_D^2}{S^2} & \text{if } K > S \end{cases} \quad (87)$$

or in our original notation

$$\text{Prob}(G_i \geq K_\alpha) \leq \frac{4\sigma_D^2}{[\min(\frac{K_\alpha}{2}, S)]^2} . \quad (88)$$

Although this demonstrates that a control on σ_D is sufficient to guarantee an arbitrarily low probability, further research should be able to provide tighter bounds. Though the bound obtained here, for the tail of gap distribution, is weaker than the bound for the previous protocol, this protocol can guarantee that the cumulative temporal distortion in a talkspurt of n packets is limited to nS.

4.3.2.3 Receive End Buffering

The limited waiting period, S, can be regarded as a delay variance reduction technique wherein we prevent temporal gaps due to a single packet from exceeding S, paying for it with a glitch silence of length Jh. An additional protocol strategy can be employed to reduce the gap variance by buffering packets at the receive end so that their total delay to the receiver is at least some minimum quantity. What is sacrificed with this technique is an increase in the average delay in return for the improved smoothness.

This modification involves buffering at the receiver end those packets whose delay does not exceed a certain appropriately chosen value w. These packets would be stored an amount of time $w-D_i$.

If $\hat{D}_i$ represents the delay for packet i as seen by the receiver terminal, then

$$\hat{D}_i = \max(w,D_i) = \begin{cases} w & \text{if } D_i \leq w \\ D_i & \text{if } D_i > w \end{cases} \tag{89}$$

where D_i is the network delay without the end buffering. If B_i represents the time packet i is buffered (we assume that no packet is discarded) then

$$B_i = \begin{cases} w-D_i & D_i \leq w \\ 0 & D_i > w. \end{cases} \tag{90}$$

(See Figure 20).

Under this protocol the average delay increases, while the variance decreases, as we now show. Let $\hat{H}(t)$ be the distribution of $\hat{D}_i$, $H(t)$ be the distribution of D_i and

$$P_s = \text{Prob}(D_i<w) = \int_o^w dH(t) \tag{91}$$

Then since

$$\hat{D}_i = \max(w,D_i) \tag{92}$$

we have

$$E(\hat{D}_i) = \int_o^\infty t d\hat{H}(t)$$

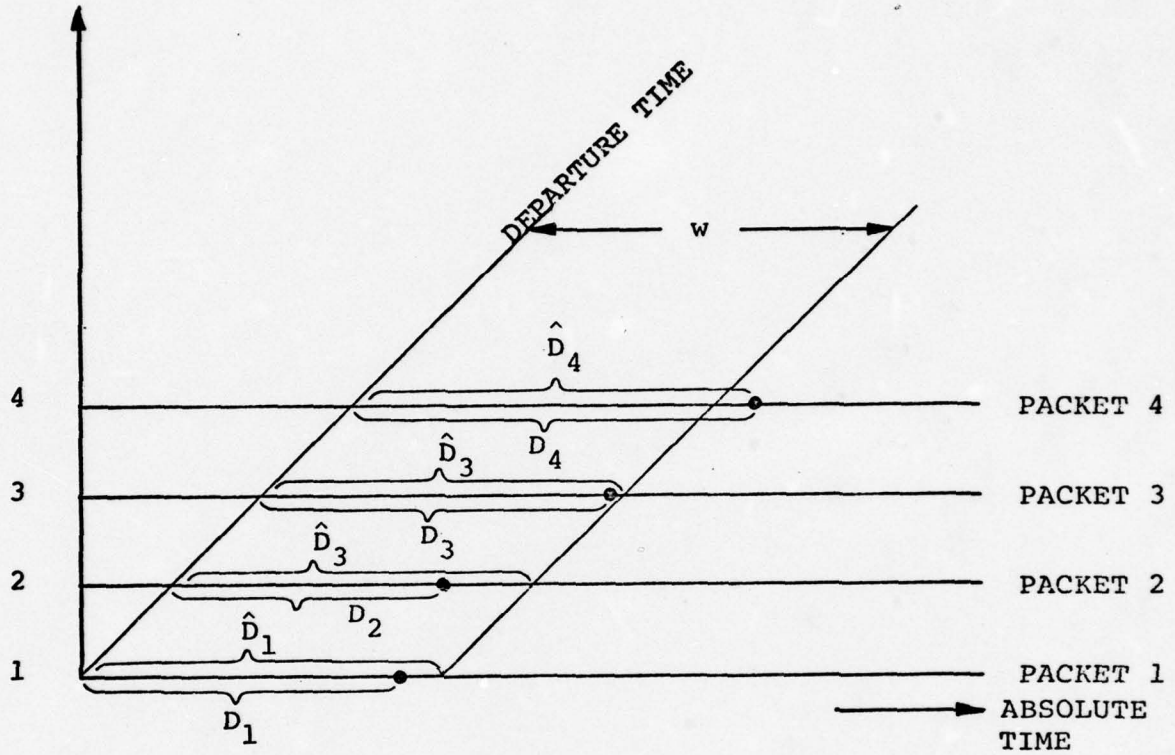


FIGURE 20: END BUFFERING

$$= w P_s + \int_w^\infty t\,dH(t)$$

$$= w\int_o^w dH(t) + \int_w^\infty t\,dH(t)$$

$$= \int_o^w w\,dH(t) + \int_w^\infty t\,dH(t)$$

$$\geq \int_o^w t\,dH(t) + \int_w^\infty t\,dH(t) = E(D_i) \qquad (93)$$

We next can show that $Var(\hat{D}_i) \leq Var\ (D_i)$ as follows:

$$Var(\hat{D}_i) = Var(\max(w, D_i))$$

$$= \int_o^\infty [t-E(\hat{D}_i)]^2\, d\hat{H}(t)$$

$$= [w-E(\hat{D}_i)]^2 \int_o^w dH(t) + \int_w^\infty [t-E(\hat{D}_i)]^2 dH(t)$$

$$= \chi(w) \qquad (94)$$

namely, a function of w.

Now $\chi(0) = [0-E(\hat{D}_i)]^2 0 + \int_o^\infty [t-E(\hat{D}_i)]^2 dH(t) = Var(D_i)$ (95)

since, if w=0 then $E(D_i) = E(\hat{D}_i)$.

Therefore, all we need to show is that $\chi(w)$ is a decreasing function of w.

Then

$$\frac{\partial \chi(w)}{\partial w} = 2[w-E(\hat{D}_i)] \int_o^w dH(t) + [w-E(\hat{D}_i)]^2 H'(w)$$

$$-[w-E(\hat{D}_i)]^2 H'(w)$$

$$= 2[w-E(\hat{D}_i)] \int_o^w dH(t). \qquad (96)$$

Now $w < E(D_i)$, except for the trivial case where Prob $(\hat{D}_i=w)=1$, therefore

$$\frac{\partial(w)}{\partial w} < 0 \text{ for all } w. \tag{97}$$

This shows that as w increases, $Var(\hat{D}_i)$ decreases from $Var(D_i)$ to 0.

We have proved that buffering the packets of the receiver end reduces the variance by truncating the left-hand tail of the distribution.

From the above discussion one is lead to the following conglomerate protocol:

1. Buffer (as explained above) packets which arrive at the destination "too early" ($D_i \leq w$). This is equivalent to chopping the left-hand tail off delay distribution.

2. Discard packets where $D_i > x$ (the x may, in fact, be equal to w). This is equivalent to chopping the right-hand tail off the delay distribution (and hence the gap distribution also).

The implementation of this or similar protocols would imply software in the nodes, a nodal processing burden on packet exit from the network, additional overhead bits on transmission to contain time stamping information, and some kind of nodal synchronization to measure the absolute delay as opposed to relative delay. Synchronization, in particular, may be difficult to achieve in practice.

4.4 LINK MODEL

4.4.1 Introduction

So far we have discussed the input to the system and the requirements to be satisified in delivering the input traffic to its destinations. We now wish to describe the system to accomplish such transfer.

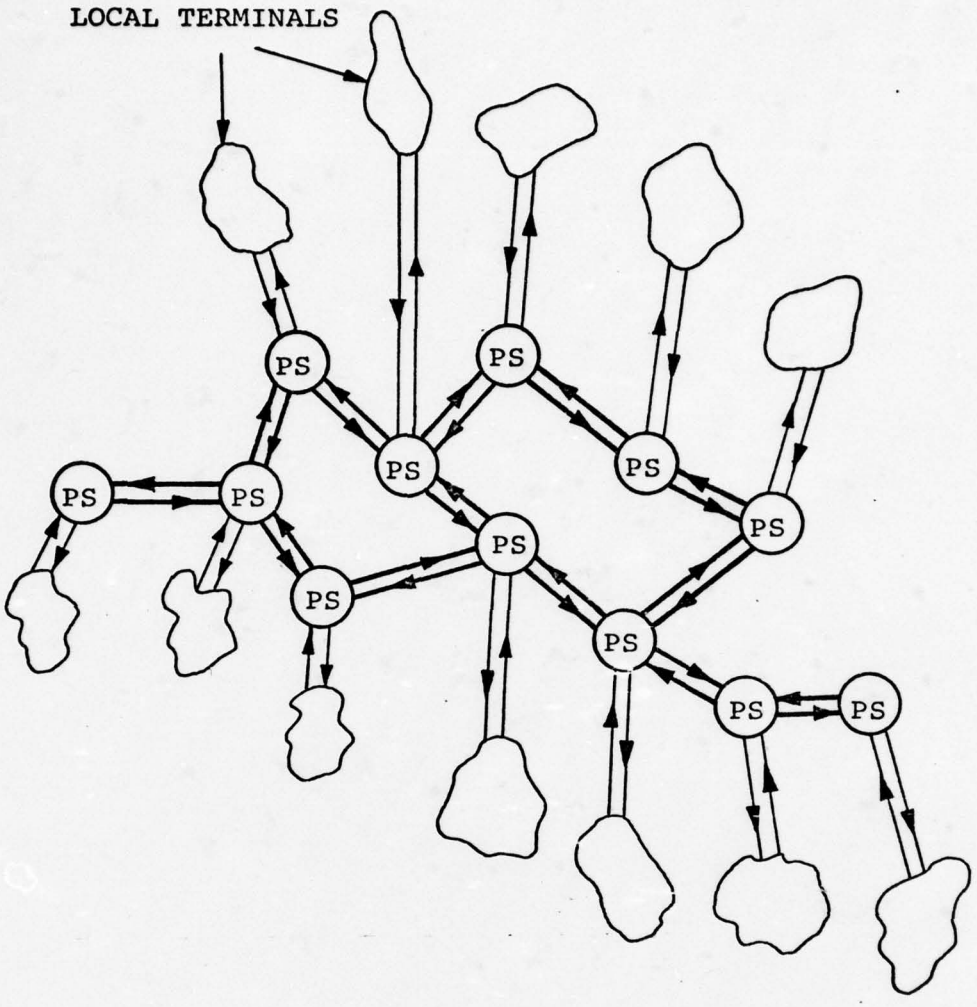
We consider a distributed population of digital voice terminals and a network of packet switches interconnected by a topology of (usually high capacity) links. Terminals are homed to a specific packet switch. Communication between remote terminals takes place via the appropriate backbone packet switches which are used for entry, exit and store-forward relay operations. Figure 21 shows a typical network. A number of qualitative and semi-quantitative studies on the general properties of a packet voice network already exist in the literature, e.g. [FORGIE, 1975], [COVIELLO, 1976].

The objective of our network model is to obtain end-to-end measures of performance, specifically: the distribution of the end-to-end delay; the percentage of lost packets; the amount of glitching; etc. Initially we shall look at a single link or channel and draw appropriate conclusions. Later the analysis will be extended to a network of tandem links and then to a more general network.

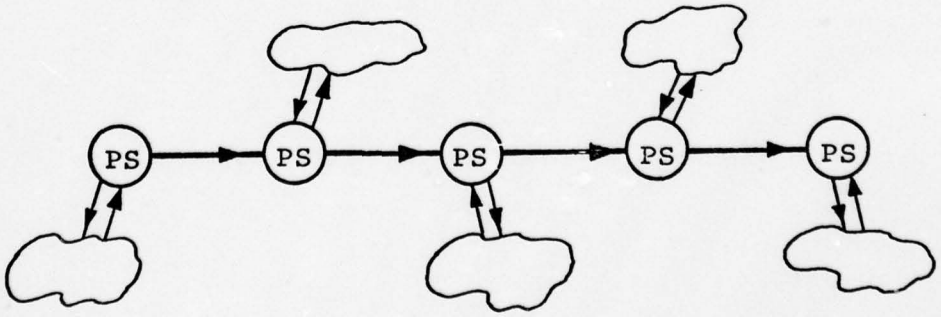
The present section looks at a single link model, representing a packet switch and one transmission channel. Section 7 models a multi-link situation.

4.4.2 Model description

We begin by identifying the underlying assumptions and introduce some descriptive notation.



A. GENERAL NETWORK



B. TANDEM NETWORK

FIGURE 21: NETWORK OF QUEUES

4.4.2.1 Traffic

Given any pair of speakers A-B, consider any one of the traffic models of Section 4.2, interfaced to the call origination model of Section 4.2.6. Let $P_i^{(n)}$ be the vector of unconditional probabilities of speaker i being in state 0,1,2,3,..., at frame n, e.g.,

$$P_i^{(n)} = (P_{i,0}^{(n)}, P_{i,1}^{(n)}, \ldots P_{i,s}^{(n)}) \tag{98}$$

where s corresponds to the number of states in the speech model and state 0 corresponds to the "on-hook" state. Let P be the transition matrix for i^{th} user which can be assumed to be identical for all users.
Let $P_i^{(0)}$, the initial state, be given, then

$$P_i^{(n)} = P_i^{(0)} \; P^n. \tag{99}$$

We are only concerned with some of the elements of this vector, namely $P_i^{(n)}$, the probability that speaker A supplies a packet at frame n, given that the pair A-B is off-hook. Then

$$P_i^{(n)} = \sum_{j|i \text{ talks}} P_{i,j}^{(n)} \Big/ \sum_{j=1}^{s} P_{i,j}^{(n)}. \tag{100}$$

$$\text{Let } Q_i^{(n)} = 1 - P_i^{(n)}. \tag{101}$$

For the two-state speech model

$$P_i^{(n)} = \frac{P_{i,1}^{(n)}}{P_{i,1}^{(n)} + P_{i,2}^{(n)}}\ ; \tag{102}$$

for the four-state model

$$P_i^{(n)} = \frac{P_{i,1}^{(n)} + P_{i,2}^{(n)}}{P_{i,1}^{(n)} + P_{i,2}^{(n)} + P_{i,3}^{(n)} + P_{i,4}^{(n)}} \cdot \tag{103}$$

Let

$$\tau_i^{(n)} = \begin{cases} 1 & \text{if terminal i is off-hook and has a packet at frame n} \\ 0 & \text{if terminal i is off-hook but has no packet at frame n.} \end{cases} \tag{104}$$

$$\left.\begin{aligned} \text{Thus } \text{Prob}(\tau_i^{(n)} = 1) &= P_i^{(n)} \\ \text{Prob}(\tau_i^{(n)} = 0) &= Q_i^{(n)}. \end{aligned}\right\} \tag{105}$$

If all users are assumed identically distributed we can drop the subscript, and to denote the steady state ($n \to \infty$) we drop the superscript. Thus $\text{Prob}(\tau=1) = P$.

Notice that the above procedure can be viewed as either considering the delay for the case when *every* subscriber assigned to the packet switch *is active and is engaged in an infinitely long* conversation ($\Omega=0$ in the call origination speaker model) - in this case m=M; or considering the delay for m users (not necessarily equal to M) given that they are off-hook and will remain so for an infinite (long) interval of time. See Section 4.2.6.

Let h be the packet time length (frame width), as before. It therefore also represents the packetizing delay; that is the delay introducing by the packetizing process itself.

4.4.2.2 Queue Operation

A. m terminals accessing the packet switch under consideration are off-hook and are engaged in (infinitely) long conversation.

B. Each terminal has two buffers for potential packets so it can start building the next packet after completion of one. Before the new one is completed the old one must be cleared or it will be overwritten. This implies a constraint on the processor.

C. We assume a scheduled arrival of potential packets; namely, terminal T_i generates completed empty or non-empty packets at times $rh+i\Delta$, $r = 0,1,2,\ldots$, where Δ is a parameter - possibly dynamically adjustable - of the entry switch. While such synchronization is impossible to arrange immediately following the off-hook condition, it can be accomplished in subsequent frames by appropriately clipping speech or adding silence. This needs to be done only once. This initial adjustment is not likely to be significant, especially when a terminal first goes off-hook. This implies, however, that the processor cyclically scans the input buffers of all off-hook terminals. An alternative would be to have the ready state of buffers generate queued interrupts for the processor's attention. The reason for not considering this alternative assumption is discussed later; it is noted however that the packetization itself generates a synchronized arrival stream, therefore the assumption is not unrealistic. Even under the interrupt alternative, the interrupts would be generated a frame length, h, apart during a talkspurt.

D. The appropriate buffer of T_i is processed at time $rh+i\Delta$. If occupied by a non-empty packet, the contents of the buffer are placed on a queue for transmission. The queue itself operates in a FIFO mode.

E. We assume that τ_i, the speaker behavior, is independent of the queue "backlog" as further explicitly discussed in Section 4.4.4.

It is clear that we must have $h \leq m\Delta$ to preserve the real time requirement - avoid extra delay or packet loss - at the entry node caused by the processor falling further and further behind in its cyclical scan for packets requiring transmission. Summarizing the above assumptions, we see that a packet joins the output transmission queue at time $rh+i\Delta$ with probability $p_i^{(r)}$. The time period, h, between the possible successive submissions of a packet by a particular terminal will still be referred to as a frame. (See Figure 22.)

There are a variety of other disciplines that can be imposed on the arrival sequence of packets from the off-hook terminals during a frame. If completely unsynchronized arrivals are permitted, then when two or more terminals start their talkspurt almost simultaneously, the latest arriving packet will be queued for transmission behind the others, suffering a delay in obtaining processor attention in the initial and *all subsequent frames* until the talkspurts ended. This delay could be avoided if we scheduled the arrival of the latest arriving packet to be later by at least the net processing time of the others. This can be accomplished by an initial adjustment (e.g., one-shot clipping), or at worst discarding the first packet - a minor effect on performance even at low terminal bit rates.

Note that, as already pointed out, packets from a specific terminal are self synchronized; namely, if the first packet is issued at some time δ then the second (if any) is issued at $\delta+h$, the third (if any) at $\delta+2h$, etc. Note this is a *consequence* (not an assumption) of the speech models of Section 4.2.

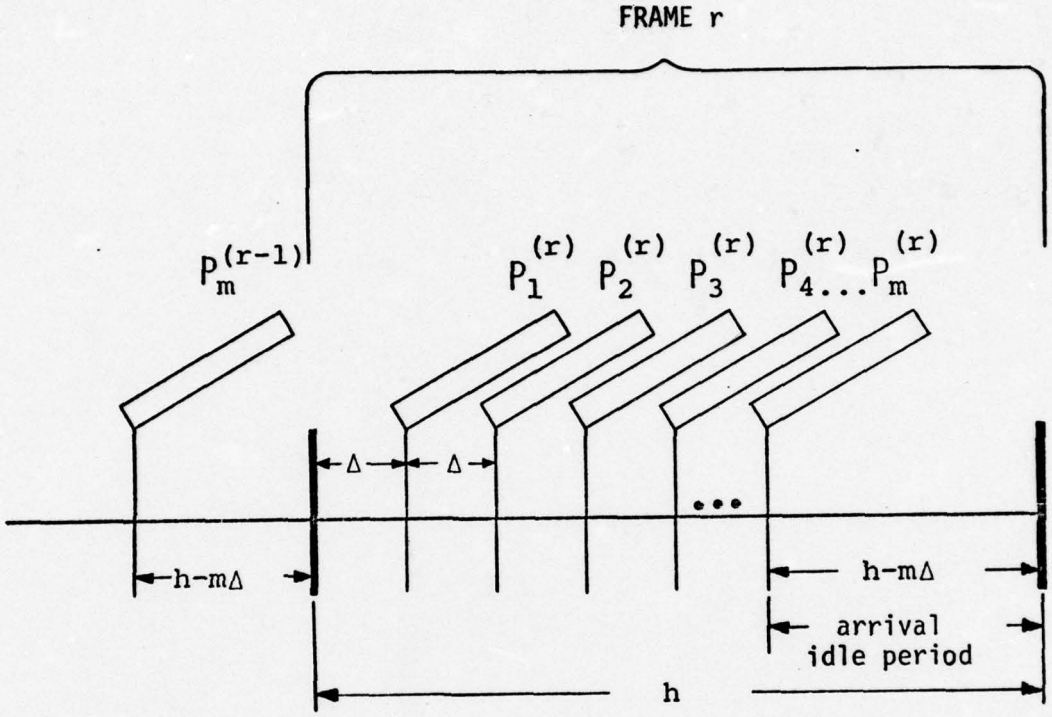


FIGURE 22: PROBABILISTIC QUEUE ARRIVAL SEQUENCE

Incorporating this idea, we have assumed an assigned fixed arrival time in a frame for the packets of each off-hook terminal - perhaps dynamically assigned as terminals change their off-hook status. Small glitches would be introduced in the dynamic case as the number of off-hook terminals fluctuated - the price for delay improvement. Under this condition, the most favorable and equitable (to all users) delay distributions will occur when the frame has scheduled packet arrivals at intervals of $\Delta = h/m$. Under this relationship the "arrival idle period" pictured in Figure 22 is eliminated and the set of random variable equations for delay will simplify.

4.4.2.3 Link Parameters

The following parameters influence the link performance:

a. m = number of off-hook users accessing the switch.

b. B = rate at which the digital voice terminal supplies bits (in bits per second).

c. C = capacity of the transmission link (in bits per second).

d. P = speech content per packet (in bits).

e. $h = P/B$ = time length of speech carried in a packet (in seconds).

f. $\emptyset$ = packet overhead (in bits) - header, etc.

g. s = the number of states in the speech model, as discussed in Section 4.2.

h. P = steady-state probability that a speaker will issue a packet in a frame. Determinable, in general, from s and the speech model state transition parameters.

i. K = number of buffers in the transmission queue.

For formulation simplicity we assume that s, B, P, $\emptyset$, and therefore h also, are identical for all terminals. From the above model parameters, it follows that the transmission service time for a packet is a *constant*, μ, computed as follows

$$\mu = \frac{\text{total packet length in bits}}{\text{line capacity}} = \frac{P+\emptyset}{C}. \tag{106}$$

An expression for line utilization, ρ, can be stated in terms of the above parameters, specifically,

$$\rho = \frac{P_m[B+(B/P)\emptyset]}{C} = \frac{P_{mB}}{C}\,[1+\emptyset/P]. \tag{107}$$

Relating the service time to the utilization we get

$$\rho = \frac{P_{mB}}{P}\mu = (\frac{P_m}{h})\mu \tag{108}$$

and for the case where $\Delta = h/m$ we get

$$\rho = P\,\frac{\mu}{\Delta}\,. \tag{109}$$

Note, in particular, that ρ, P, and μ/Δ are all dimensionless quantities - likely to be the critical system parameters. Another physical interpretation of the above relationship is that P/Δ is the average packet arrival rate and ρ/μ is the average packet departure rate, and the two must be equal in the steady state. We have the usual stability requirement, $\rho<1$, for the existence of the steady state.

4.4.3 Delay Analysis

We begin with the simplest possible model. Let $U_i^{(r)}$ be the output transmission delay (waiting plus service) that would be faced by a packet arriving on the queue at time $rh+i\Delta$. Namely, $U_i^{(r)}$ is the amount of time by which the system is behind just after terminal T_i would complete its generation of the r^{th} frame packet. If T_i has a non-empty packet ready at time $rh+i\Delta$, then $U_{i-1}^{(r)}-\Delta$ would be that portion of the queueing delay seen by T_{i-1} that remained at time $rh+i\Delta$. Thus $U_{i-1}^{(r)}-\Delta$ represents how much the sytem is "behind" at the instant that T_i's packet readiness state is examined. Thus we have the following random variable relationships:

$$U_{i\,i=1,\ldots,m}^{(r)} = \begin{cases} \mu & \text{if } \tau_i^{(r)}=1,\ U_{i-1}^{(r)}-\Delta \leq 0 \\ \mu + U_{i-1}^{(r)}-\Delta & \text{if } \tau_i^{(r)}=1,\ U_{i-1}^{(r)}-\Delta > 0 \\ 0 & \text{if } \tau_i^{(r)}=0,\ U_{i-1}^{(r)}-\Delta \leq 0 \\ U_{i-1}^{(r)}-\Delta & \text{if } \tau_i^{(r)}=0,\ U_{i-1}^{(r)}-\Delta > 0 \end{cases} \tag{110}$$

$$U_0^{(r)} = \begin{cases} 0 & \text{if } U_m^{(r-1)} \leq h-m\Delta \\ U_m^{(r-1)}-(h-m\Delta) & \text{if } U_m^{(r-1)} > h-m\Delta \end{cases} \tag{111}$$

where $U_0^{(r)}$ is the delay of a symbolic terminal placed in the arrival schedule at the beginning of the frame; it represents how much backlog transmission service from frame (r-1) is carried forward into frame r. $U_i^{(r)}$ is the delay backlog faced by terminal i in frame r.

The steady state delay is $U_i^{(\infty)}$. It has been determined by numerical computations that $U_i^{(r)}$ converges rapidly - particularly for low utilization - to $U_i^{(\infty)}$. In fact, r=10 generally yields a highly accurate result. The solution procedure involves obtaining the distribution for $U_i^{(r)}$, i=0,1,...m, r=0,1,...r_{max}, computationally, by performing the appropriate convolutions as required by the sequential relationships of the $U_i^{(r)}$.

Under the scheduled arrival assumption $\Delta=h/m$ of the last subsection, the set of equations simplify by having

$$U_0^{(r)} = U_m^{(r-1)}. \tag{112}$$

An intuitive interpretation is that the arrivals are scheduled so that the arrival idle periolds following each packet are identical for all users. With this assumption, all users have the same delay distribution in the steady state and we can denote the steady state random variable $U_i^{(\infty)}$ simply as U.

A simple augmentation of the above model can take into account finite buffer storage on the output transmission queue. If we have

$$\frac{U_{i-1}^{(r)}-\Delta}{\mu} \geq K \tag{113}$$

then the i^{th} terminal in the r^{th} frame is assumed to be blocked from placing its packet on the output queue and the blocked packet is assumed to be subsequently overwritten or otherwise discarded. The random variable relationships with $\Delta=h/m$ now become

$$U_{i\ i=1,\ldots,m}^{(r)} = \begin{cases} \mu & \text{if } \tau_i^{(r)}=1, \quad U_{i-1}^{(r)}-\Delta \leq 0 \\ \mu+U_{i-1}^{(r)}-\Delta & \text{if } \tau_i^{(r)}=1, \quad 0<(U_{i-1}^{(r)}-\Delta)<\mu K \\ 0 & \text{if } \tau_i^{(r)}=0, \quad U_{i-1}^{(r)}-\Delta \leq 0 \\ U_{i-1}^{(r)}-\Delta & \text{if } \tau_i^{(r)}=0, \quad U_{i-1}^{(r)}-\Delta>0 \\ & \quad\ \tau_i^{(r)}=1, \quad U_{i-1}^{(r)}-\Delta \geq K \end{cases} \tag{114}$$

$$U_0^{(r)} = U_m^{(r-1)} \tag{115}$$

We remind the reader that all m off-hook terminals are assumed to be conversing with terminals at a *different* packet switch, since otherwise their packets would not enter the transmission queue. Locally connected speakers would have no particular need of being "scheduled" within the arrival frame.

4.4.4 Solution Strategy

The solution strategy involves computationally convoluting the random variables $U_k^{(r)}$ where the relationships, more compactly expressed, have now simplified to

$$U_i^{(r)} = \begin{cases} \mu\tau_i^{(r)} + \max(0, U_{i-1}^{(r)}-\Delta), & i=1,\ldots,m \\ U_m^{(r-1)}, & i=0 \end{cases} \tag{116}$$

for the infinite buffer case. Many interesting system measures can be directly computed from the U distribution, including the important link delay distribution, as will be shown. The computational process can be started by assuming $U_0^{(1)}=0$ with probability 1, or any other convenient initial distribution.

We make an additional important simplifying assumption at this point to carry out the computations. We assume that $\tau_i^{(r)}$ and $U_{i-1}^{(r)}$ are independent. This will be a weak assumption if there is only one user, for example. However, if there are several users (m, in our notation), a given individual's influence on the total accumulated delay will not be significant, so this independence assumption will not distort the accuracy of the solution. This assumption permits the calculation of the joint probability of $\tau_i^{(r)}$ and $U_{i-1}^{(r)}$ as the product of their respective probabilities. The error - though slight - will tend to under estimate delay. High values of accumulated delay push the odds in favor of speech, since speakers tend to repeat their actions from frame-to-frame, and high delay implies a high proportion of recent speech activity.

In Table 1 we present the distribution of $U_0^{(15)}$ using the above methodology for the following conditions

m = 18 off-hook users
B = 5 KBS digitization
C = 50 KBS capacity
P = 1000 bits length
$\emptyset$ = 0 bits overhead
s = 2 states
K = ∞ buffers.

Note that s=2 implies that $P = \frac{1}{2}$. Also, h=200 ms. and the 18 users generate an average of 45 KBS with no overhead assumed, yielding a 90% utilization. The reader can verify that additional iteration(s) will have negligible effect on the $U_i^{(r)}$ distribution and steady-state is essentially achieved by the 15^{th} frame, even in this case of high utilization. Actually, the values of E(U) and V(U) can be accurately estimated much earlier. In this case E(U)=50 ms. (or 2.5 times the service time) and V(U)=2.0 ms.

POINT #	DELAY	PROBABILITY
0	0.000	7.74506265E-02
1	0.002	5.31273574E-03
2	0.004	7.57071149E-03
3	0.007	1.30809263E-02
4	0.009	4.33114250E-02
5	0.011	8.17372662E-03
6	0.013	1.06259583E-02
7	0.016	1.51414883E-02
8	0.018	2.61625277E-02
9	0.020	8.66224477E-02
10	0.022	1.63439163E-02
11	0.024	2.12481548E-02
12	0.027	3.02831468E-02
13	0.029	5.23238699E-02
14	0.031	1.83492445E-02
15	0.033	2.20637263E-02
16	0.036	2.73551405E-02
17	0.038	3.44061814E-02
18	0.040	1.80239817E-02
19	0.042	2.03449281E-02
20	0.044	2.28693611E-02
21	0.047	2.44283280E-02
22	0.049	1.64868853E-02
23	0.051	1.77122934E-02
24	0.053	1.86294871E-02
25	0.056	1.83843782E-02
26	0.058	1.44545402E-02
27	0.060	1.49474891E-02
28	0.062	1.50661186E-02
29	0.064	1.43762230E-02
30	0.067	1.23433224E-02
31	0.069	1.24213900E-02
32	0.071	1.21996927E-02
33	0.073	1.15070610E-02
34	0.076	1.03695657E-02
35	0.078	1.02370695E-02
36	0.080	9.89201292E-03
37	0.082	9.31805605E-03
38	0.084	8.62324494E-03
39	0.087	8.40000622E-03
40	0.089	8.05185956E-03
41	0.091	7.60236004E-03
42	0.093	7.13376771E-03
43	0.096	6.87949447E-03
44	0.098	6.56847045E-03
45	0.100	6.20745961E-03
46	0.102	5.87139418E-03
47	0.104	5.62969537E-03
48	0.107	5.36370574E-03
49	0.109	5.08363568E-03

TABLE 1: DELAY DISTRIBUTION

POINT #	DELAY	PROBABILITY
50	0.111	4.82983922E-03
51	0.113	4.61373036E-03
52	0.116	4.38416161E-03
53	0.118	4.16503276E-03
54	0.120	3.95468494E-03
55	0.122	3.77396698E-03
56	0.124	3.58405954E-03
57	0.127	3.40893294E-03
58	0.129	3.24377706E-03
59	0.131	3.09500529E-03
60	0.133	2.93863678E-03
61	0.136	2.79069998E-03
62	0.138	2.65970058E-03
63	0.140	2.52703944E-03
64	0.142	2.40325177E-03
65	0.144	2.28089423E-03
66	0.147	2.17525158E-03
67	0.149	2.07114851E-03
68	0.151	1.97223781E-03
69	0.153	1.87186002E-03
70	0.156	1.78099262E-03
71	0.158	1.69871739E-03
72	0.160	1.60897287E-03
73	0.162	1.53020986E-03
74	0.164	1.45212462E-03
75	0.167	1.38699274E-03
76	0.169	1.31960011E-03
77	0.171	1.25644531E-03
78	0.173	1.19203162E-03
79	0.176	1.13657450E-03
80	0.178	1.08439545E-03
81	0.180	1.02377238E-03
82	0.182	9.73468857E-04
83	0.184	9.22900523E-04
84	0.187	8.83403038E-04
85	0.189	8.40281617E-04
86	0.191	7.99548696E-04
87	0.193	7.57842281E-04
88	0.196	7.25499631E-04
89	0.198	6.92681861E-04
90	0.200	6.50447983E-04
91	0.202	6.18092076E-04
92	0.204	5.85322712E-04
93	0.207	5.62110661E-04
94	0.209	5.34366642E-04
95	0.211	5.07799108E-04
96	0.213	4.80783248E-04
97	0.216	4.63159926E-04
98	0.218	4.42607547E-04
99	0.220	4.12299101E-04

TABLE 1 (Cont'd)

POINT #	DELAY	PROBABILITY
100	0.222	3.91428839E-04
101	0.224	3.70187248E-04
102	0.227	3.57003184E-04
103	0.229	3.39101200E-04
104	0.231	3.21620140E-04
105	0.233	3.04105881E-04
106	0.236	2.95270391E-04
107	0.238	2.82693956E-04
108	0.240	2.60455301E-04
109	0.242	2.46937041E-04
110	0.244	2.33153814E-04
111	0.247	2.26001581E-04
112	0.249	2.14462585E-04
113	0.251	2.02868301E-04
114	0.253	1.91503126E-04
115	0.256	1.87443189E-04
116	0.258	1.80170588E-04
117	0.260	1.63668250E-04
118	0.262	1.54872727E-04
119	0.264	1.45937092E-04
120	0.267	1.42283816E-04
121	0.269	1.34892140E-04
122	0.271	1.27169196E-04
123	0.273	1.19800604E-04
124	0.276	1.17989153E-04
125	0.278	1.14227198E-04
126	0.280	1.02001598E-04
127	0.282	9.65043353E-05
128	0.284	9.04942190E-05
129	0.287	8.87717943E-05
130	0.289	8.41002166E-05
131	0.291	7.89677442E-05
132	0.293	7.42097946E-05
133	0.296	7.32584040E-05
134	0.298	7.17000230E-05
135	0.300	6.27632098E-05
136	0.302	5.90302548E-05
137	0.304	5.53361110E-05
138	0.307	5.46112128E-05
139	0.309	5.17230783E-05
140	0.311	4.83498807E-05
141	0.313	4.53066637E-05
142	0.316	4.46080530E-05
143	0.318	4.42751230E-05
144	0.320	3.78900822E-05
145	0.322	3.54865952E-05
146	0.324	3.31595152E-05
147	0.327	3.29060758E-05
148	0.329	3.11750646E-05
149	0.331	2.90055034E-05

TABLE 1 (Cont'd)

POINT #	DELAY	PROBABILITY
150	0.333	2.70924074E-05
151	0.336	2.64786902E-05
152	0.338	2.66849402E-05
153	0.340	2.22616093E-05
154	0.342	2.07477713E-05
155	0.344	1.93190017E-05
156	0.347	1.92615792E-05
157	0.349	1.82647029E-05
158	0.351	1.69145274E-05
159	0.353	1.57439947E-05
160	0.356	1.52239742E-05
161	0.358	1.55558175E-05
162	0.360	1.26078862E-05
163	0.362	1.16874924E-05
164	0.364	1.08420180E-05
165	0.367	1.08502018E-05
166	0.369	1.03032498E-05
167	0.371	9.49951550E-06
168	0.373	8.81005315E-06
169	0.376	8.41778797E-06
170	0.378	8.68413611E-06
171	0.380	6.81093070E-06
172	0.382	6.27817207E-06
173	0.384	5.80170615E-06
174	0.387	5.82218411E-06
175	0.389	5.53853107E-06
176	0.391	5.08594326E-06
177	0.393	4.69937805E-06
178	0.396	4.43927496E-06
179	0.398	4.59435944E-06
180	0.400	3.47123125E-06
181	0.402	3.18144808E-06
182	0.404	2.92874341E-06
183	0.407	2.94469641E-06
184	0.409	2.80664324E-06
185	0.411	2.56810426E-06
186	0.413	2.36403932E-06
187	0.416	2.21166926E-06
188	0.418	2.27872113E-06
189	0.420	1.65071688E-06
190	0.422	1.50432207E-06
191	0.424	1.37956594E-06
192	0.427	1.38872804E-06
193	0.429	1.32606104E-06
194	0.431	1.20957536E-06
195	0.433	1.10920277E-06
196	0.436	1.02959359E-06
197	0.438	1.04766005E-06
198	0.440	7.24337255E-07
199	0.442	6.56351212E-07

TABLE 1 (Cont'd)

POINT #	DELAY	PROBABILITY
200	0.444	5.99575564E-07
201	0.447	6.03902677E-07
202	0.449	5.77505297E-07
203	0.451	5.25334052E-07
204	0.453	4.79806566E-07
205	0.456	4.42149091E-07
206	0.458	4.40970236E-07
207	0.460	2.89861152E-07
208	0.462	2.61117357E-07
209	0.464	2.37498123E-07
210	0.467	2.39204006E-07
211	0.469	2.28931633E-07
212	0.471	2.07737902E-07
213	0.473	1.88909750E-07
214	0.476	1.72362681E-07
215	0.478	1.67380232E-07
216	0.480	1.03136055E-07
217	0.482	9.23417307E-08
218	0.484	8.35563352E-08
219	0.487	8.40975813E-08
220	0.489	8.04735327E-08
221	0.491	7.28600194E-08
222	0.493	6.59483304E-08
223	0.496	5.91381983E-08
224	0.498	5.56757587E-08
225	0.500	3.05224002E-08
226	0.502	2.60963564E-08
227	0.504	2.34827680E-08
228	0.507	2.36065092E-08
229	0.509	2.25668051E-08
230	0.511	2.03937198E-08
231	0.513	1.83792923E-08
232	0.516	1.60736557E-08
233	0.518	1.46881484E-08
234	0.520	7.10294118E-09
235	0.522	5.01355735E-09
236	0.524	3.71918343E-09
237	0.527	3.73423448E-09
238	0.529	3.56544549E-09
239	0.531	3.21797877E-09
240	0.533	2.89089450E-09
241	0.536	2.46477458E-09
242	0.538	2.20249161E-09
243	0.540	1.20941651E-09
244	0.542	8.61032835E-10
245	0.544	9.05798957E-10

EXPECTED DELAY = 0.050 VARIANCE = 0.002034

TABLE 1 (Cont'd)

Rewriting our previous relationship for the dimensionless system parameters as follows

$$\frac{\rho}{P} = \frac{\mu}{\Delta} \tag{117}$$

and defining $\tilde{\rho} = \rho/P$ and $\tilde{U}_i^{(r)} = U_i^{(r)}/\Delta$ we get for the simpler infinite buffer case

$$\tilde{U}_i^{(r)} = \begin{cases} \tilde{\rho}\tau_i^{(r)} + \max(0, \tilde{U}_{i-1}^{(r)} - 1) & i=1,\ldots, m \\ \tilde{U}_m^{(r-1)} & i=0. \end{cases} \tag{118}$$

Thus, all values of $\tilde{U}$ which can have non-zero probability are of the form

$$k\tilde{\rho} - n \geq 0, \tag{119}$$

where k and n are positive integers. Conceptually, $k\tilde{\rho}$ represents an accumulation of workload and n represents the time erosion of the workload by the server (transmission facility).

In Figure 23 we have drawn a general schematic representation of the successive iteration scheme for computing $\tilde{U}$. The schematic is strictly only valid for $\tilde{\rho} \geq 1$, but $\tilde{\rho} < 1$ is the case where $\mu < \Delta$ and is not of interest since a queue never forms. Initially all the probability is in state zero. Downward transition lines take P of the probability from the state and upward transition lines take (1-P), the remaining probability. Note that for any $\tilde{\rho} > 1$ - the range of validity - at least two new states are obtained with each iteration, corresponding to a generated packet at each opportunity including the current one, and each opportunity except one. Many equivalent states are generated during the iterations as indicated by the ovals in Figure 23. Depending on the value of $\tilde{\rho}$, many of the states - specifically those whose state value is not positive - are equivalent to the zero state. Each new state generated eventually

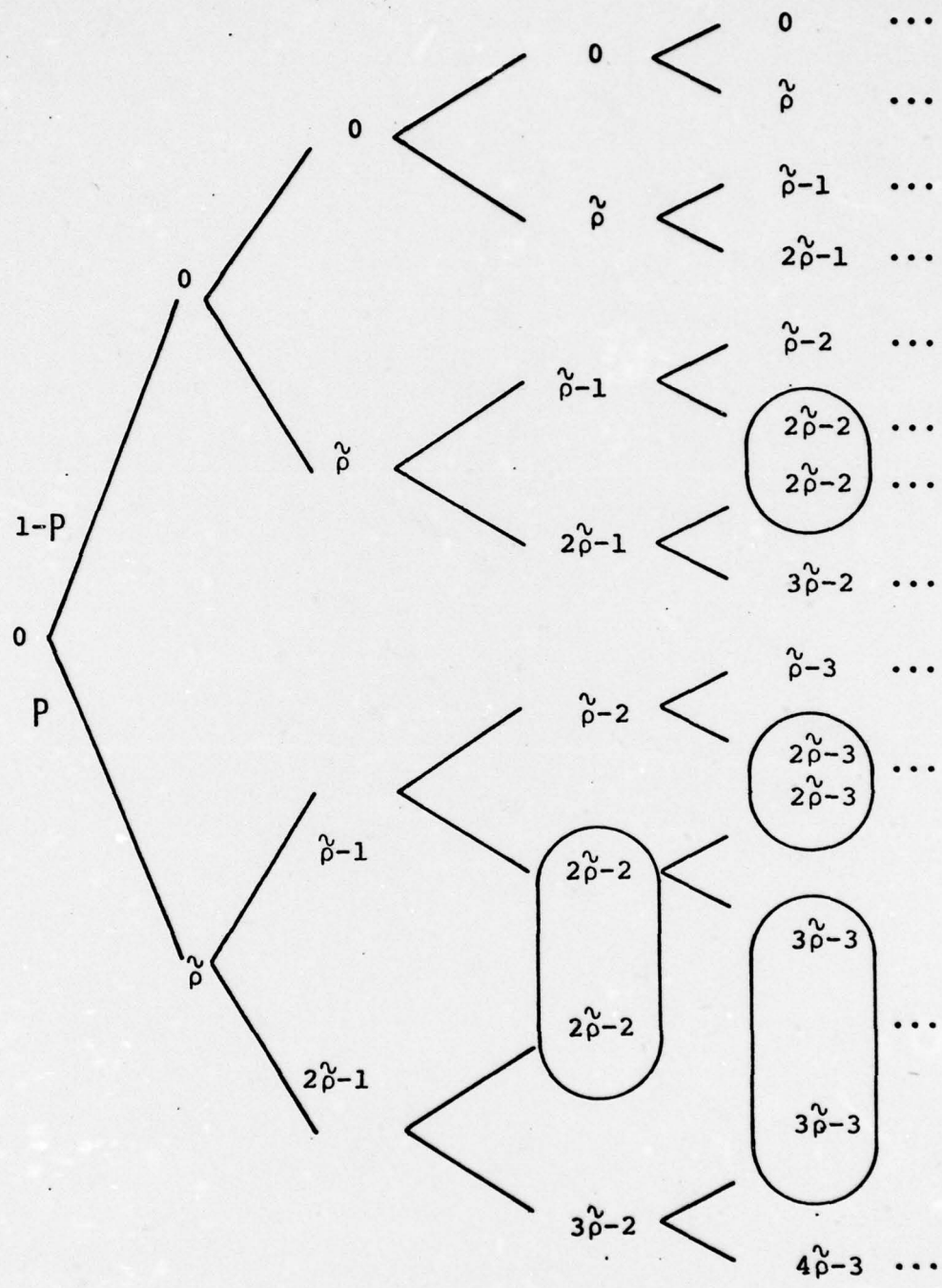


FIGURE 23: ITERATION SCHEMATIC

"transfers" its probability down to one or more zero equivalent states. The number of iterations it takes to do this depends only on the magnitude of $\tilde{\rho}$. Thus $\tilde{\rho}$ is a crucial convergence parameter. Also, the probability of the "bottom" states in the schematic, at each iteration, is a progressive power of P. Thus the significance of new states decreases geometrically with each iteration.

There appears to be little hope of obtaining a closed form expression for the U distribution. State probabilities are fed down to zero equivalent states and then eventually fed back up from the zero state indicating very involved equilibrium relationships.

Formulation of the state probability relationships from a difference equation point of view did not yield solutions. The difference equations are of high order, four different forms are involved, and there is a problem of equivalent states since $k\tilde{\rho}-n = 0$ has an infinite number of solutions.

4.4.5 Link Performance Variables

Several random variables of primary interest to the network designer can be derived directly and simply, once the distribution of $U_i^{(r)}$ is obtained. We define the following random variables:

$$d_i^{(r)} = U_i^{(r)} \mid (\tau_i^{(r)}=1) \tag{120}$$

$$D_i^{(r)} = U_i^{(r)}+h \mid (\tau_i^{(r)}=1). \tag{121}$$

$d_i^{(r)}$ is the delay - queueing plus service time - experienced by a packet and $D_i^{(r)}$ includes the packetizing delay. We require the random variable $D_i^{(r)}$ explicitly because the distribution of $U_i^{(r)}$ is an implicit function of h if there is a non-zero overhead. Thus $D_i^{(r)}$ is needed if system performance is to be studied as a function of h.

Measures of interest include the first and second moments of these distributions as well as tail probabilities such as

$$\text{Prob}(D_i^{(\infty)} > K_\alpha) \leq \alpha \tag{122}$$

with either α or K_α as the dependent variable. Another important measure is

$$\text{Prob}\left[U_{i-1}^{(r)} > K\mu+\Delta \,|\, (\tau_i=1)\right] \tag{123}$$

which is the probability of a buffer overflow when only K buffers are available.

Another important distribution is the time of departure from the queue of a generated packet. This gives us a view of the single link model from the output side. A description of the output process is particularly important because it gives the information needed to describe the input to nodes and links embedded in a network.

The delay experienced by a packet generated by user i in the r^{th} frame is

$$U_i^{(r)} \,|\, (\tau_i^{(r)}=1) = \begin{cases} \mu & \text{if } U_{i-1}^{(r)}-\Delta \leq 0 \\ \mu + U_{i-1}^{(r)} - \Delta & \text{if } U_{i-1}^{(r)}-\Delta > 0 \end{cases} \tag{124}$$

Therefore the departure time of this packet is

$$rh + i\Delta + U_i^{(r)} \,|\, (\tau_i^{(r)}=1) \tag{125}$$

Then the joint probability, $Q_{i,r}(t)$, of a departure at time t and terminal i having a packet in frame r is

$$Q_{i,r}(t) = \text{Prob}(rh+i\Delta+U_i^{(r)}=t \,|\, \tau_i^{(r)}=1)\ \text{Prob}(\tau_i^{(r)}=1). \tag{126}$$

The probability that a packet leaves the queue at t is

$$Q(t) = \sum_r \sum_i Q_{i,r}(t). \tag{127}$$

Finally, the probability that a packet departs in the interval $\gamma \leq t \leq \delta$ is

$$\sum_{\gamma \leq t \leq \delta} Q(t).$$

Several numerical examples of this and appropriate conclusions are described in Section 4.5.2.

The model presented in this section assumed that potential packet arrivals occur only at times $rh+i\Delta$, $r=1,2,\ldots$, $i=1,2,\ldots,m$. This assumes that the packet switch can control the synchronization of the packetization process - as opposed to the terminals - and that m does not fluctuate. While the second condition is satisfied in the present formulation, the question arises as to how to deal with the case when m is allowed to vary, as for the model of Section 4.2.6.

If M represents the total terminals homed to packet switch - in other words M is the maximum value of m - the frame interval can be permanently divided into h/M subintervals. Terminals can be preassigned to a specific arrival time within the frame or given one as they go off-hook. Thus, with h/M arrival times available, no dynamic adjustment would be required as m fluctuated.

In this case, however, we need to redefine the arrival probabilities as

$$R_i^{(n)} = P_i^{(n)} \sum_{j=1}^{s} P_{i,j} \tag{128}$$

$$S_i^{(n)} = 1-R_i^{(n)} \tag{129}$$

$$\Delta = \frac{h}{M} \tag{130}$$

so that $R_i^{(n)}$ represents the probability of a packet arrival in the r^{th} by the i^{th} terminal but includes the possibility of the terminal being on-hook. With this minor modification one can study the dynamic behavior of the single link transmission queue as calls arrive and depart from the switch. Naturally, if Δ were computed as

$$\Delta^{(n)} = \frac{h}{m^{(n)}} , \qquad (131)$$

a slightly improved performance would be achieved over the performance obtainable with the constant Δ. In view of the queuing model complexity introduced by this variable Δ, this latter approach was not followed.

No numerical computation has been carried out when $\Omega \neq 0$ in the call origination model (i.e., $m^{(n)}$ is a function of n, but Δ is considered to be h/M); however, we can make a worst case assumption, m=M, and be confident that a resulting network design is conservative and can handle peak requirements.

4.4.6 More General Arrival Scheme

Consider a fixed population of m off-hook pairs of speakers. The model employed thus far assumed that packet arrivals occur at uniformly spaced times

$$rh+i\Delta \qquad \begin{array}{l} r=1,2,... \\ i=1,2,...,m \\ \Delta=h/m \end{array}$$

or more generally allowing arrivals at $rh+i\Delta_i$ with

$$\sum_{i=1}^{m} \Delta_i = h. \qquad (132)$$

In this subsection we allow more freedom in the arrival scheme, by modeling a random (uniform) distribution of the packet departure from the digital voice terminal in the first frame. Packet departures - if any - at subsequent frames are simply offset by rh, r=1,2,..., from the first departure.

To approximately achieve this, we can postulate that for the initial frame, m arrivals are allowed to occur at any one of m

points out of a totality of M discrete points, $M \geq m$; where M is suitably large. We now have the possibility, for example, that all m packets can come in toward the beginning (or the end) of the frame.

Let

$$\chi = \{\ell\Delta \mid \ell=1,2,\ldots, M \text{ and } \Delta = h/M\} \tag{133}$$

be the set of available packet arrival points within a frame. With m off-hook users we can assume that each of the points in χ is equally likely to be an actual potential packet arrival point. Let

$$\chi_m = \{\text{all subsets of } \chi, \text{ having cardinality } m\}$$

Each element $a \in \chi_m$ has probability

$$\frac{1}{\binom{M}{m}},$$

the reciprocal of the number of equally likely elements in χ_m.

Using a slight modification of the technique of Sections 4.4.3 and 4.4.4 we could calculate the distribution of ${}_aV_k^{(r)}$, the delay for the k^{th} user in frame r assuming the potential packet arrival pattern given by $a \in \chi_m$. We could perform the calculation for every $a \in \chi_m$ and then combine the distributions under the equal likelihood assumption to obtain $V_i^{(r)}$.

The procedure is straightforward but computationally complex. No numerical evaluation has been attempted but we would expect the delay to increase slightly from our previous simpler model.

As a final observation, random (uniform) potential packet arrival times within a frame result in an interarrival length distribution given by

$$\text{Prob}(\Delta_k > x) = (1-\frac{x}{h})^m \qquad (134)$$

where Δ_k is the time between the k^{th} and $(k-1)^{st}$ potential packet arrivals. Recall that these times would remain identical frame-to-frame as long as the set of m off-hook terminals remained stable.

Of mathematical interest only, since it is physically unrealistic, is the fact that as h becomes infinite, the interarrival times approach an exponential distribution for which well-known queuing theory results are available. (See Section 4.6).

4.5 SINGLE LINK BEHAVIOR

The model of the previous section can be used to study the performance of a single link network carrying only packetized speech traffic. In this section we study the following for the steady state situation: the delay distribution faced by a typical packet - both for the case of infinite buffer capacity and the case of finite buffer; the output distribution of the link; the effect of the speech model on the delay and on the utilization of the link; the relationships between the overhead, the digitization rate and the link capacity on the mean and variance of the delay (actually the entire distribution is obtained in every case); finally, we determine optimal packet lengths - with respect to a specific performance criterion - as a function of the number of users, the overhead, the digitization rate and the capacity of the transmission line. We also use the model to address the complex transient issues.

4.5.1 Properties of the Delay Distribution

As indicated in Section 4.3, we are, in general, interested in studying the characteristics of the delay distribution. A concentration of probability around the delay mean necessarily implies a related concentration of probability near zero for the gap structure.

A "concentration-of-probability" study can take two approaches: if only the mean and variance of the distribution are available, one can resort to a (generalized) Chebychev technique; otherwise, when the distribution itself is available, one can carry out the analysis directly on the distribution. With a computer program available implementing the model of Section 4.4 we are, fortunatelv, in the latter situation. The advantage of having the entire distribution is that precise statements about the tail structure can be

made (e.g., the number of standard deviations from the mean one must accept from the delay distribution before insuring that there is a sufficiently small residual probability in the tail.) The Chebychev inequality implies, for example, that one must go out 4.47 standard deviations from the mean before covering 95% of the area, and 10 standard deviations to cover 99% of the area. While these estimates are very conservative, in general, there are distributions which are widely dispersed and actually have such long tails. For comparison, for an exponential (standard normal) random variable, the 95^{th} percentile is at 2(2) standard deviations, and the 99^{th} at 3.6(2.58) standard deviations.

A distribution with a high probability tail is poor, in the sense that one must pay a high price in terms of network control and slack facility utilization to ensure a certain performance confidence. Packet speech is particularly vulnerable to high probability tails in the delay distribution because the value of its information is perishible. If a network is not designed properly - say much hopping, and a very high utilization of line capacity - the variance may be such that one would be forced to wait several seconds before insuring that the fraction of the packets needed for fidelity are received. This would violate the tight delivery time requirements previously described.

Because of the correspondence between the tail probability and the variance - as made precise by the Chebychev inequality - in the sequel we sometimes use the two concepts interchangeably.

As a first order approximation, the end-to-end variance can be taken as the sum of the links variances (and is bounded by $L^2\sigma_M^2$ where L is the number of links and σ_M^2 is the largest link variance). Therefore, the variance of the end-to-end delay can be directly related to the variance of the delay on a single link. A key issue is: "What is the single link delay distribution for a packet voice network?" To answer this question we explored a wide range of cases. In each instance we examined the total delay - including packetizing delay which is deterministic - as discussed in Section 4.3.1.1.

Generally speaking, the delay distribution is a function of five parameters driving our model; namely, the line capacity C, the packet length P, the packet overhead $\emptyset$, the digitization bit rate B, the number of users and/or the speech model. Hopefully, the distribution will be a function of a reduced set of (dimensionless) parameters such as ρ (which is invariant, for example, under a doubling of m and C), μ (which is invariant under a doubling of P, $\emptyset$ and C) and h (which is invariant under a doubling of P and B). Possible relationships will be examined and conclusions made. The inclusion of the packetizing delay in the total delay complicates the relationship between the delay and the driving variables; because of this complication we do not attempt in this section to vary all parameters influencing the distribution. Our methodology examines the trends when the utilization is increased by increasing just the number of users - holding everything else constant. The resulting increase in the tail probabilities is typical of what occurs when the utilization increases independently of the cause.

Figure 24 shows the distribution of total delay for five values of ρ for the case $P=800$, $\emptyset=200$, C=50 KBS, B=5 KBS, and a symmetric two-state speaker model. Naturally, as the utilization increases the distribution becomes more dispersed. For comparison Table 2 and Figure 25 show the expected value and the 95^{th} percentile with the mean as predicted from an M|M|1 approximation* as a comparison reference. For the more structured input process of speech and constant packet length, there is a slower degradation in expected delay (and percentiles) as $\rho \to 1$ than in the case of data (assumed M|M|1). In terms of concentration of probability, the distributions under consideration are comparable to exponential distributions. Figure 26 compares the obtained distribution to an exponential distribution

*In all curve fittings and approximations to follow, we fit or approximate to $d^{(\infty)}-\mu$ the waiting delay (queueing time without service time) but plot $D^{(\infty)} = d^{(\infty)} + h$, the total delay, which is the variable of primary interest. (See Section 4.4.4.5.)

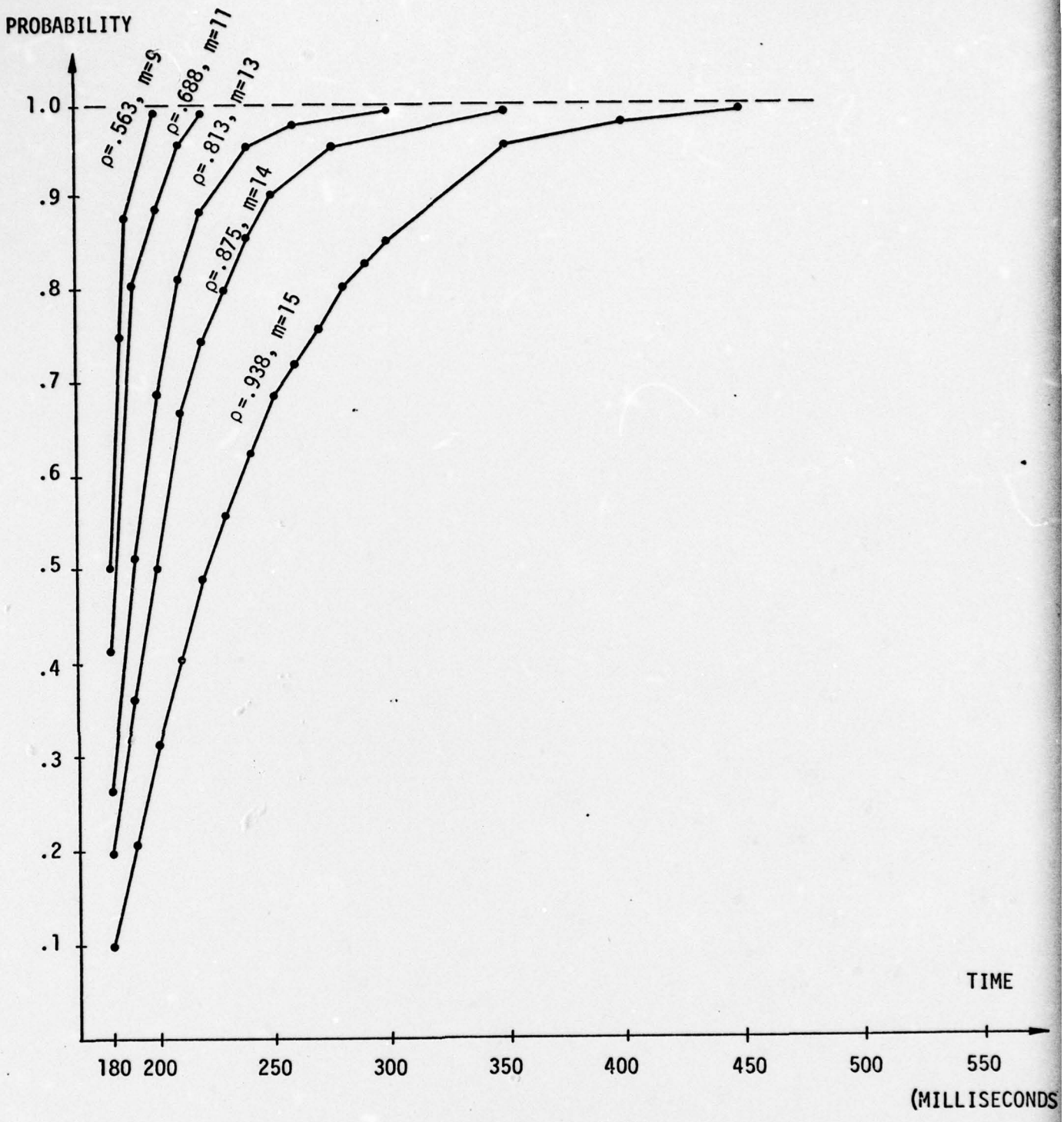


FIGURE 24: TYPICAL TOTAL DELAY DISTRIBUTIONS

m	ρ	E(D)	E(D) (M|M|1 Appx)	95th
9	.563	.183	.206	.188
10	.625	.184	.213	.196
11	.688	.187	.224	.207
12	.750	.191	.240	.220
13	.813	.196	.261	.238
14	.875	.209	.320	.271
15	.938	.239	.480	.357

TABLE 2: MEAN AND 95th PERCENTILE AS A FUNCTION OF ρ

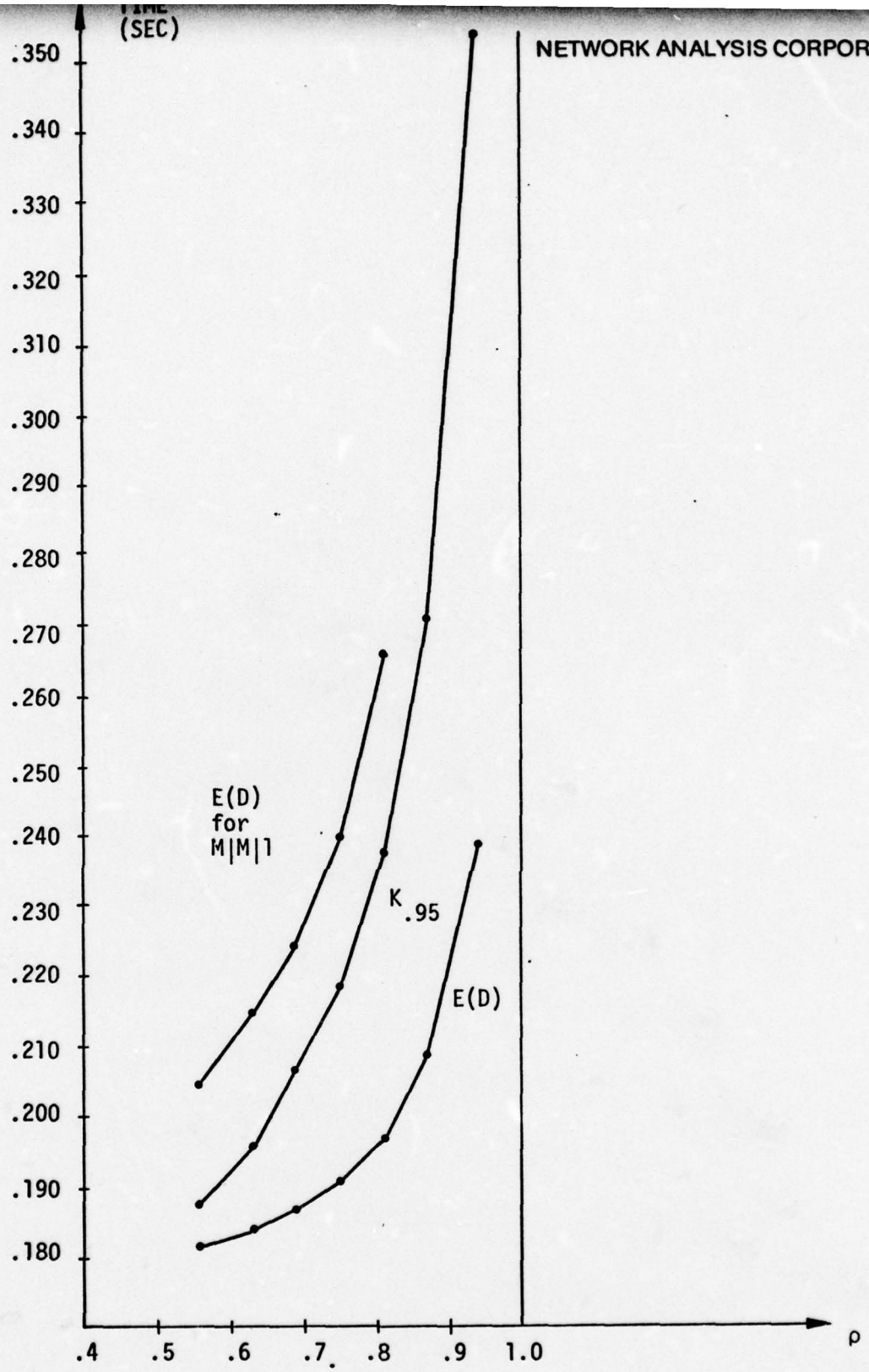


FIGURE 25: EXPECTED DELAY AS A FUNCTION OF ρ

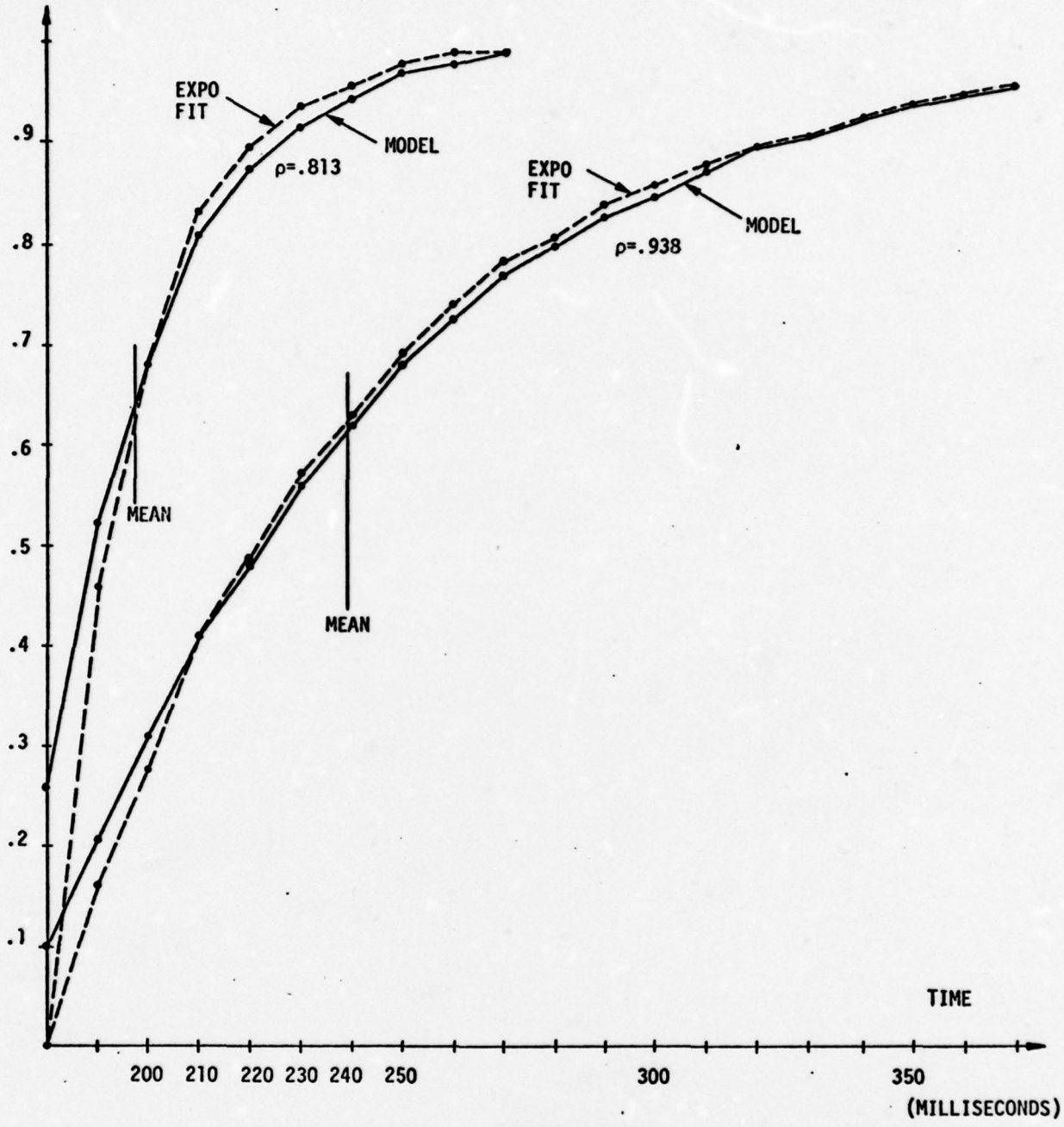


FIGURE 26: COMPARISON BETWEEN MODEL DELAY DISTRIBUTION AND EXPONENTIAL APPROXIMATION

for the cases of ρ=.813 and ρ=.938. The agreement is excellent. From this observation it appears that we can approximately chacterize the single link delay distribution with one parameter - its expected value. Later sections are devoted to characterizing the mean - since a close approximation to the entire distribution is then available.

For the infinite buffer case the following conclusions can be made:

1. The exponential distribution accurately fits the actual delay distribution.

2. As a consequence of good exponential fit, the 95^{th}, 99^{th}, and 99.9^{th} percentile of the model's total delay distribution, are comparable to those of the exponential distribution (in the number of standard deviations); the model delay tends to have a slightly longer tail. See the table below. This can be observed in Figure 26.

PERCENTILE	STANDARD DEVIATIONS FROM MEAN FOR EXPONENTIAL	STANDARD DEVIATIONS FROM MEAN FOR MODEL DELAY
95^{th}	2	Circa 2
99^{th}	3.6	Circa 3.6
99.9^{th}	5.9	Circa 6.2

3. The standard deviation, σ_D, of the model delay is slightly higher than the standard deviation, σ_E, of an exponential random variable having the same mean.

4. The model delay has an accumulation of probability at an initial point, indicating the possibility of finding a server idle; this feature is not captured by the exponential random variable. Observe however (Figure 24) that as the utilization increases this probability decreases; since most of the times we operate at high utilization, this discrepancy is expected to have minor implications.

On the other hand, the finite buffer case reduces the expected delay and contracts the delay distribution as compared to the infinite buffer case, at the expense of blocking some packets. In Section 4.5.5 we address this issue in more detail.

4.5.2 Output Distribution

There are various ways of describing the output process of a queue. We have selected an approximate description that is especially suitable for later work on tandem and general link models.

A fixed line capacity, C, and fixed packet length, $P+\emptyset$, implies a deterministic service time, μ. We assume for simplicity that h/μ is an integer. We see that if the output of the line is fed to a single receiving buffer, this buffer needs to be serviced every μ seconds to avoid overflow. More precisely, if a packet emerges from the line at t_0, there is no need to revisit the buffer before $t_0+\mu$ because it is impossible for a new packet to emerge at the line output before μ seconds are up. On the other hand, if one waits more than μ seconds there is danger of overwriting. The only problem is when to initiate the first scan. We fix the first scan at t=0.

Let T_i be the emergence time of packet i; let $H_i(t)$ be its distribution. Then

$$\int_{(j-1)}^{j\mu} dH_i(t) \qquad j = 1,2,3... \tag{135}$$

is the probability that packet i emerges between $(j-1)\mu$ and $j\mu$; then

$$p_j = \sum_{i=1}^{\infty} \int_{(j-1)\mu}^{j\mu} dH_i(t) \tag{136}$$

is the probability that a packet emerges between $(j-1)\mu$ and $j\mu$. The p_j's are a description of the output process; we do not need to determine $H_i(t)$ explicitly; all we need to construct a tandem queue model is the probability that at scan j a packet is found in the incoming line buffer (see Section 4.7) - this is p_j.

As one would expect, the output departure times are correlated; if a packet was just received there is some information on the arrival of the next packet. A number of studies reported on the literature for the case of M|M|1 and/or M|D|1 queues have shown that such correlation is extremely small and can thus be ignored. See [KING, 71], [DALEY, 68], [FINCH, 59], [JENKINS, 66]. We expect a similar behavior for our model.

It turns out that the behavior of the p_j is cyclic. Consider integers K_μ and K_Δ such that

$$K_\mu \mu = K_\Delta \Delta \tag{137}$$

Then $p_{j+K_\mu} = p_j$ for all j, once we reach the steady state. Thus the output process as we have described it has periodicity K_μ. Such periodicity is induced by beats of two process, arrival and service, being performed at different rates. The following two observations hold:

1. $\sum_{j=1}^{K_\mu} p_j = K_\mu \rho$; or, in other words, on the average p_j is equal to ρ.

2. if $0<\Delta<<\frac{h}{m}$ then the interdeparture times are *not* cyclic and $p_j = \rho$ for all j.

Table 3 shows an example of the output process for a particular case, after the steady state has been reached. $K_\mu=3$, and $\rho = .667$ for this example. Observe the departure cyclical behavior and the fact that the sum of the p_j's is 3ρ. Table 4 shows a situation with a high K_μ.

The finite queue buffer case exhibits the same characteristics relating to periodicity and correlation except that

$$\sum_{j=1}^{K_\mu} p_j = K_\mu \rho \chi \qquad (138)$$

where χ = Prob(Buffer not full).

Naturally, as expected, the p_j's are less for the finite buffer case. However, it is not true that each p_j is precisely reduced by the χ factor. For comparison we show in Table 5 the cases for buffer sizes $K=\infty,3,2$.

The above characterization of the output distribution is quite useful since

1. It allows us to compute the arrival distribution at the next queue, in a simple way.

2. It contains some information on correlation.

3. An approximation for the output process can be obtained (other information missing) by taking $p_j=\rho$ for all j.

j	PROBABILITY A PACKET LEAVES BETWEEN jμ AND (j+1)μ	
20	.50000	
21	.85186	
22	.64767	PERIOD
23	.50000	
24	.85203	
25	.64774	
26	.50000	
27	.85212	
28	.64777	
29	.50000	
30	.85221	
31	.64781	
32	.50002	

m=2
P=100
B=5000
C=15000
Ø=100

TABLE 3: EXAMPLE OF OUTPUT PROCESS

j	PROBABILITY A PACKET LEAVES BETWEEN jμ AND (j+1)μ
37	.68547
38	.80090
39	.69265
40	.80716
41	.68523
42	.81078
43	.68905
44	.80597
45	.69157
46	.80826
47	.68864
48	.80971
49	.69017
50	.80772
51	.69122
52	.80868
53	.68994
54	.80932
55	.69062
56	.56949
57	.83444
58	.66552
59	.77862
60	.69718
61	.80279
62	.67349
63	.81436
64	.68563
65	.80085
66	.69284

PERIOD

m=4
P=900
B=5000
C=15000
Ø=100

TABLE 4: LONG PERIOD OUTPUT PROCESS

INFINITE BUFFER		3 BUFFERS		2 BUFFERS	
j	PROBABILITY A PACKET LEAVES BETWEEN jμ AND (j+1)μ	j	PROBABILITY A PACKET LEAVES BETWEEN jμ AND (j+1)μ	j	PROBABILITY A PACKET LEAVES BETWEEN jμ AND (j+1)μ
56	.82691	71	.81025	27	.81027
57	.85776	72	.84562	28	.67523
58	.75038 PERIOD	73	.72878 PERIOD	29	.78376 PERIOD
59	.72740	74	.82255	30	.63546
60	.72740	75	.70387	31	.77382
61	.82700	76	.81205	32	.81027
62	.85784	77	.84562	33	.67523
63	.75048	78	.72878	34	.78376
PROB BUFFER NOT FULL = 1.0		PROB BUFFER NOT FULL = .9782		PROB BUFFER NOT FULL = .9196	

4.110

m=7
P=900
B=2000
C=10000
Ø=100

TABLE 5: EFFECT OF FINITE BUFFER SIZE ON OUTPUT PROCESS

4.5.3 Delay Mean and Variance

It has been indicated in Section 4.5.1 that an exponential distribution with the same mean as that of our single link model is a good match. It thus becomes important to characterize the mean of our model. The variance can be considered to be equal to the square of the mean, as the exponential predicts. We show below that the expected delay can be expressed as an explicit function of three variables: μ, h and ρ. Table 6 shows no change in the mean as m, P, $\emptyset$, B, and C are varied with μ, h and ρ being held constant. Note that for the two-state speaker model - used throughout this subsection -

$$m = \frac{2\rho h}{\mu} \tag{139}$$

so m is implicitly also held constant. Furthermore

$$C = \frac{hB+\emptyset}{\mu} \tag{140}$$

$$P = hB. \tag{141}$$

Thus given ρ, h, μ it follows that C and P are determined by fixing $\emptyset$ and B and vice versa.

We can verify by numerical computation that

$$E(D) = f_1(h,\mu,\rho) \tag{142}$$

for some function f_1.

To determine the functional form a three dimensional parameter variation was performed - similar to a Karnaugh map analysis of logical variable states. Table 7 shows the results: across each row, ρ is constant; down each column, μ is constant; h is one constant on the left half, and another constant on the right half.

ø = 200 B = 5000 m = 15 C = 72,500 P = 1250	E(D) = .273
ø = 100 B = 5000 m = 15 C = 67,500 P = 1250	E(D) = .273
ø = 100 B = 2500 m = 15 C = 36,250 P = 625	E(D) = .273
ø = 100 B = 10000 m = 15 C = 130,000 P = 2500	E(D) = .273

μ = .020
ρ = .6
h = .250

TABLE 6: CONSTANT μ, ρ, h.

	h = .250			h = .500		
	μ = .100	μ = .020	μ = .010	μ = .100	μ = .020	μ = .010
ρ = .6	m = 3 Ø = 100 B = 5000 C = 13,500 P = 1250 E(D) = .367	m = 15 Ø = 100 B = 5000 C = 67,500 P = 1250 E(D) = .273	m = 30 Ø = 100 B = 5000 C = 135,000 P = 1250 E(D) = .262	m = 6 Ø = 200 B = 5000 C = 27,000 P = 2500 E(D) = .617	m = 30 Ø = 300 B = 4000 C = 115,000 P = 2000 E(D) = .523	m = 60 Ø = 200 B = 5000 C = 270,000 P = 2500 E(D) = .512
ρ = .8	m = 4 Ø = 200 B = 1000 C = 4500 P = 250 E(D) = .428	m = 20 Ø = 300 B = 5000 C = 77,500 P = 1250 E(D) = .286	m = 40 Ø = 100 B = 1000 C = 35,000 P = 250 E(D) = .268	m = 8 Ø = 400 B = 10,000 C = 54,000 P = 5000 E(D) = .678	m = 40 Ø = 100 B = 2500 C = 67,500 P = 1250 E(D) = .536	m = 80 Ø = 200 B = 5000 C = 270,000 P = 2500 E(D) = .518

TABLE 7: PARAMETER VARIATION FOR FUNCTIONAL DEPENDENCIES

NETWORK ANALYSIS CORPORATION

If we are interested in just the queueing delay, we can express E(D) as follows:

$$E(D) = f_2(\mu,\rho) + h \qquad (143)$$

where f_2 is the expected value of the link delay, E(d) and h, the packetizing delay, is put in a separate term. We then express f_2 as

$$f_2(\mu,\rho) = \mu f_3(\rho). \qquad (144)$$

Table 8 shows the experimentally computed values for f_2 and f_3 and supports our conjecture for the functional form of the dependencies of E(D) on μ and ρ. This form of dependence is commonly encountered in a variety of similar queueing situations.

The remaining task to obtain the functional form for $f_3(\rho)$. A more detailed series of experiments was employed with the results summarized in Table 9. Various functional forms were tested. The results are contained in Table 10. Values of the coefficients were determined which give the best fit. All the functional forms shown except form 5 had an asymptotic growth of too high an order. In view of its excellent fit to actual model measurements in the high (>.5) utilization range, for a wide range of system parameters, as shown in Table 11, form 5 can be used as the empirical expression for the mean of the exact model. This will prove to be a valuable asset in doing network design.

We thus have the following empirical expression for total delay:

$$E(D) = \frac{.358 - .266\rho^2}{1-\sqrt{\rho}}\ \mu + h \quad , \quad \rho > .5 \qquad (145)$$

Note that the rise to infinity as $\rho \to 1$ is slower than that for M|M|1 queue but the ratio of the two delays approaches $2(I-J)=.184$ as $\rho \to 1$. (See Table 12.) The regularity of the traffic, intrinsic to a voice environment permits a higher line utilization for the same average delay than in a bursty data environment.

	$\mu = .1$	$\mu = .02$	$\mu = .01$
$\rho = .6$	.117 $= .1(.117)$	.023 $.02(1.17)$	.012 $.01(1.17)$
$\rho = .8$	.178 $= .1(1.78)$	.036 $.02(1.78)$	.018 $.01(1.78)$

$$f_2(\mu,\rho) = \mu f_3(\rho)$$

TABLE 8: EXPERIMENTAL VALUES FOR f_2 AND f_3

	ρ	E(D)	$f_3(\rho)$
1	.563	.182	1.10
2	.625	.184	1.20
3	.688	.187	1.35
4	.750	.191	1.55
5	.813	.197	1.85
6	.875	.209	2.45
7	.938	.239	3.95

$h = .160$

$\mu = .020$

$f_3(\rho) = \frac{E(D)-h}{\mu}$

TABLE 9: EXPERIMENTAL VALUES FOR f_3

		FORM 1	FORM 2	FORM 3	FORM 4	FORM 5
		$\frac{A-B\rho}{1-\rho}$	$\frac{C-D\sqrt{\rho}}{1-\rho}$	$\frac{E-F\sqrt{\rho}}{1-\sqrt{\rho}}$	$\frac{G-H\rho}{1-\sqrt{\rho}}$	$\frac{I-J\rho^2}{1-\sqrt{\rho}}$
ρ	$f_3(\rho)$	A=.870 B=.639	C=1.367 D=1.127	E=.821 F=.706	G=.511 H=.403	I=.358 J=.266
.563	1.10	1.17	1.19	1.16	1.13	1.10
.625	1.20	1.26	1.27	1.25	1.24	1.21
.688	1.35	1.38	1.39	1.38	1.37	1.36
.750	1.55	1.56	1.56	1.56	1.56	1.55
.813	1.85	1.87	1.87	1.87	1.87	1.85
.875	2.45	2.49	2.50	2.48	2.45	2.40
.930	3.95	4.36	4.44	4.34	4.22	3.95

TABLE 10: FIT OF FUNCTION FORMS OF f_3

NETWORK ANALYSIS CORPORATION

$$E(D) = \frac{.358 - .266\rho^2}{1-\sqrt{\rho}}\,\mu + h$$

	EXACT	PREDICTED
B=2500, C=50000, Ø=100, m=18, P=100	.051	.051
B=2500, C=50000, Ø=100, m=18, P=112	.054	.054
B=2500, C=50000, Ø=100, m=18, P=125	.058	.058
B=10000, C=200000, Ø=100, m=18, P=150	.017	.017
B=10000, C=200000, Ø=100, m=18, P=125	.015	.015
B=10000, C=200000, Ø=100, m=18, P=100	.013	.013

TABLE 11: COMPARISON OF E(D) WITH VALUE PREDICTED USING FORM 5

	E(D) (for μ=1, h=0)	
	M|M|1 (Data Traffic)	Empirical Function & Model
ρ = .9	10	3
ρ = .99	100	20
ρ = .999	1,000	185
ρ = .9999	10,000	1,841
ρ = .99999	100,000	18,401
ρ = .999999	1,000,000	184,001

TABLE 12: HIGH UTILIZATION BEHAVIOR OF EMPIRICAL FUNCTION COMPARED TO M|M|1

Thus, with excellent agreement with our model (see 4.5.1) this analysis results in

$$\text{Prob}\ (D<t) \approx 1 - \exp\ -\left(\frac{1-\sqrt{\rho}}{.358-.266\rho^2}\,\mu\right)(t-\mu-h)\ ,\ t>\mu+h \qquad (146)$$

This expression for the (approximate) distribution can be utilized to find the "optimal" packet length (see Section 4.5.4) by expressing ρ and μ as a function of P and then differentiating the mean (or any percentile) to obtain the value of the packet length which minimizes the mean (or percentile).

A packet length optimization calculation is performed in Section 4.6.2; however, a different approximation is used there since the result of this section, involving ρ^2 and $\rho^{1/2}$, leads to algebraic difficulty.

4.5.4 Optimal Packet Length

4.5.4.1 Approach

The model of Section 4.4 can be used to determine the "optimal" packet length where "optimal" needs to be precisely defined. While our *results* apply only to the pendant link case, the *methodology* can be applied to any network configuration, and in particular to the tandem link model of Section 4.7.

The optimally criterion needs to be defined with respect to two issues. The basic concept is to select a packet length which minimizes the network delay. But, since the delay is a random variable, we need to specify some functional of the random variable; that is, a single valued function into the real numbers. For example, given a distribution for D, we can define the optimal packet length as that packet length which minimizes E(D), or K_α where

$$\text{Prob}\ (D \leq K_\alpha) = \alpha. \qquad (147)$$

The second issue is: given a network with a delay distribution D_{ij} defined between every pair of packet switches i and j (whether i and j are adjacent or not), which selection of D_{ij}'s should be used to perform the minimization. Clearly, the packet length which is optimal on a path from i_1 to j_1 may not be optimal on a path from i_2 to j_2.

As we demonstrate below, an important component of the total delay is the packetizing delay experienced at the source. For a single link packet length optimization this effect is more influential than for a more general network where several backbone hops are traversed.

It turns out, in practice, that the criteria based on the α^{th} percentiles ($\alpha \leq .975$) and the mean are roughly equivalent; if the delay is assumed to be exponential with mean γ, all these criteria are <u>exactly</u> equivalent since $K_\alpha = \gamma(-\ln(\alpha-1))$ - directly proportional to the mean. The goal of this investigation, and also of the network designer, is to obtain the range of the optimum packet length rather than the precise value. In our case, we were somewhat limited by the fact that a numerical search was involved necessitating substantial computation for each trial packet length value. For the cases investigated in this section, the optimal length was usually determined to within an interval of 5 to 10 bits.

4.5.4.2 Results

Numerical investigations of the model have shown that the curve of the expected total delay (and also the α^{th} percentiles) is a convex function of P, the packet length excluding overhead. This is expected from the fact that if the overhead is fixed,

1. $E(D) \to \infty$ as $P \to P_{MIN}$, where $P_{MIN}+\emptyset$ is the packet length which results a utilization of 1.

2. $h \to \infty$ as $P \to \infty$, since the packetizing delay, h, is a linear function of P; namely, $h = \frac{P}{B}$.

Figure 27 illustrates the typical behavior of each term in the total delay separately, and the consequent behavior of the total delay. Figure 28 shows the curve obtained for an example with m=18; B=5 KBS; C=100 KBS, $\emptyset$=100; Table 13 contains the corresponding values. Note that the three criteria (E(D), $K_{.01}$, $K_{.05}$) are roughly equivalent and that the optimal packet length is somewhere between 100 bits and 107 bits.

Observation and certain of the approximate theoretical models of Section 4.6 led to the conjecture of the following form for functional dependence of P_{opt}:

$$P_{opt} = \emptyset \cdot f(L,A) \tag{148}$$

where

$$A = \frac{B}{C} \tag{149}$$

$$L = P_m \frac{B}{C} = P_m A \tag{150}$$

and f is a yet unspecified function.

In a series of studies, L and A were kept constant, but the parameters determining L and A were varied. The optimal packet length varied linearly with $\emptyset$; however, the range of P which gave the same value of the functional depended slightly on the optimality criteria, E(D) or K_α. Representative results are shown in Table 14.

A closer look at the functional relation between P_{opt} and L, A is undertaken with the "map" of Table 15. No closed form expression was sought; Section 4.6 supports this decision, since an approximate expression for f(L,A) derived there shows a complicated relationship.

Generally speaking, the optimal packet length as determined from the single link case - and thus we may infer for the whole network - is quite small, around 100 - 150 bits. This is a distinguishing feature of low bitrate packet voice networks, and must be taken into account when developing a design methodology. In

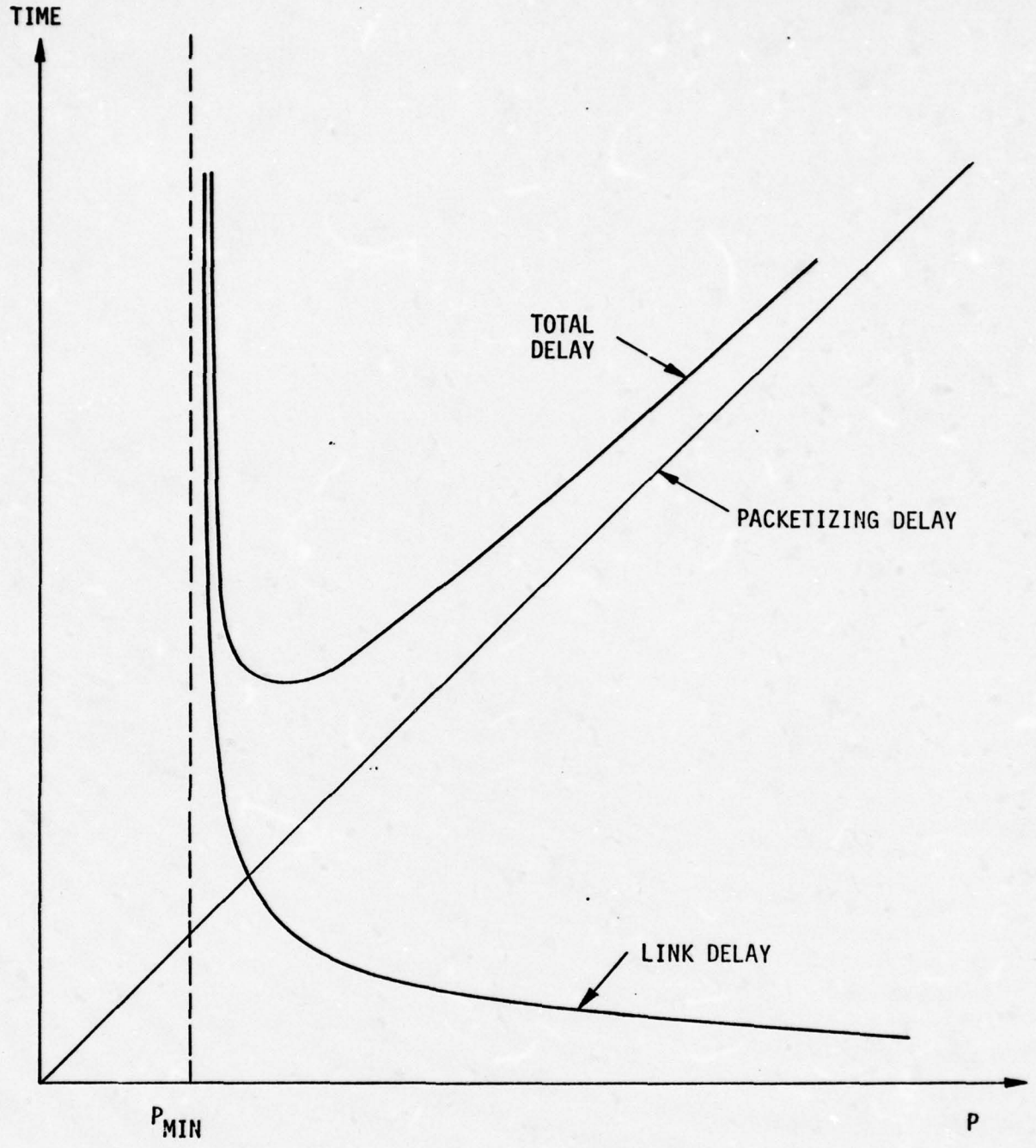


FIGURE 27: CONVEXITY OF TOTAL DELAY VS. PACKET LENGTH

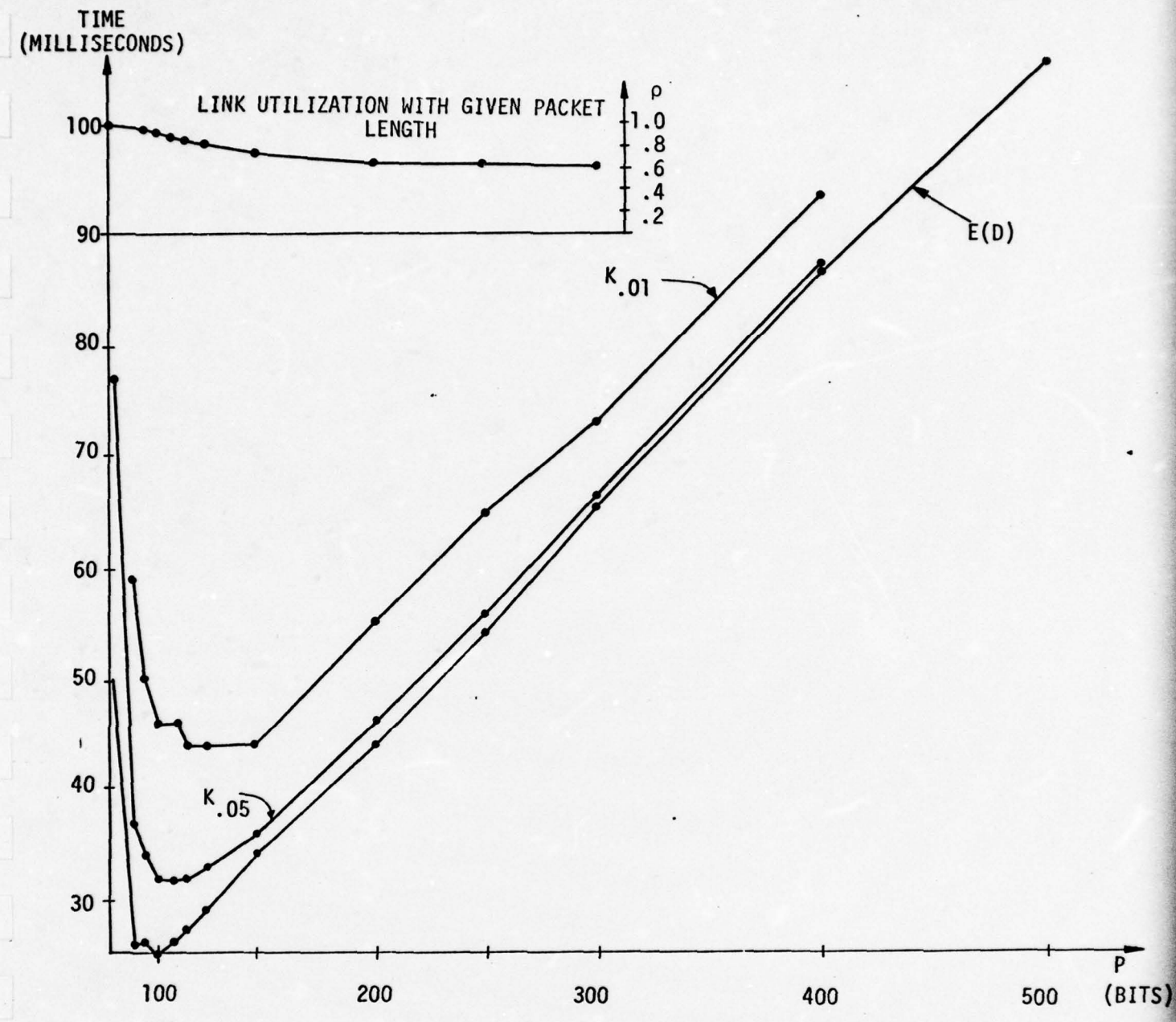


FIGURE 28: REPRESENTATIVE OPTIMIZATION CURVE FOR THE PACKET LENGTH

P	E(D)	95th	99th	ρ
500	.106	.107	.112	.540
400	.086	.087	.093	.563
300	.065	.066	.073	.600
250	.054	.056	.065	.630
200	.044	.046	.055	.675
150	.034	.036	.044	.750
125	.029	.033	.044	.810
112	.027	.032	.044	.852
107	.026*	.032*	.046*	.871
100	.025*	.032*	.046*	.900
95	.026	.034	.051	.924
90	.026	.037	.057	.950
80	∞	∞	∞	1+

m=18, C=100,000, B=5000, Ø=100
Optimum: between 100 and 107 bits

TABLE 13: PACKET LENGTH OPTIMIZATION

m = 18 B = 5,000 C = 100,000 ø = 100 ⩾ L = .45 A = .05 Popt = 107	m = 18 B = 10,000 C = 200,000 ø = 100 ⩾ L = .45 A = .05 Popt = 107	m = 18 B = 2,500 C = 50,000 ø = 100 ⩾ L = .45 A = .05 Popt = 107
m = 18 B = 5,000 C = 100,000 ø = 200 ⩾ L = .45 A = .05 Popt = 214		

TABLE 14: INVARIANCE OF OPTIMAL PACKET LENGTH WHEN L AND A ARE HELD CONSTANT

← A →		← A/2 →	
L	L/2	L	L/2
m = 18 B = 5000 C = 100,000 Ø = 100 ⇒ L = .45 A = .05 [107]	m = 9 B = 1000 C = 20,000 Ø = 100 L = .225 A = .05 [38]	m = 36 B = 2,500 C = 100,000 Ø = 100 ⇒ L = .45 A = .025 [100]	m = 18 B = 500 C = 20,000 Ø = 100 L = .225 A = .025 [33]
m = 18 B = 5000 C = 100,000 Ø = 200 ⇒ L = .45 A = .05 [212]	m = 9 B = 1000 C = 20,000 Ø = 200 ⇒ L = .225 A = .05 [75]	m = 36 B = 2,500 C = 100,000 Ø = 200 ⇒ L = .45 A = .025 [200]	m = 18 B = 50 C = 20,000 Ø = 200 ⇒ L = .225 A = .025 [65]

[XXX] $\Rightarrow P_{opt}$ = XXX

TABLE 15: "MAP" VARIATIONS FOR L, A, AND Ø EFFECTS ON P_{opt}

contrast, a data packet network such as the ARPANET, and the present implementation of packet radio networks, operate with much *longer* packet lengths (around 1000 bits); the primary tradeoff in data being between overhead bits - minimized by long packets - and error retransmission - minimized by short packets. This feature has been generally recognized in the literature [HUGGINS, 1976], [FORGIE, 1975], [COVIELLO, 1976] but no specific values have yet been provided. Even so, the two existing experimental implementations of packet voice networks (ISI and Lincoln Labs) use higher packet lengths, 268 ms. and 134 ms. respectively.

4.5.5 Finite Buffer Case

The finite buffer case is interesting from an engineering point of view. First, there is the issue of how many buffers to provide at a (single link) packet switch; secondly, what is the buffer overflow probability, given that a fixed number of buffers is provided. The modified version of the model of Section 4.4 can be employed to successfully address and answer these issues.

The examples of Section 4.5.1 were rerun with 2, 3 and 4 buffers and the results analyzed.

Figure 29 is a plot of some of these results. The delay distributions for finite buffer cases are contrasted with the infinite buffer situation. Finite buffers situations have a delay distribution which is much less dispersed; the long tail of the delay distribution is cut off. This has the beneficial effect of improving the performance of those packets that are not blocked, at the expense of dropping a few packets. Table 16 compares the mean and the 95^{th} percentile of those packets which are successfully delivered; the improvement effect mentioned above can be seen particularly for high utilization. Note that now the largest possible delay is $K\mu+h$.

We have not yet shown the probability that a packet is blocked - this is expected to increase as the number of buffers decrease. The significant observation is that such probability is low, even when the

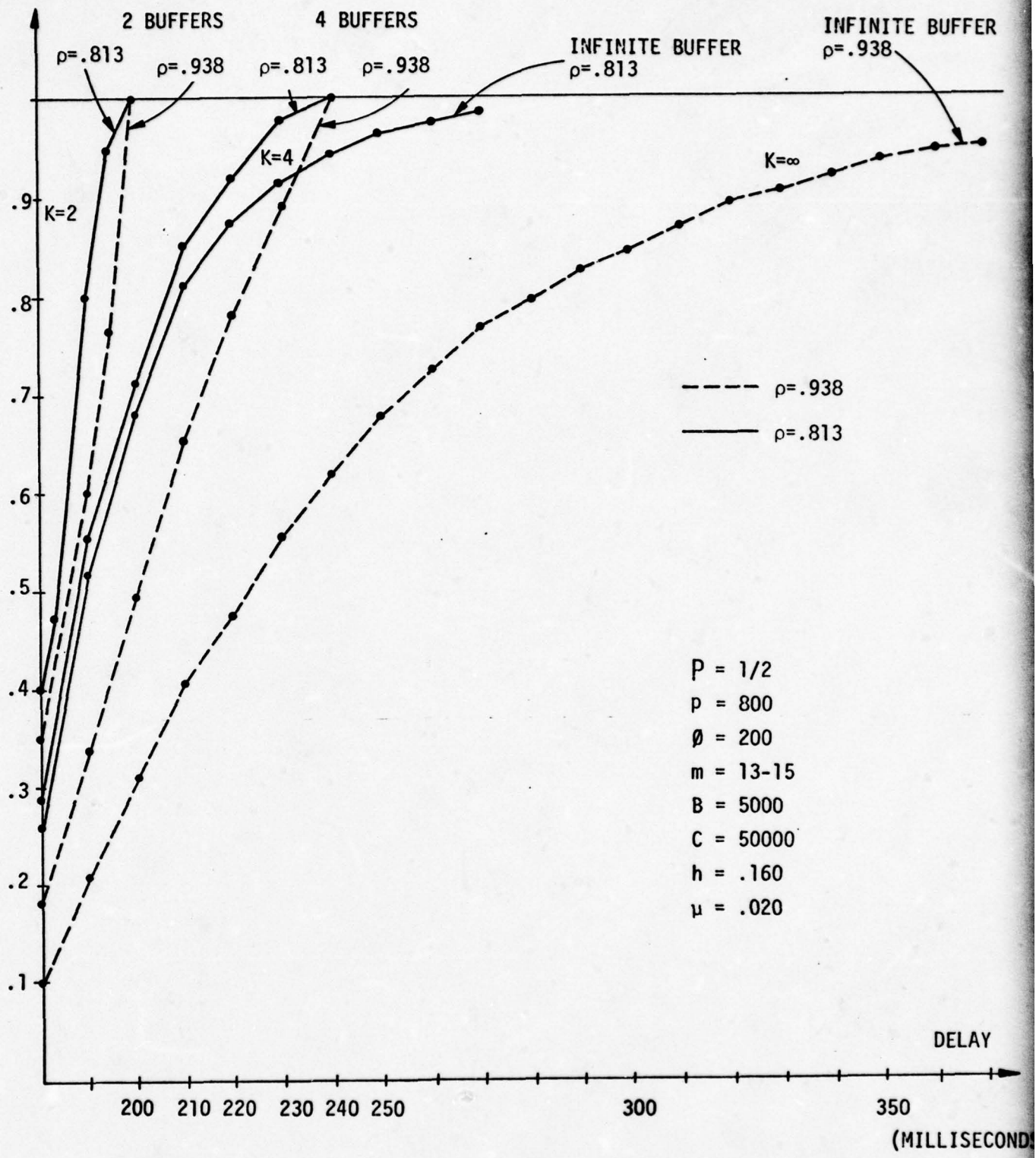


FIGURE 29: COMPARISON BETWEEN THE DELAY DISTRIBUTION FOR INFINITE AND FINITE BUFFER FACILITY

BUFFER SIZE / UTILIZATION	K = ∞		K = 4		K = 3		K = 2	
	E(D)	$K_{.95}$	E(D)	$K_{.95}$	E(D)	$K_{.95}$	E(D)	$K_{.95}$
.563	.182	.188	.182	.188	.182	.188	.182	.188
.688	.187	.207	.187	.205	.187	.203	.185	.190
.813	.197	.238	.194	.226	.191	.213	.186	.196
.938	.239	.357	.203	.234	.196	.216	.188	.198

TABLE 16: EFFECT OF FINITE BUFFER

the number of buffers is small. Table 17 depicts results which are typical for any model parameter selection. Thus, even at ρ=.938, four buffers are sufficient to guarantee that 95% of the packets are not blocked.

The effect of finite buffers on the output distribution has already been discussed in Section 4.5.2: a decrease in the effective utilization of the outgoing link by the factor X, the probability that the buffer is not full.

4.5.6 Effect of Speech Models

Thus far we have kept the speaker model frozen with respect to two factors:

1. The number of states in the speech chain.

2. The transition matrix of the speech chain.

Neither our formulation nor the software package developed to evaluate the single link model are restricted to this particular situation. In this section we explore the implications of varying the speaker model.

4.5.6.1 Transition Matrix Changes for the Two-State Model

Because of the symmetry requirement, the steady state distribution is invariant under changes in the transition matrix. This follows, quite simply, because the steady state distribution of the arrival process was shown in Section 4.2 to be

$$P(\mathcal{Z}^{(\infty)}=k) = \binom{m}{k}\left(\frac{1}{2}\right)^m \qquad (151)$$

for a two-state chain. Changes in the transition matrix affect only the transient behavior unless the symmetry requirement is eliminated.

	K = ∞	K = 4	K = 3	K = 2
ρ=.563	1.0	1.0	1.0	.9995
ρ=.688	1.0	.9996	.9967	.9653
ρ=.813	1.0	.9905	.9738	.9139
ρ=.938	1.0	.9557	.9275	.8653

TABLE 17: FRACTION OF PACKETS NOT BLOCKED

NETWORK ANALYSIS CORPORATION

4.5.6.2 Three-State Speaker Model

A brief investigation of a three-state speaker model was undertaken. It has already been indicated that the two-state model was very conservative - predicting more traffic than actual. The average number of packets per frame supplied by a population of m users under the two and three-state model are, respectively,

$$\frac{m}{2}$$

and

$$\frac{mx}{2x+p}$$

where the three-state transition matrix is

$$\begin{pmatrix} 1-p & p & 0 \\ x & 1-2x & x \\ 0 & p & 1-p \end{pmatrix}.$$

Typical values of the transition probabilities are p=.1 - high tendency to continue speech, once initiated; x=.46 - high tendency to break away from mutual silence. Thus for m terminals we would expect .45m packets compared with .5m as for the two-state. With the assumption of a three-state speech model, 10% more users can be accommodated (if x=.23, we could accommodate 20% more speakers). The delay distribution for the examples of Section 4.5.1 have been obtained for the three-state model and are depicted in Figure 30 for infinite buffers. Note that the same number of terminals - 13 and 15 - give a smaller utilization than before. We observe that the delay distribution is fairly similar to the distribution obtained with a two-state model at the same utilization. Compare: ρ=.733, S=3; ρ=.845, S=3; and ρ=.813, S=2.

The effect of buffer size is similar to that described in Section 4.5.5. For the three-state speaker model with the same value of m,

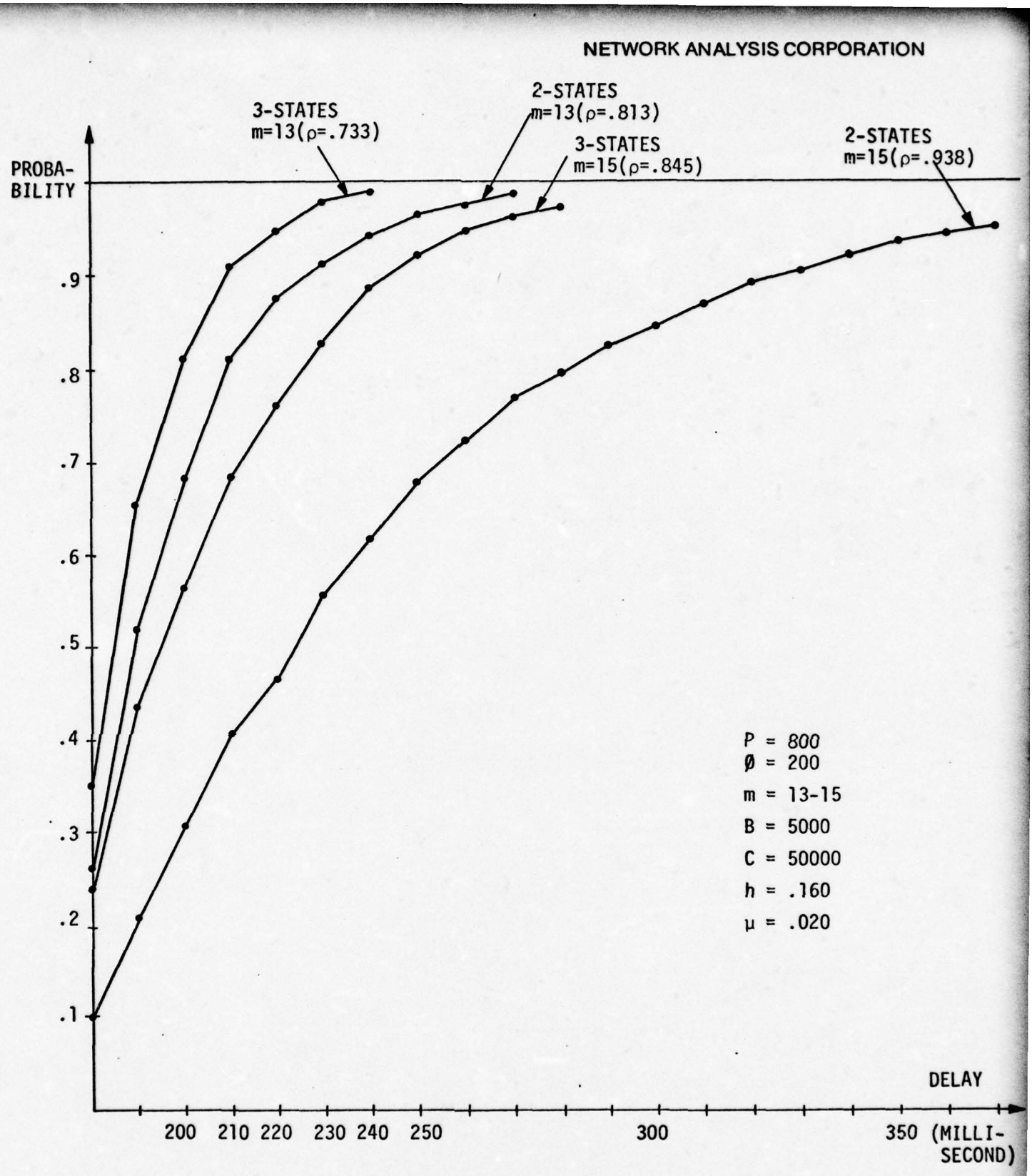


FIGURE 30: COMPARISON BETWEEN THE DELAY DISTRIBUTION FOR A 2-STATE SPEECH MODEL AND A 3-STATE SPEECH MODEL

fewer packets arrive at the packet switch so the number of buffers necessary for acceptable operation is the same or reduced. Additionally, the three-state speaker model has little or no effect on the steady state output distribution, reflecting only the change in utilization. The explicit dependence of E(D) on ρ, h and μ continues to hold - however the form of the functional relation was not investigated; packet size optimization with three-state speakers yield similar results.

4.5.7 Transient Behavior

The model of Section 4.4 can be employed to study the *transient* behavior of the single link. This reveals the time duration of degraded performance when a "bad" state is entered - i.e., unusual period of high speech activity in one direction.

The basic procedure involves finding the steady state delay, then perturbing the speaker activity states and observing the effect on the system. As an extreme, at a particular frame, every terminal is forced to supply a packet by re-starting each speaker Markov chain in state 1 (active) with probability 1. This situation produces the worst case transient which can arise. Out of the 2^m arrival patterns, it occurs with probability $(\frac{1}{2})^m$. On the average we would expect it only every $(.5)^{-m}$ frames; for m=20, this is about one in 1,000,000 frames; in other terms, for a frame length of .100 seconds, this transient would occur once every 30 hours, on the average.

This perturbation overloads the system for a period of time, causing a related increase in the delay faced by a typical incoming packet. Figure 31 shows the instantaneous expected delay as a function of time for various examples. The following observations can be made:

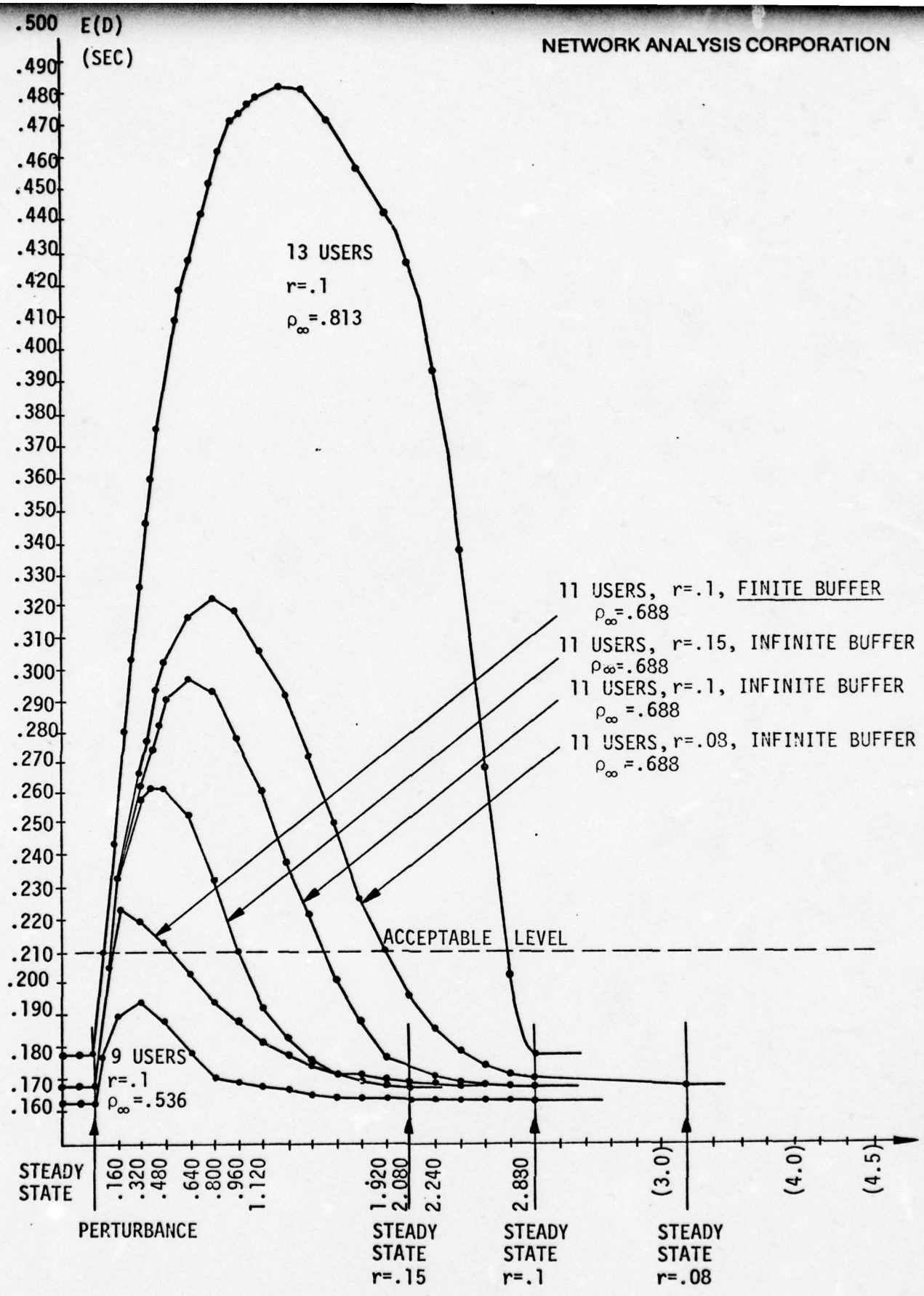


FIGURE 31: TRANSIENT BEHAVIOR

1. The maximum delay *perturbation* caused by the transient is very dependent on the steady state link utilization and the speaker behavior parameter, r.

2. The *duration* of the transient is a function of r only. As r decreases there is more frame-to-frame correlation as shown in Figure 31; as this frame-to-frame correlation increases (r=.15, r=.1, r=.08) the duration of the transient increases.
   The transient dies off rather rapidly (2-3 seconds) for typical values of r, $r \geq .08$.

3. One example was run with 4 buffers. Finite buffer size has a damping effect on the transient. This behavior is intuitive. The system becomes overloaded: if the buffer is infinite, there is no other way to unload but to pump the packets out the line; in the case of finite buffer, an unloading takes place when the packets are blocked because there is no room in the buffer. The blocked packets do not impact the future.
   Figure 32 shows increased packet blocking rate for this example immediately following the overload.

Typical transient delay distributions are shown in Figure 33(a) and (b) while Figure 34 depicts the "overload" situation. Note that utilization is temporarily pushed over 1.

Because packet voice networks would be designed to be driven at high utilization we must guard against transients or at least be aware of their effect. Figure 31 shows a radical departure

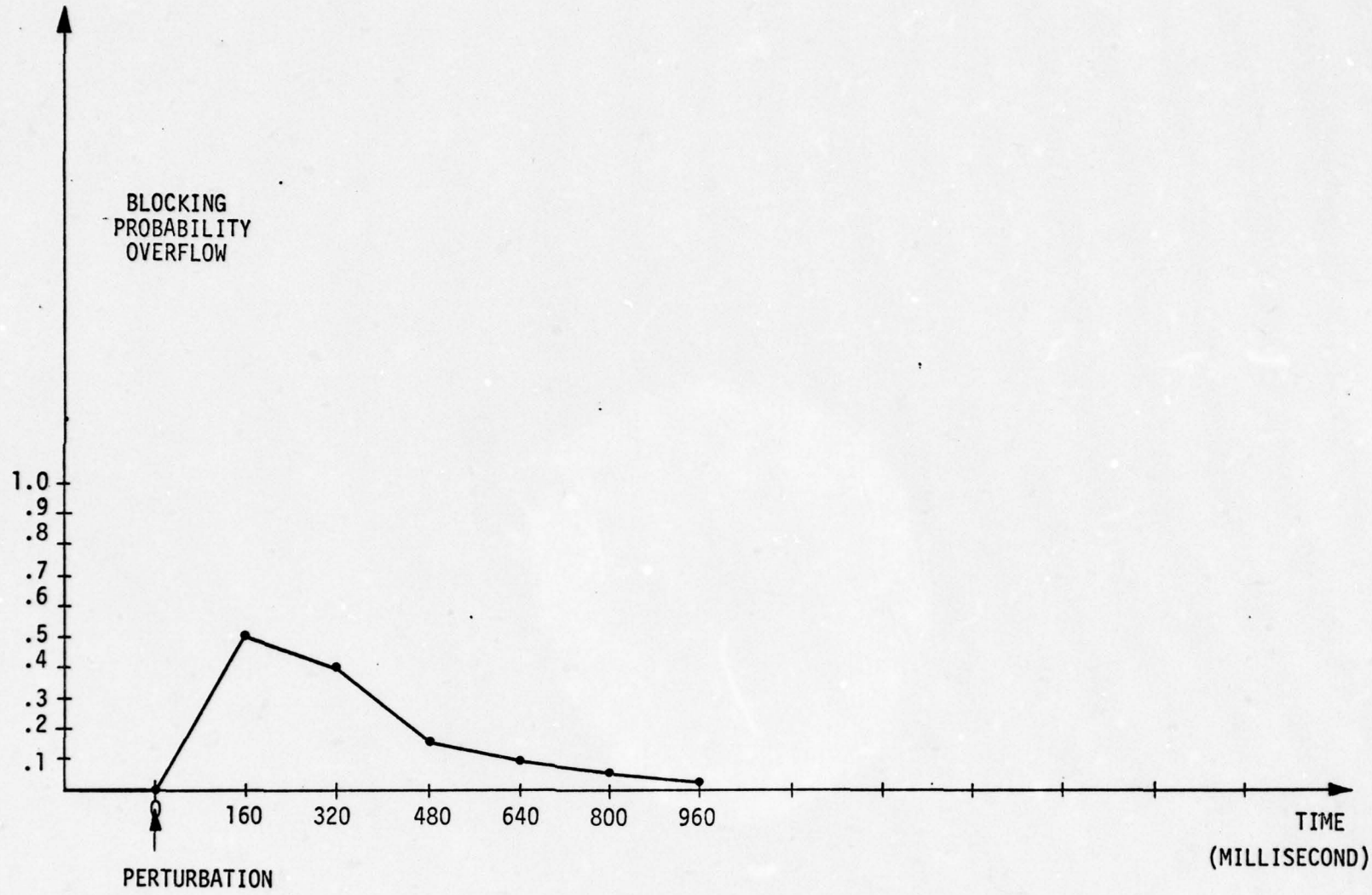


FIGURE 32: TRANSIENT BLOCKING RATE FOR A FOUR BUFFER SYSTEM

NETWORK ANALYSIS CORPORATION

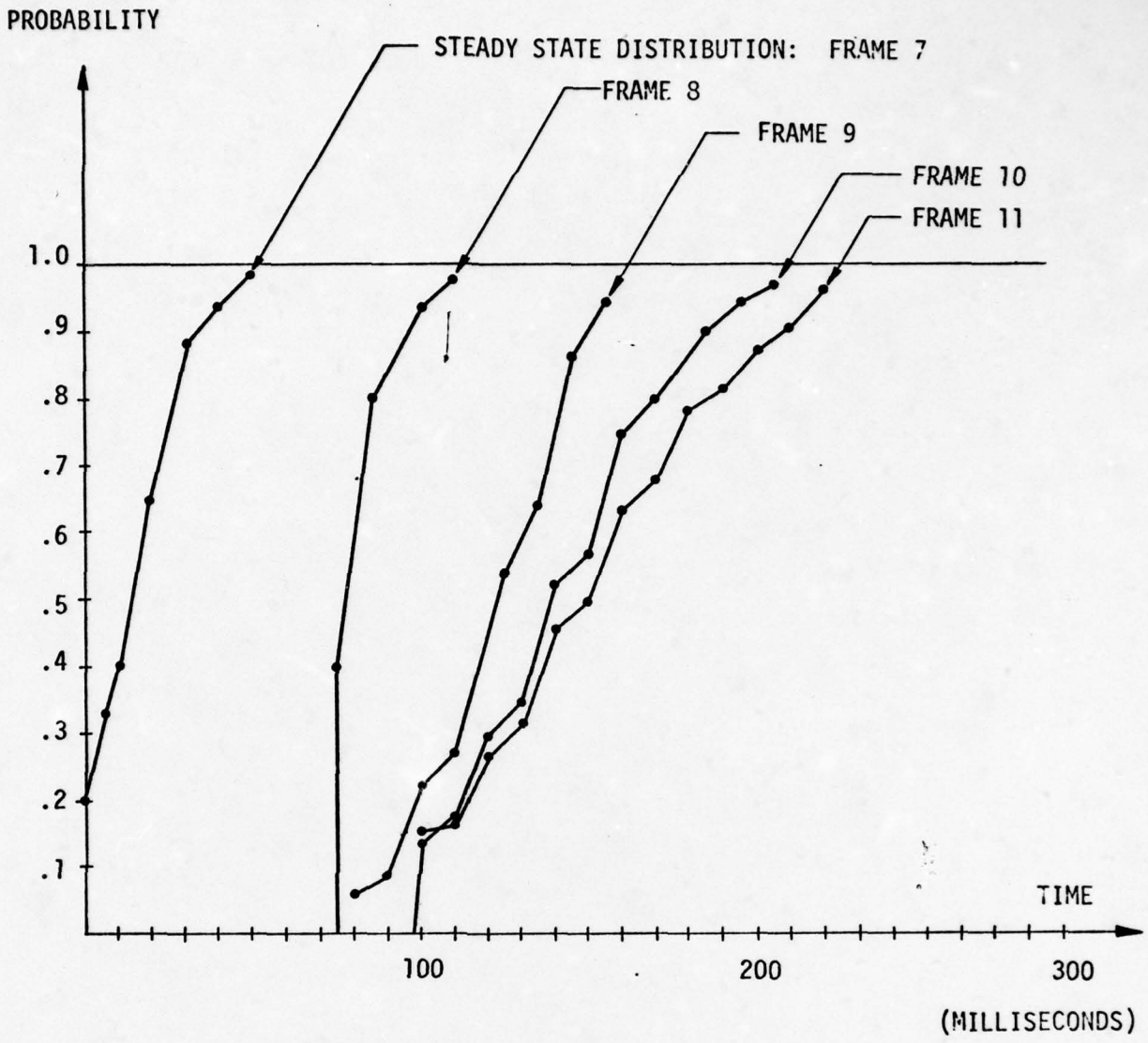


FIGURE 33(a): TRANSIENT DELAY DISTRIBUTION

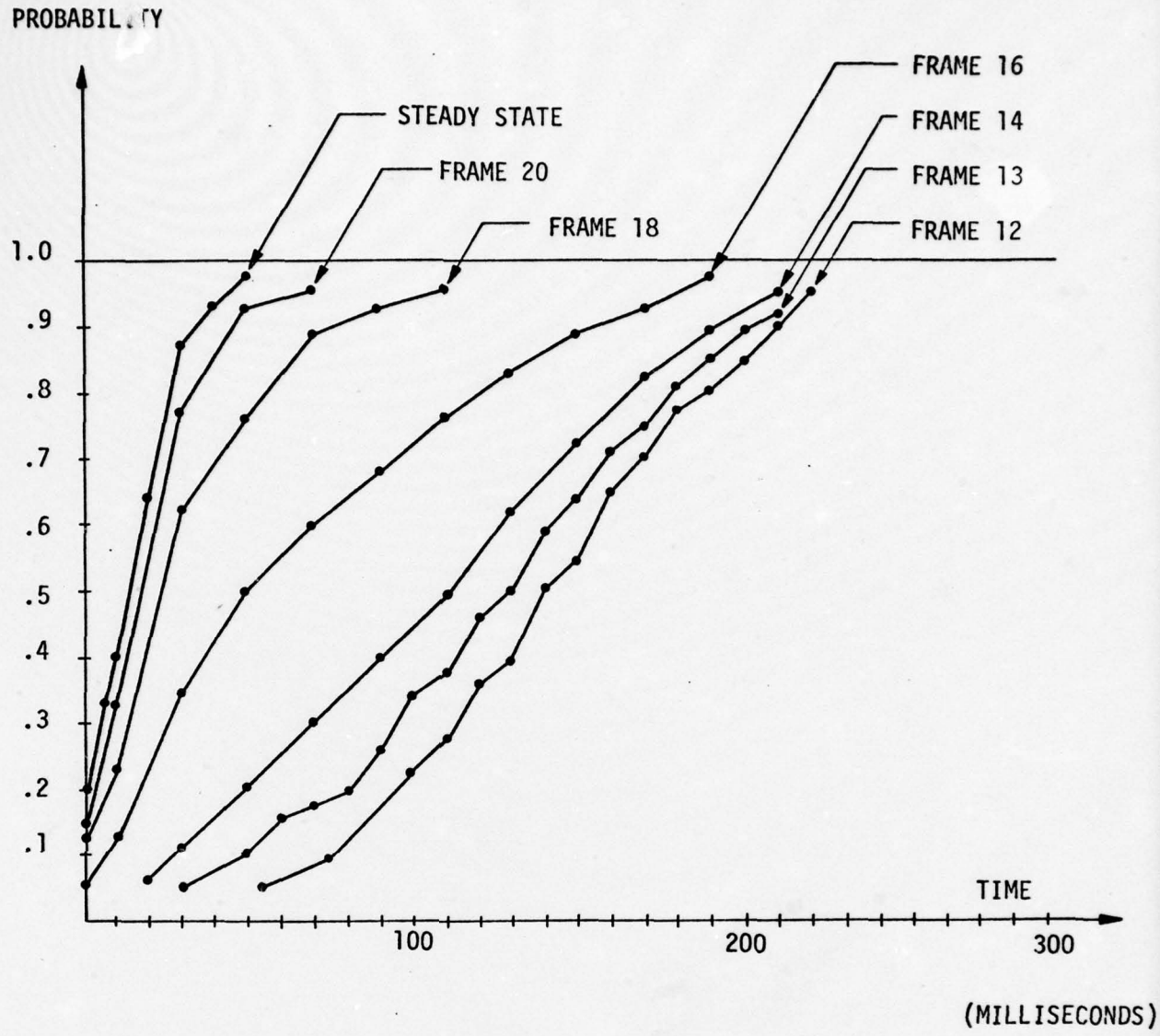


FIGURE 33(b): TRANSIENT DELAY DISTRIBUTION

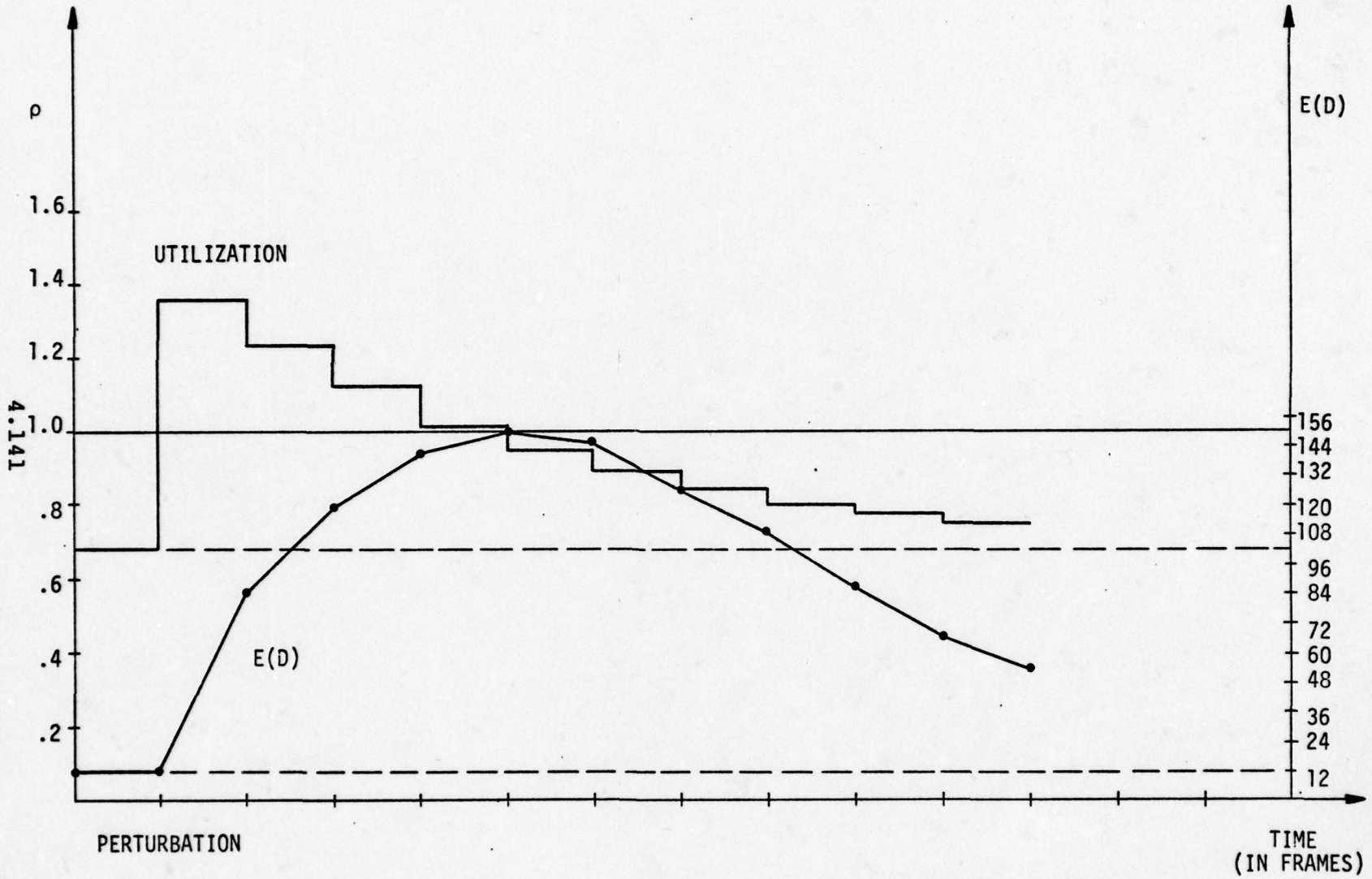


FIGURE 34: INSTANTANEOUS UTILIZATION

NETWORK ANALYSIS CORPORATION

from steady state for a utilization as low as .813. A finite buffer situation has the beneficial effect of preventing extremely high instantaneous delays; note, however, that for a typical speaker parameter r, even a severe transient - as the one discussed above - is damped to acceptable levels in 2-3 seconds.

The length of the transient can be computed for the two-state speaker model. We already have from Section 4.2 that with symmetry

$$p_{11}^{(n)} = \frac{1}{2} + \frac{(1-2r)^n}{2} . \qquad (152)$$

We are interested in obtaining the n which makes $p_{11}^{(n)}$ close to $p_{11}^{(\infty)}$, say by no more than 1/2%.

We require

$$\frac{(1-2r)^n}{2} \leq .005 \qquad (153)$$

or for the cases presented in Figure 31:

r	n
.15	13
.10	21
.08	27

A typical example of transient output process behavior is shown in Figure 35.

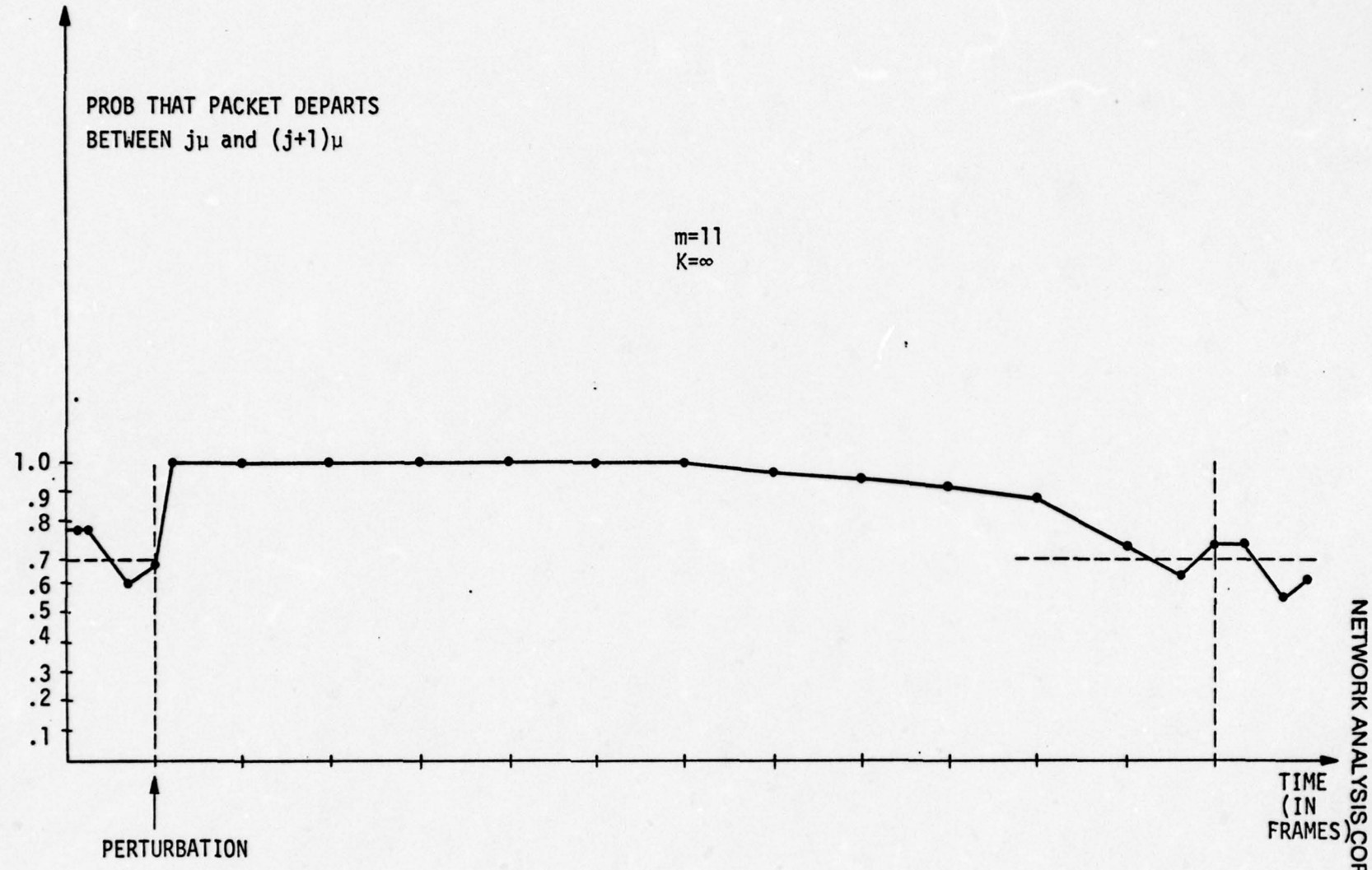


FIGURE 35: TRANSIENT DEPARTURE PROBABILITIES

NETWORK ANALYSIS CORPORATION

4.6 APPROXIMATIONS

In spite of the wealth of results obtainable with the detailed link model of Section 4.4, its complete interface with any traffic model, its relative computational tractability (for $\rho \leq .9$), and its ability to assess the effect of parameter variations on performance, it is instructive to examine standard queues in an attempt to acquire further insight. The purpose of this section is to supply more approximate, but closed form, expressions for the link behavior. This will be done by using exact solutions of queueing models which only approximate the packet voice situation. These exact solutions are mainly adopted from standard references on queueing: [SAATY, 1961] for Section 4.6.1; [KLEINROCK, 1976] for Sections 4.6.2 and 4.6.3. Comparisons between the closed form approximations and results from our more precise model are carried out. All of the approximations consider only the infinite buffer case; therefore, to examine the effect of finite buffers we must resort to the Section 4.4 model.

4.6.1 M|D|1 Approximation

In standard queueing terminology, the model which comes closest to representing a pendant link under the traffic generated by a population of off-hook packet voice terminals in the steady state is a Y|D|1 queue, where Y is a scaled geometric random variable representing the packet interarrival statistics and D represents the deterministic (i.e., constant) service time distribution of the single server packet transmission process. We will first characterize the interarrival distribution Y; then, to avoid deriving in closed form the distribution of delay for this queue model - clearly a complex task - we instead approximate the queue with an M|D|1 model, for which closed form solutions are available. We also discuss a result for a special case of geometric interarrival.

4.6.1.1 Input Distribution Y

In our previous development we have always imposed a synchronization on the stream of arrivals from each terminal. We can assume any of the traffic models discussed in Section 4.2. In the steady state, a particular user chain, X_i, is in the active state (supplies a packet) with probability P_i and is inactive (does not supply a packet) with probability $Q_i = 1 - P_i$. We again make the assumption that these probabilities are independent of i - a crucial assumption for the development, but a reasonable one - so that each terminal supplies a packet with probability P, and does not supply a packet with probability Q. Retaining our assumptions of synchronized potential packet arrival scheduling, constant frame length, and uniform Δ-spacing ($\Delta=h/m$), we have, as before, the arrival of a packet (if any) from the i^{th} user in the r^{th} frame at time $rh + i\Delta$. Then the interarrival time of packets to the queue can be obtained as follows.

Assume that at time $k\Delta$ a packet arrives on the queue; such packet must have been issued by user $\Theta(k)$ where

$$\Theta(k) = \begin{cases} \text{mod}(k,m) & \text{if mod}(k,m) \neq 0 \\ m & \text{if mod}(k,m) = 0 \end{cases} \qquad (154)$$

at frame

$$\Psi(k) = \sup_{n \leq \frac{k}{m}} n, \qquad n \text{ integer.} \qquad (155)$$

The interarrival time, Y, is

Δ if $\Theta(k+1)$ issues a packet at frame $\Psi(k+1)$

2Δ if $\Theta(k+1)$ does not issue a packet at frame $\Psi(k+1)$ but $\Theta(k+2)$ issues a packet at frame $\Psi(k+2)$

3Δ if ...

etc.

Then

$$\left.\begin{aligned} Y &= \Delta \quad \text{with prob } P \\ Y &= 2\Delta \quad \text{with prob } P\,Q \\ &\vdots \\ Y &= n\Delta \quad \text{with prob } P\,Q^{n-1}. \end{aligned}\right\} \tag{156}$$

Since a geometric random variable B has distribution

$$\text{Prob}(B=i) = p\,q^{i-1} \qquad i = 1,2,\ldots\infty, \tag{157}$$

with $p + q = 1$ and $0 \leq p \leq 1$,

we see that

$$Y = \Delta B. \tag{158}$$

We call Y a scaled geometric random variable. The mean interarrival time is

$$E(Y) = \Delta E(B) = \frac{\Delta}{P} \tag{159}$$

and the variance is

$$V(Y) = \Delta^2 V(B) = \Delta^2 \frac{Q}{P^2}. \tag{160}$$

The cumulative distribution of Y is

$$\text{Prob}(Y \leq i\Delta) = P \sum_{j=1}^{i} Q^{j-1} = 1-Q^i. \tag{161}$$

A closed form solution for a somewhat related queueing situation was obtained by [SCHMOOKLER, 1970]. His solution is valid under the following conditions:

1. The service time μ is a integer multiple of the arrival interval Δ.

2. Arrivals are uncorrelated (not precisely true in our situation).

For geometric input with $\mu=K\Delta$, K integer and μ and Δ fixed (deterministic service), he obtained the result

$$E(d) = \left(\frac{2K-\rho(K+1)}{2(1-\rho)}\right)\Delta = \frac{1-\frac{\rho}{2}\left(\frac{K+1}{K}\right)}{1-\rho}\,\mu \,. \tag{162}$$

For the case K=1 we get $\mu=\Delta$, no queue is formed and the formula trivially reduces to $E(d)=\mu$. Since we have

$$\rho = P\left(\frac{\mu}{\Delta}\right) = P_K \tag{163}$$

the conditions are only satisfied when

$$P = \frac{\rho}{K} \,. \tag{164}$$

For the case K=2 this relationship cannot be satisfied except for high ρ and many-state speaker models. For this case we can get the following expression for f_3 of Section 4.5.3

$$f_3(\rho) = \frac{\frac{\rho}{4}}{1-\rho} \qquad P = \frac{\rho}{2} \tag{165}$$

which has reasonable agreement with our previous result in the range of its validity.

The cases of $K \leq 3$ would require too low a P to represent realistic speaker models. Thus this result has a very limited range of applicability to our situation.

4.6.1.2 Exponential Interarrival Approximation

The above derivation of Y has shown that the interarrival distribution is discrete in the sense that an arrival must occur *exactly* on a multiple of Δ.

An exponential approximation to Y with the same average interarrival time, would have a continuous distribution with more variance than the original geometric distribution. In particular, if M is a exponential random variable with parameter $\frac{P}{\Delta}$ the distribution density function is

$$f_M(t) = \frac{P}{\Delta} e^{-\frac{P}{\Delta} t} \tag{166}$$

and $E(M) = \frac{\Delta}{P}$ as arranged, but

$$V(M) = (\frac{\Delta}{P})^2 \tag{167}$$

and

$$\frac{V(Y)}{V(M)} = Q < 1, \tag{168}$$

or rewriting the inequality

$$V(Y) < V(M). \tag{169}$$

Figure 36 compares the cumulative distribution functions for the two distributions. Note a reasonable agreement. We expect the exponential approximation to be conservative in the sense of higher delay estimates because of the higher variance. The M|D|1 queue results are well known [SAATY, 1961]. It is shown there that the distribution of the queue waiting time is

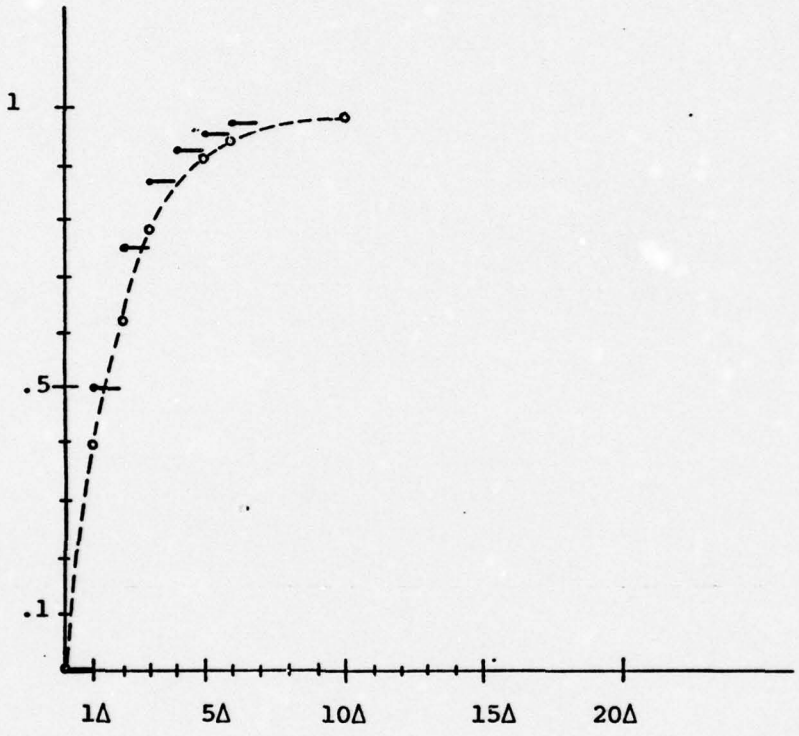


FIGURE 36: EXPONENTIAL APPROXIMATION TO GEOMETRIC

$$P(W_q<t) = (1-\rho) \sum_{i=0}^{k} e^{(t/\mu-i)} \frac{[-\rho(t/\mu-i)]^i}{i!} \qquad (170)$$

where k is the largest integer less then t/μ, and $\rho = P(\frac{\mu}{\Delta})$ as before. We also have

$$E(W_q) = \frac{\rho\mu}{2(1-\rho)} \qquad (171)$$

so that

$$E(d) = \frac{\rho\mu}{2(1-\rho)} + \mu = \left(\frac{1-\frac{\rho}{2}}{1-\rho}\right)\mu \qquad (172)$$

where *d* is, as usual, the time (waiting plus service) in the system. While we now have a closed form expression for the delay distribution, the complexity in computing it is non-trivial. Both the distribution and the variance require complex computation to evaluate and perform the comparison with our more exact model.

The expression for average delay is simpler and can be specialized to our problem by straightforward substitution of our link parameters as follows.

We have from previous calculations

$$\mu = \frac{P+\emptyset}{C} \qquad (173)$$

and

$$\rho = \frac{P_{mB}}{C}\left(1+\frac{\emptyset}{P}\right) \qquad (174)$$

yielding

$$E(d) = \left[\frac{P+\emptyset}{C}\right]\left[\frac{P(1-\frac{P_{mB}}{2C})-\frac{P_{mB}}{2C}\emptyset}{P(1-\frac{P_{mB}}{C})-\frac{P_{mB}}{C}\emptyset}\right]. \qquad (175)$$

Letting $L = P_{mB}/C$, the zero overhead utilization, we get

$$E(d) = \left[\frac{P+\emptyset}{C}\right]\left[\frac{P(1-\frac{L}{2})-\frac{L}{2}\emptyset}{P(1-L)-L\emptyset}\right], \qquad 0 \leq L \leq 1. \qquad (176)$$

4.6.1.3 "Optimal" Packet Length for M|D|1 Approximation

We wish to use this M|D|1 result to find an "optimal" packet length. "Optimal" is very loosely used here since the model is approximate and minimizing the mean delay is not necessarily the best performance criterion. However, we have shown in Section 5 that the criteria for mean and the K^{th} percentile ($K \leq 97.5$) yield similar answers to the optimal packet length problem. The total delay, D, is composed of the queueing delay (waiting plus service times) and the packetizing delay. Thus

$$E(D) = E(d) + \frac{P}{B}$$

$$= \frac{P^2[B(1-\frac{L}{2}) + C(1-L)] + P\emptyset[B(1-L)-CL] - \emptyset^2(\frac{BL}{2})}{BC[P(1-L)-L\emptyset]} \qquad (177)$$

Taking $\frac{dE(D)}{dP}$ and setting it equal to zero gives the following equation:

$$0 = P^2(1-L)[B(1-\frac{L}{2})+C(1-L)]-P\emptyset(2L)[B(1-\frac{L}{2})+C(1-L)]+\emptyset^2(L)[CL-\frac{B}{2}(1-L)]. \qquad (178)$$

If we now let $A=\frac{B}{C}$ and $\tilde{P}=\frac{P}{\emptyset}$, we get an equivalent equation by dividing through by the factor $\emptyset^2 C$. In our new terms we have

$$0 = \tilde{P}^2(1-L)[A(1-\frac{L}{2})+(1-L)] - \tilde{P}(2L)[A(1-\frac{L}{2})+(1-L)] + (L)[L-\frac{A}{2}(1-L)]. \qquad (179)$$

Since $L=(\rho_m)A$, we have the important conclusion that the "optimal" packet length is a multiple, $\tilde{P}$, of the overhead, and the value of this multiple is a function only of the ratio B/C and the effective number of speakers, ρ_m. Note that the exact "optimal" length may turn out *not* to be an integral number of bits.

We now pursue this approach and obtain closed form expressions for the "optimal" packet length and the optimal mean delay. The above equation can be divided through by the leading coefficient; since $L<1$ and $A>0$, the coefficient is assured to be positive. We get

$$0 = \tilde{P}^2 - 2\left(\frac{L}{1-L}\right)\tilde{P} + \left(\frac{L}{1-L}\right)\left[\frac{L-\frac{A}{2}(1-L)}{(1-L)+A\left(1-\frac{L}{2}\right)}\right]. \tag{180}$$

The range of interest is where A is not more than 2L (i.e., $\rho_m \geq \frac{1}{2}$) and in this range the right-hand term is always positive. Thus the roots are of the same sign. Also, the roots sum to $2\left(\frac{L}{1-L}\right)$, the coefficient of the middle term. So one root must be larger than $\frac{L}{1-L}$ and one root smaller. From the expression

$$L\left(1+\frac{\phi}{P}\right) = \rho \tag{181}$$

and the fact that $\rho \leq 1$ for stability, we substitute $\tilde{P}=\frac{P}{\phi}$ and get

$$\tilde{P} > \frac{L}{1-L} \tag{182}$$

for stable operation. Thus we are interested only in the larger root of the quadratic equation in $\tilde{P}$. After some manipulation we get

$$\tilde{P}_{opt} = \left(\frac{L}{1-L}\right)(1+F) \tag{183}$$

where

$$F = \sqrt{1 - \frac{1-\frac{A}{2}\left(\frac{1-L}{L}\right)}{1+A\left(\frac{1-\frac{L}{2}}{1-L}\right)}} \tag{184}$$

for the optimal value of $\tilde{P}$. Furthermore

$$P_{opt} = \phi\left(\frac{L}{1-L}\right)(1+F) \tag{185}$$

and

$$E(D)\Big|_{P=P_{opt}} = \left[\frac{\phi}{B}\,\frac{\left(\frac{L}{1-L}\right)(F+1)^2\left[A\left(\frac{1-\frac{L}{2}}{1-L}\right)+1\right] + (F+1)\left[A-\frac{L}{1-L}\right] - \frac{A}{2}}{F}\right]. \tag{186}$$

We illustrate these relationships with an example.

Given

$$P = \frac{1}{2}$$

$$m = 18$$

$$B = 5 \text{ KBS}$$

$$C = 50 \text{ KBS}$$

we can calculate A = .1 and L = .9. Using the equation for $\tilde{P}_{opt}$ we get

$$\tilde{P}_{opt} = 14.388.$$

Suppose the protocols adopted required $\emptyset$ = 100 bits, then we would select 1439 as the best packet length for the speech portion. This assumes compatibility with the speech terminals and ignores the issue of line errors and possible retransmissions. The resulting mean delay is .714 sec., which, of course, is unacceptable in practice. The utilization is a very high .9625. However, as our results predict, trying to reduce the mean delay by decreasing the packet size will only worsen the situation. Increasing the packet length, say, to 2000 bits (packet speech length = 1900 bits, packet overhead = 100 bits) reduces the utilization to .947 but the corresponding increase in packetizing delay now gives a mean delay of .78 sec. - about 10% worse!

Of course, even the mean delay experienced at the optimal packet length is in the unacceptable range. Thus the 50 KBS line cannot adequately support the 18 users with the given characteristics.

If we now assume the line is upgraded to a capacity of 100 KBS, for the same assumption of 100 bits of overhead, we get the following results:

P_{opt} = 107 bits

E(D) = .030 sec.

ρ = .87

which is very acceptable performance from the point of view of delay - but a heavy price is paid in terms of overhead. The L factor - representing the effective utilization - has been reduced in half from 90% to 45%.

The results on optimal packet length obtainable with this formula are in excellent agreement with those obtainable empirically from the detailed model. Table 18 compares values of optimal packet length computed from the detailed model with values obtained from the formula. Nominal overhead values of 100 and 200 are used - these overheads are not included in the packet lengths reported in the tables. Excellent agreement is observed.

The closed form of the approximation result permits an evaluation of the optimal packet length for a wide range of typical situations of interest. Table 19 shows the formula-computed optimal packet lengths (without overload) and the delays incurred for two digitization notes: a 2.5 KBS (vocoder) rate and a 50 KBS (PCM) rate. A wide range of cases is spanned by our selection of A and L. Note that the 2.5 KBS digitization rate cannot tolerate the total delay across even a single link with packet size optimized, if the value of A or L are high with 100 bits of overhead.

4.6.2 M|M|1 Approximation

The approximation of the previous subsection was in good agreement with the detailed model since we only approximated the input process - everything else was kept as in the detailed model.

	m (terminals)	18	9	36	18	18	9	36	18
	B (KBS)	5	1	2.5	5	5	1	2.5	5
	C (KBS)	100	20	100	20	100	20	100	20
	ø (bits)	100	100	100	100	200	200	200	200
Optimal Packet Length (bits w/o overhead)	Model	107	38	100	33	212	75	200	65
	Formula	106.9	39.7	99.8	36.6	213.8	79.4	199.6	73.2

TABLE 18: COMPARISON OF OPTIMAL PACKET SPEECH LENGTH DETERMINATIONS

NETWORK ANALYSIS CORPORATION

A \ L	.5				.6				.7				.8			
	m	P_{opt}	E(D) (sec) 2.5KBS	50KBS	m	P_{opt}	E(D) (sec) 2.5KBS	50KBS	m	P_{opt}	E(D) (sec) 2.5KBS	50KBS	m	P_{opt}	E(D) (sec) 2.5KBS	50KBS
.05*	20	130	.074	.004	24	196	.112	.006	28	309	.183	.009	32	537	.345	.017
.1	10	141	.094	.004	12	213	.114	.007	14	336	.238	.012	16	536	.413	.023
.2	5	155	.129	.006	6	233	.200	.010	7	354	.289	.014	8	650	.680	.034
.4	2.5	170	.194	.010	3	255	.302	.005	3.5	400	.529	.026	4	701	1.090	.054
1**	1	189	.378	.018	1.2	280	.596	.029	1.4	435	1.020	.052	1.6	753	2.28	.114

*High Capacity
50 KBS @ B=2.5K
1 MBS @ B=50K

**Low Capacity
2.5 KBS @ B=2.5K
50 KBS @ B=50K

Ø=100 BITS

TABLE 19: OPTIMAL PACKET SPEECH LENGTH AND AVERAGE DELAY FROM FORMULA

NETWORK ANALYSIS CORPORATION

In this subsection we relax the conditions even more, approximately both the input and the service processes. It is clear that with this approach we lose considerable accuracy. The arrival and service time distributions used in the approximation are quite different from the distributions of the Section 4.4 model. What we gain, however, is simplicity in the delay distribution. Since the M|M|1 queue has been extensively studied, there exists a wealth of closed form results which are extendable to our model, as an approximation.

With μ, as the *average* service time, we obtain immediately

$$E(W_q) = \frac{\rho\mu}{1-\rho} \qquad E(d) = \frac{\mu}{1-\rho} \tag{187}$$

$$P(W_q \leq y) = 1-\rho e^{-(1-\rho)y/\mu} \qquad y \geq 0 \tag{188}$$

$$P(d \leq y) = 1-e^{-(1-\rho)y/\mu} \qquad y \geq 0. \tag{189}$$

With the exception of the accumulation of probability at the origin we note that the distributions are essentially exponential. We can apply the technique used above to solve for an "optimal" packet length, using the somewhat different expression for $E(W_q)$. Furthermore, because of the very tractable form for the cumulative distribution function we can get an explicit closed form expression for "optimal" packet length where "optimal" refers to a specific percentile (i.e., cut-off point on the tail of the delay distribution).

4.6.3 G|G|1 Heavy Traffic Approximation

We try, as a third alternative approximation, closed form results which do not require any assumptions about the distributions of service or interarrival times - thus the G, for general, in the subsection heading. Under appropriate heavy utilization assumptions -

warranted in our case because this is the operating situation we normally expect to design a packet voice network for - the distribution of delay, is shown to be approximately exponential, with an appropriately selected parameter. In this case we can also obtain, in approximate form, the variance of the delay.

It is shown in [KLEINROCK, 1976] that, under heavy traffic ($\rho \approx 1$), the waiting time distribution is approximately exponentially distributed with mean waiting time

$$E(W_q) = \frac{\sigma_A^2+\sigma_B^2}{2(1-\rho)\bar{t}} \tag{190}$$

where σ_A^2 and σ_A^2 are the respective variances of the interarrival and service time distributions and $\bar{t}$ is the mean interarrival time. This approximation is valid when the denominator is small compared to the square root of the numerator in the above expression for $E(W_q)$. Accepting the approximation, we obtain

$$P(W_q<t) \simeq 1-\exp\left\{-\frac{2\bar{t}(1-\rho)}{\sigma_A^2+\sigma_B^2}t\right\} \tag{191}$$

$$V(W_q) = \left(\frac{\sigma_A^2+\sigma_B^2}{2(1-\rho)\bar{t}}\right)^2 . \tag{192}$$

In our case, with $\Delta = h/m$

$$\bar{t} = \frac{\Delta}{P} = \frac{P}{P_m B} \tag{193}$$

$$\sigma_A^2 = \frac{(1-P)P^2}{P^2 m^2 B^2} \tag{194}$$

$$\sigma_B^2 = 0. \tag{195}$$

Then in terms of our system parameters

$$E(W_q) = \left(\frac{1-P}{P}\right)\left(\frac{P}{mB}\right)\left(\frac{1}{1-\frac{PmB}{C}\ [1+\emptyset/P]}\right). \qquad (196)$$

This model is a good complement to our detailed model of Section 4.4; the detailed model converges slower as $\rho \to 1$ - but as ρ approaches 1 this approximation becomes increasingly accurate.

In more detail, the G|G|1 heavy traffic approximation is only valid [KLEINROCK 1976] when

$$\sqrt{\sigma_A^2+\sigma_B^2} \gg 2(1-\rho)\ \bar{t} \qquad (197)$$

or, when translated into our parameters, equivalently

$$\rho \gg 1 - \sqrt{\frac{1-P}{2}}\ . \qquad (198)$$

Thus the range of validity depends only on the speaker model parameter, P. If $P = \frac{1}{2}$ the condition becomes

$$\rho \gg 1 - \frac{1}{2\sqrt{2}} = .65.$$

4.6.4 Empirical Approximation

Section 4.5.1 indicated that the distribution of delay, as obtained from the detailed model, was very close to an exponential distribution. Section 4.6.3 also implied a similar result for high utilization. It turns out however (see the following Section 4.6.5) that using the mean indicated in Section 4.6.3 does not produce close agreement. To compensate for this deficiency we can revert back to the approximate empirical mean obtained in Section 4.5.3 and then make the assumption of an exponential delay distribution. Let Ω be the assumed exact mean queueing delay. The following empirical approximation can be made:

$$P(W_q \leq t) = 1-e^{-\Omega t}.$$

The following section indicates that the agreement is excellent when the fit was examined for numerous specific cases.

4.6.5 Comparison

We now compare the distributions obtained with the above approximations to the distribution of the detailed model.

Two study cases are employed; one has a utilization of .86, the other has utilization of .95. Recall that the approximation of this section are restricted to the infinite buffer case.

Figures 37 and 38 compares the cumulative distribution obtained via the approximations of Sections 4.6.1, 4.6.2. 4.6.3 and 4.6.4 with the detailed model distribution.

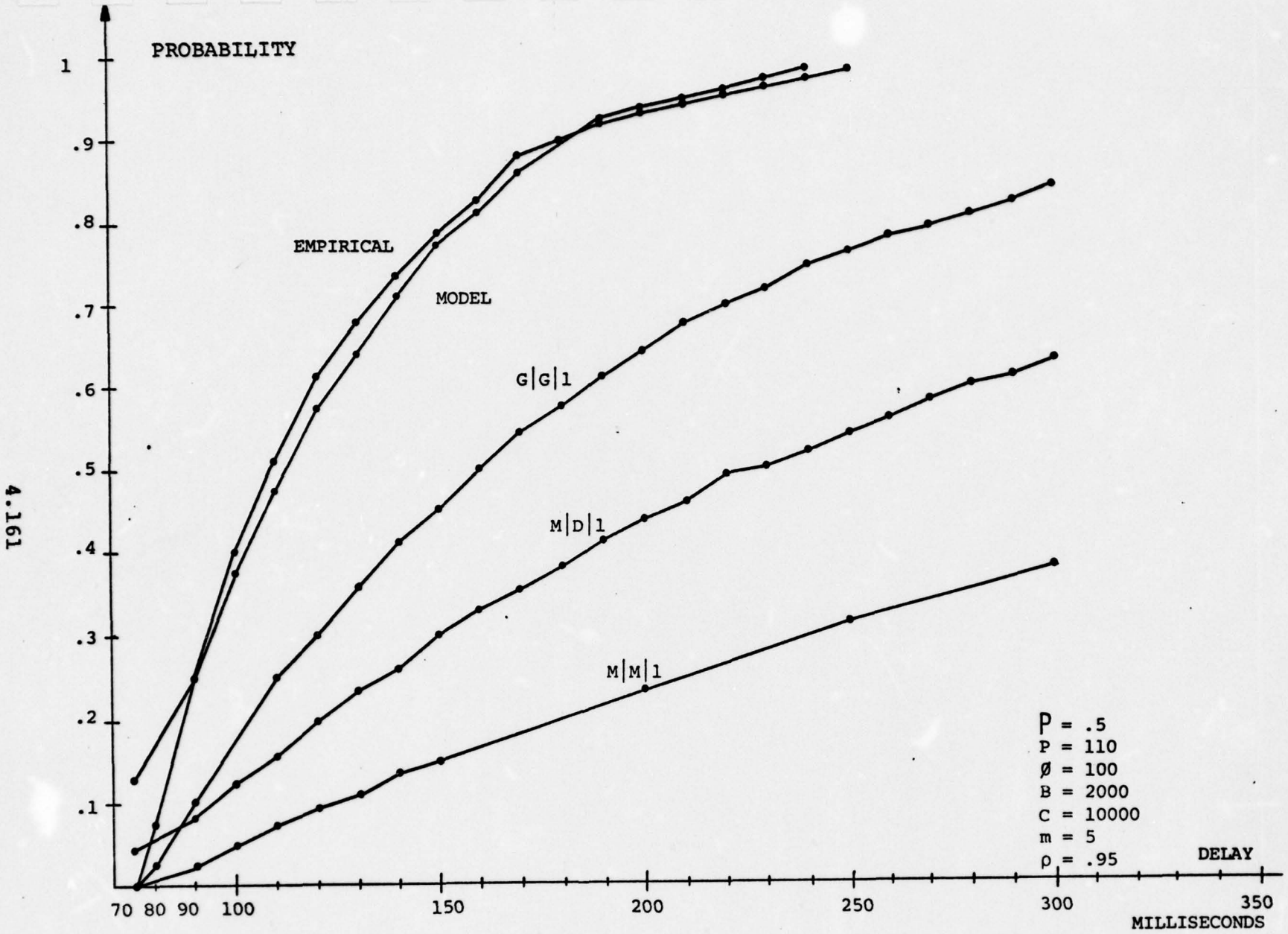
We noted first that, as predicted, the empirical distribution is quite close to the actual distribution; however, as already indicated, use has been made of the exact expected value.

As anticipated the other approximations turn out to be conservative (actually, too conservative) in that these distributions imply a larger mean value and a higher probability tail. Ranking in the order of accuracy is as follows:

1. Detailed

2. Empirical

3. G|G|1 - heavy traffic



4.161

NETWORK ANALYSIS CORPORATION

FIGURE 37: COMPARISON OF APPROXIMATIONS

NOTE: ONLY THE DOTS SHOULD BE COMPARED .

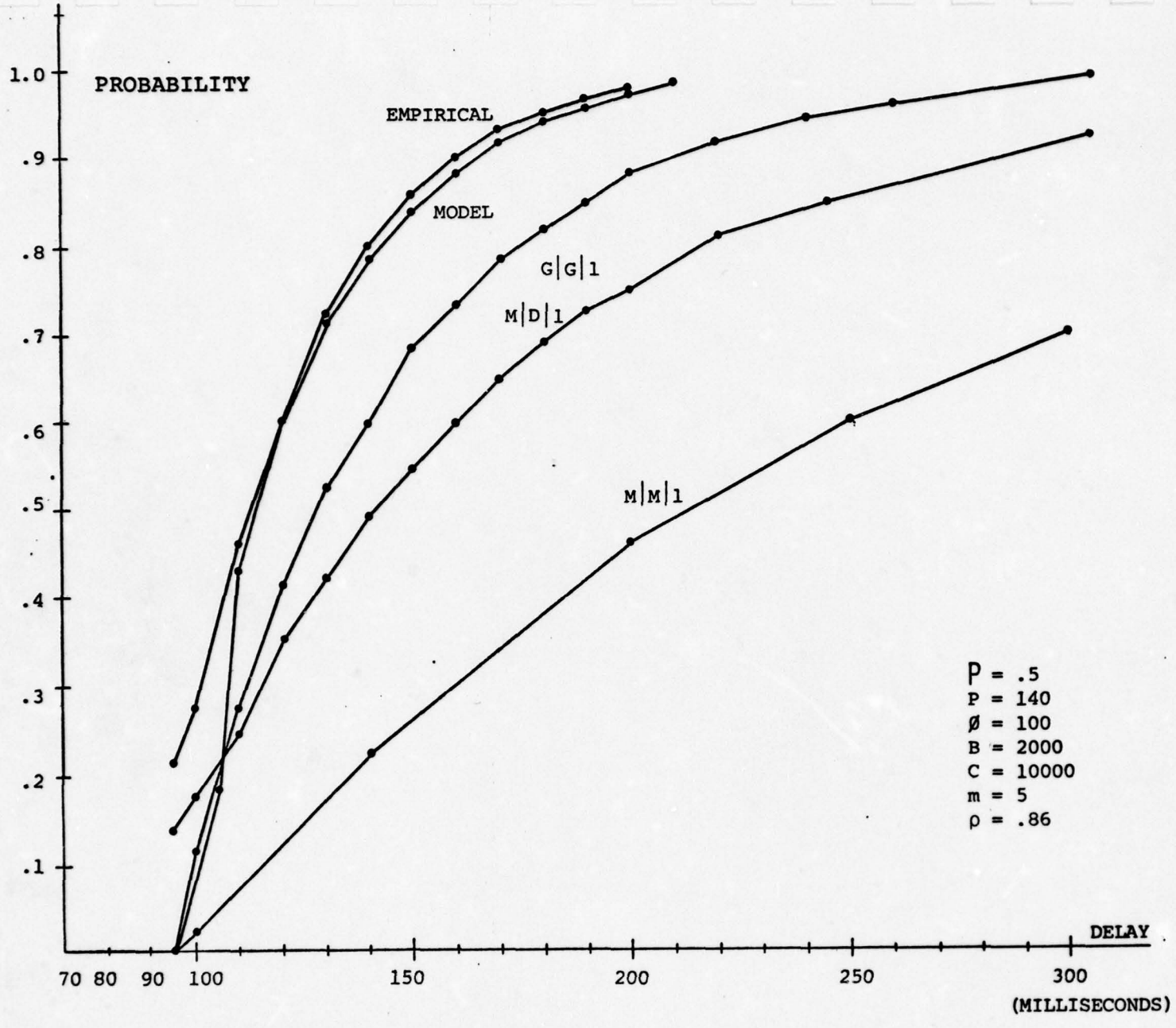


FIGURE 38: COMPARISON OF APPROXIMATIONS

NOTE: ONLY THE DOTS SHOULD BE COMPARED

NETWORK ANALYSIS CORPORATION

4. M|D|1

5. M|M|1

The only surprise is that the M|D|1 approximation, with its exact service process representation, appears inferior to the G|G|1 approximation. Of course, the G|G|1 approximation is specialized to the heavy traffic case and uses two moments of the interarrival distribution.

4.6.6 Approximation Conclusions

The results of Section 4.6.5 show that the detailed model is necessary since the approximations that are independent of it - all except the empirical - are much too conservative and that their overestimate of the performance degradation becomes more serious with high utilization.

The G|G|1 heavy traffic approximation was fairly successful and complements the detailed model; its range of validity, $\rho >> .65$, is the range where the detailed model takes more computation to converge. This approximation uses the exact first and second moments of both the interarrival and service distributions. The M|D|1 approximation, which models the service process exactly is not as accurate. It did, however, provide a closed form expression for optimal packet length which agreed well with results from the detailed model. We plan to derive similar relationships for the G|G|1 approximation and compare the "optimal" packet length results. The M|M|1 approximation is the most analytically tractable but, using only first moment information on the service and interarrivals, turned out to give intolerably poor performance estimates. This particularly highlights the behavior differences and the need for developing special techniques for analysis of packet voice networks, as opposed to packet data networks.

The empirical approximation is accurate and can be described in closed form which is significant if network design techniques ultimately require differentiation or convolution operations.

4.7 TANDEM LINK MODEL

4.7.1 Introduction

In this section we develop a tandem (or series) link model for a network carrying packetized voice traffic. A few (mild) approximations will be made in the analysis. Two approaches can be taken: first, one can feed into the link under study, the more exact output distribution from the previous link in the tandem series (i.e., cyclical packet arrival probabilities); second, one can use instead the approximate output distribution discussed in Section 4.5.2 (i.e., non-cyclical packet arrival probabilities). It is expected that in the steady state the difference would be small. We develop both cases; however, for actual implementation, the second case is easier.

No numerical evaluations have been made yet of the model described here. Future effort will be directed at studying this model, and a general network model, in an attempt to develop design methodology for obtaining cost-effective packet voice networks.

4.7.2 Model Assumptions and Notations

1. We assume the existence of a series of packet switches, PS_i, i=1,2,...M, connected by tandem full duplex transmission lines.

2. Each PS_i has four queues as follows (see Figure 39):

   a. $\vec{Q}_i$: the queue for the line to PS_{i+1}.

   b. $\vec{R}_i$: the queue for incoming local packets from PS_{i-1} destined to a local terminal.

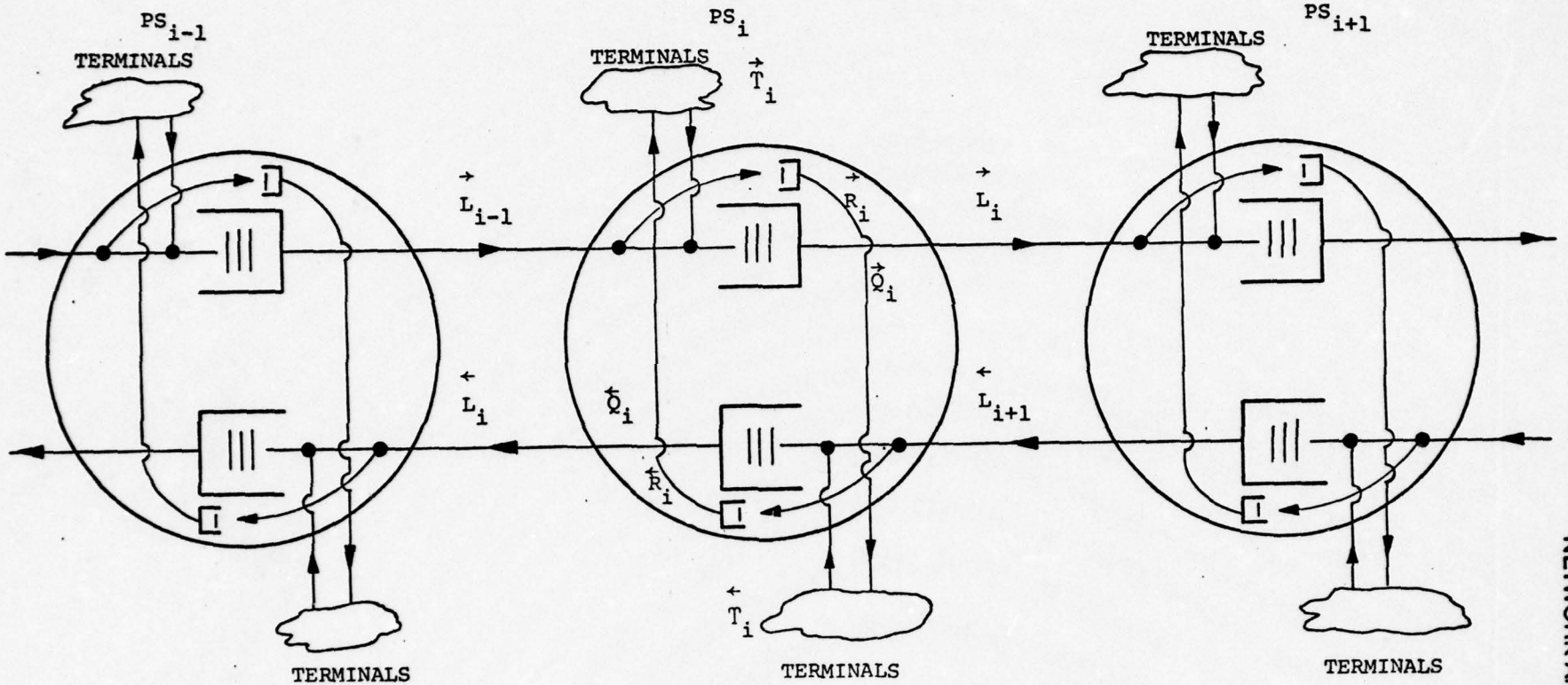


FIGURE 39: FULL DUPLEX TANDEM LINKS

NETWORK ANALYSIS CORPORATION

c. $\overleftarrow{Q}_i$: the queue for the line to PS_{i-1}.

d. $\overleftarrow{R}_i$: the queue for the incoming packets from PS_{i+1}, destined for a local terminal.

3. We again exclude the situation where a terminal connected to PS_i wishes to communicate with another terminal connected to PS_i - this assumption is not strictly required, but is made for simplicity.

4. A terminal $T_{n,i}$ connected to PS_i, is assigned to queue $\overrightarrow{Q}_i$ ($\overleftarrow{Q}_i$) if it wishes to communicate with some at PS_j, $j>i$ ($j<i$).

5. Let $\overrightarrow{T}_i$ be the set of terminals connected to PS_i and assigned to $\overrightarrow{Q}_i$; similarly for $\overleftarrow{T}_i$. Note that the terminals in $\overrightarrow{T}_i$ are independent of terminals $\overrightarrow{T}_j$, $i\neq j$. Similarly, for $\overleftarrow{T}_i$, $\overleftarrow{T}_j$. (This would not be true in case of conferencing). However, in general, the terminals in $\overrightarrow{T}_i$ are not independent of the terminals in $\overleftarrow{T}_j$.

   The above follows since for each speaker-listener pair consisting of terminals $T_{n,i}$ $T_{m,j}$, either $T_{n,i} \in \overrightarrow{T}_i$ and $T_{m,j} \in \overleftarrow{T}_j$ or $T_{n,i} \in \overleftarrow{T}_i$ and $T_{m,j} \in \overrightarrow{T}_j$.

6. The four queues in a packet switch are independent of each other with respect to the arrival stream and with respect to their operation.

7. At each PS_i there is an incoming line from PS_{i-1}, $\overrightarrow{L}_{i-1}$; an outgoing line to PS_{i+1}, $\overrightarrow{L}_i$; an incoming

line from PS_{i+1}, $\overleftarrow{L}_{i+1}$; an outgoing line to PS_{i-1}, $\overleftarrow{L}_i$.

8. Line $\vec{L}_i$ has capacity $\vec{C}_i$; line $\overleftarrow{L}_i$ has capacity $\overleftarrow{C}_i$.

9. Since we will only be concerned with computing the one-way delay from some PS_i to a PS_{i+j}, $j>0$, we can reduce the complexity by considering only the simplex portion of the full duplex system in Figure 39. This simplex, one-way, system is shown in Figure 40.

10. There is a requirement matrix defined as follows. With M as the total number of packet switches, define an MxM matrix V where v_{ij} is the number of terminals homed to PS_i wishing to communicate with a terminal homed at PS_j, $i<j$.

$$V = \begin{pmatrix} 0 & v_{12} & v_{13} & v_{14} & \cdots \\ 0 & 0 & v_{23} & v_{24} & \cdots \\ 0 & 0 & 0 & v_{34} & \cdots \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \end{pmatrix} \qquad (199)$$

We need not be concerned with those terminals wishing to communicate with a PS_j, $i>j$, since these, by assumption do not contribute to the delay on the simplex path being considered.

Clearly $\sum_{j=1}^{M} v_{ij} = \vec{v}_i$. is the number of terminals homed at i, engaged in conversation with

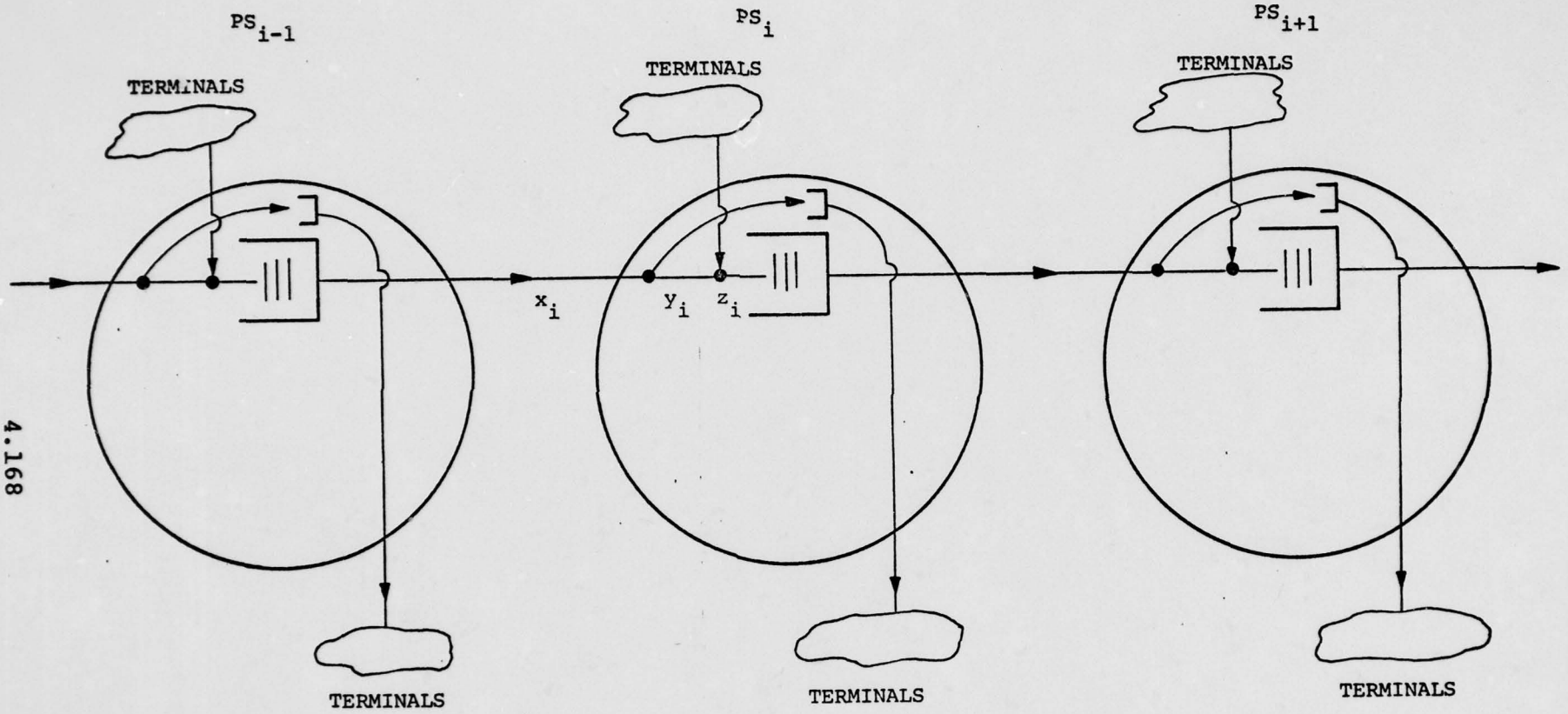


FIGURE 40: SIMPLEX TANDEM LINKS

NETWORK ANALYSIS CORPORATION

a $PS_\ell, \ell>i$ (namely assigned to $\vec{Q}_i$). Also

$\sum_{i=1}^{M} \overleftarrow{v}_{ij} = v_{.j}$ is the number of terminals which send traffic to $\vec{R}_j$.

11. Let X_i, Y_i, Z_i be the cumulative number of terminals which are communicating downstream of the respective points x_i, y_i, z_i of Figure 37.

Then

$$\left.\begin{aligned} X_i &= \sum_{j<i} \vec{v}_{j.} - \sum_{k<\ell<i} v_{k\ell} \\ Y_i &= \sum_{j<i} \vec{v}_{j.} - \sum_{k<\ell\leq i} v_{k\ell} \\ Z_i &= \sum_{j\leq i} \vec{v}_{j.} - \sum_{k<\ell\leq i} v_{k\ell}. \end{aligned}\right\} \quad (200)$$

12. Note that $X_i = Z_{i-1}$ for $i=1,\ldots,M$ (define $Z_0 = 0$).
The following assumption is made: a packet entering PS_i from the line has probability

$$\theta_i = \frac{Y_i}{X_i} \quad (201)$$

of being on route to some PS_j, $j>i$ and probability

$$\eta_i = 1-\theta_i \quad (202)$$

of being a packet destined to a terminal which is in $\overleftarrow{T}_i$.

13. We assume infinite (or large) storage at each $\vec{Q}_i$.

14. All assumptions and notation developed for the single link model (Section 4.4) apply to this case as well.

4.7.3 Methodology

We will be concerned only with the simplex situation in the remainder of this section so we omit the upper arrow notation from the sequel without ambiguity. Also let v_i replace $v_{i.}$ for notational simplicity. For each queue Q_i we compute the distribution of delay δ_i and the output distribution, based solely on the interarrival to Q_i (which is partially formed from the output distribution of Q_{i-1}). Because of the assumption of infinite storage at each switch, we may not need the joint distribution of $\delta_1, \delta_2, \ldots \delta_M$ but can consider each queue separately, and then combine the individual delay distributions. This issue will be addressed in a future report. The emergence time from an incoming line can be considered to have the cyclical phenomenon discussed in Section 4.5.2 - at the cost of increased complexity - or considered to be uniform over time. Both approaches are outlined in Section 4.7.6. We will obtain the end-to-end delay distribution between any pair of packet switches by convoluting or otherwise combining the intervening delay distributions.

In the actual implementation we will sequentially find the individual delay distributions δ_i at queue Q_i, $i = 1,2,\ldots,M$ along with the output distribution to be fed into link i+1; finally combining these distributions to obtain the distribution of end-to-end delay.

4.7.4 Line Utilization

We wish to calculate the line utilization, ρ_i, of L_i in terms of the parameters at the i^{th} switch and the utilization, ρ_{i-1}, of preceding line L_{i-1}. Each of the V_i off-hook terminals at PS_i produce

speech information bits at the rate P_B, which with overhead, leads to a total average bit rate from the local terminals of

$$P_B(1+\frac{\emptyset}{P})v_i.$$

The average bit rate of traffic coming in off L_{i-1} and being passed on to L_i is

$$\rho_{i-1}\, C_{i-1}\, \Theta_i.$$

Combining these two traffic components we get

$$\rho_i = \frac{\rho_{i-1}\, C_{i-1}\, \Theta_i + P_B\left(1+\frac{\emptyset}{P}\right)v_i}{C_i}, \quad i=1,\ldots,M \tag{203}$$

with $\rho_0 \triangleq 0$, $C_0 \triangleq 0$.

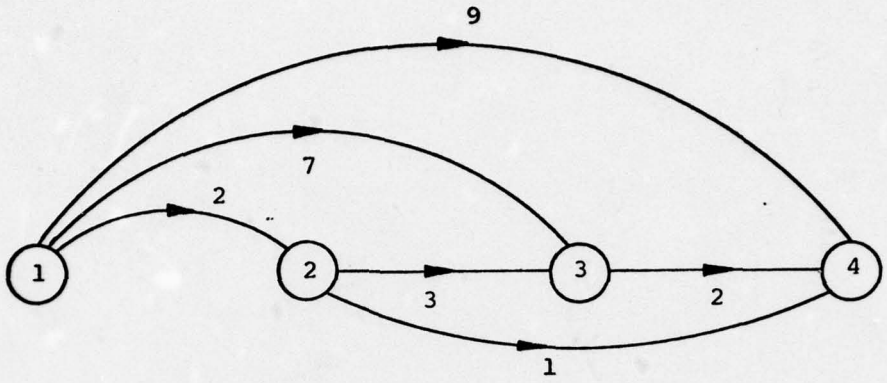
We illustrate these concepts with an example. Let

$M = 4$ $\quad C_1 = 50$ KBS
$P = 1$ KBS $\quad C_2 = 100$ KBS
$B = 5$ KBS $\quad C_3 = 25$ KBS
$\emptyset = 0$
$P = .5$

$$V = \begin{bmatrix} 0 & 2 & 7 & 9 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{aligned} v_1 &= 18 \\ v_2 &= 4 \\ v_3 &= 2 \\ v_4 &= 0. \end{aligned}$$

See Figure 41 for a flow diagram form of V and the calculation of the important quantities. We get

$\rho_1 = .9$
$\rho_2 = .5$
$\rho_3 = 1.2$ (Overload!).



$X_1 = 0$	$X_2 = 18$	$X_3 = 20$	$X_4 = 12$
$Y_1 = 0$	$Y_2 = 16$	$Y_3 = 10$	$Y_4 = 0$
$Z_1 = 18$	$Z_2 = 20$	$Z_3 = 12$	$Z_4 = 0$
$\Theta_1 = 0$	$\Theta_2 = \frac{8}{9}$	$\Theta_3 = \frac{1}{2}$	$\Theta_4 = 0$

FIGURE 41: FLOW GRAPH OF TANDEM LINK TRAFFIC

4.7.5 Queue Operation

4.7.5.1 Locally Originating Traffic

The identical assumptions are made for packets from local terminals joining Q_i as were made for the single link case. Namely,

1. At each frame r, $T_{n,i}$ supplies a packet with probability P (note the worst case assumption).

2. There is self-synchronization in the sense that if $T_{n,i}$ supplies a packet at $r_1 h+\zeta_n$, then any future packet (if any) must be supplied at $(r_1+k)h+\zeta_n$, for some integer k.

3. We impose an assigned arrival scheduling (if necessary, by one-shot initial adjustment) in the sense that the ζ_n are computed as a function of n and V_i by the switch rather than determined by the terminal. The potential packet from $T_{n,i}$ in the r^{th} frame is examined by the switch processor at time $rh+\zeta_n$.

4.7.5.2 Incoming Line

We assume a partition of the frame time width h into subslots of width

$$\mu_{i-1} = \frac{P+\emptyset}{C_{i-1}} . \qquad (204)$$

Let

$$e_i = \frac{h}{\mu_{i-1}} \qquad (205)$$

be the number of such subslots in a frame with e_i assumed integral for convenience. The output stream from a link can be characterized by specifying the probability that a packet is delivered to the store/forward switch during each of these subslots. An arriving packet is delivered by the link to an input buffer, which is processed at the end of each subslot. During a subslot at most one packet can be delivered by the line.

4.7.5.3 Combined Stream

It is operationally infeasible to synchronize the arrival process for packets coming from the line; however, the switch processor can sample the input buffer e_i times during a frame at equally spaced intervals. Packets can then be thought of as arriving (or not arriving) at points $rh+\ell\mu_{i-1}$, $\ell=1,2...e_i$, with appropriate probability, (See Figure 42). We call these points the arrival set for the line.

On the other hand, it is possible to control the arrival time (scheduling) for local packets. One could, for example, force each terminal to supply a packet (if any) exactly at $rh+\frac{h}{2}$ or any other point. It is, however, convenient (in order to reduce the queueing delay) to require that the terminals supply packets at equally spaced out points throughout the frame. More specifically, it will be assumed that $T_{n,i}$ supplies a packet (if any) at $rh+j\frac{h}{v_i}$, $j=1,2,...v_i$. (See Figure 42). We call these points the arrival set for the local terminals.

We then superimpose the two arrival streams to obtain a single stream, whose arrival set is the union at the two arrival sets described above. (See Figure 42). This formulation is equivalent to assuming independent arrival streams for the two packet sources.

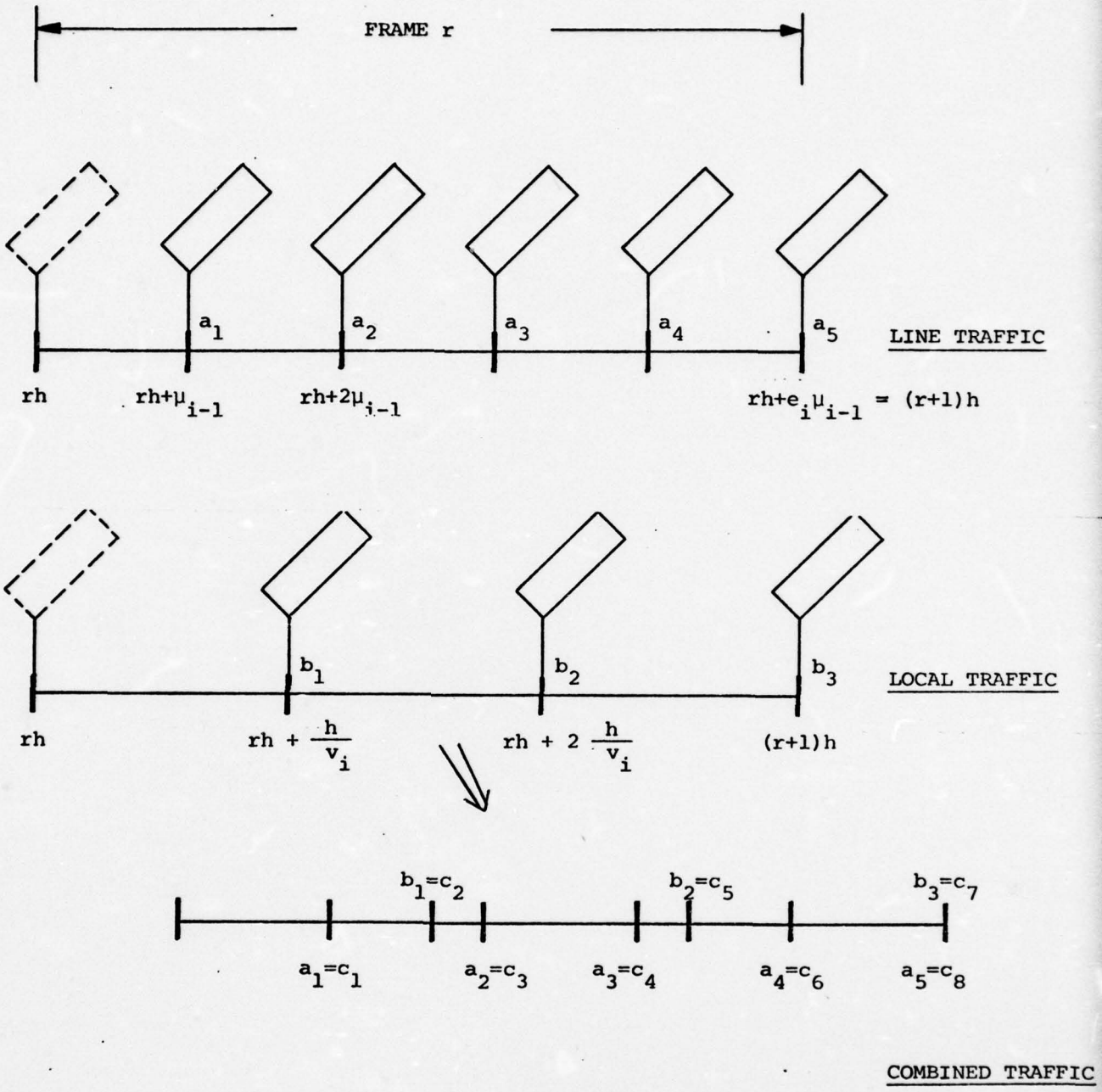


FIGURE 42: PACKET ARRIVALS

4.7.5.4 Combined Interarrival Time

$$\left.\begin{aligned} \text{Let } A^{(r)} &= \{ rh+\ell\mu_{i-1} \mid \ell = 1, 2,\ldots e_i \} \\ B^{(r)} &= \{ rh+j\frac{h}{v_i} \mid j = 1, 2,\ldots v_i \}. \end{aligned}\right\} \quad (206)$$

Let $C^{(r)} = A^{(r)} \cup B^{(r)} = \{C_k^{(r)}\}$ and order the $C_k^{(r)}$ by magnitude; if a tie occurs take the point from $A^{(r)}$ first. There are e_i+v_i points in $C^{(r)}$. This quantity represents the maximum number of packets that can be supplied to the queue Q_i during frame r.

$$\text{Define } \Delta_k = C_k^{(r)} - C_{k-1}^{(r)} \qquad k = 1,2,\ldots,e_i+v_i. \quad (207)$$

where $C_0^{(r)} = C_{e_i+v_i}^{(r-1)}$.

Then Δ_k $k = 1,2,\ldots,e_i+v_i$, is the interarrival intervals of the combined stream of potential packets. This concept will be used below.

4.7.6 Queue Delay Distribution

For expository convenience we introduce a symbolic "k^{th} device" which refers to either a local terminal or the incoming line which potentially supply a non-empty packet at the end of interval Δ_k. As before, let $U_k^{(r)}$ be the remaining delay at time $C_k^{(r)}$, $k=1,2,\ldots,e_i+v_i$. Then the following random variable relationships hold:

$$U_k^{(r)} \underset{k=1,2,..e_i+v_i}{=} \begin{cases} 0 & \text{if } U_{k-1}^{(r)} - \Delta_k \leq 0 \text{ and device k has no packet for } Q_i \\ \mu_i & \text{if } U_{k-1}^{(r)} - \Delta_k \leq 0 \text{ and device k has a packet for } Q_i \\ U_{k-1}^{(r)} - \Delta_k & \text{if } U_{k-1}^{(r)} - \Delta_k > 0 \text{ and device k has no packet for } Q_i \\ U_{k-1}^{(r)} - \Delta_k + \mu_i & \text{if } U_{k-1}^{(r)} - \Delta_k > 0 \text{ and device k has a packet for } Q_i \end{cases} \quad (208)$$

To complete the model we need to obtain the probability that device k supplies a packet. When device k is a terminal this probability is simply P for steady state analysis. If device k is the line whose input buffer is examined at time $rh+\ell\mu_{i-1}$, $1 \leq \ell \leq e_i$, we can follow two approaches:

Alternative 1 - Smooth Line Process

Let p_w be the probability that a non-empty packet is ready in the line input buffer at time $rh+\ell\mu_{i-1}$, $1 \leq \ell \leq e_i$, $r=0,1...$, where $w=\ell+re_i$. Because the Δ-interval and the μ-intervals are not, in general, integer ratios there is a cyclical pattern in the sequence $\{p_w\}$ proportional to the difference frequency $(\frac{1}{\Delta} - \frac{1}{\mu})$ - as observed for the single link (Section 4.5.2). This cycle frequency does not in general correspond to the frame frequency $\frac{1}{h}$ and they may not be simply related. For simplicity we can assume that

$$\sum_{re_i \le w < (r+1)e_i} p_w \approx e_i \rho_{i-1}, \quad \text{for any } r; \tag{209}$$

that is, we assume that each frame period has the same expected number of line arrivals. This approximation is not too radical since by substituting the $\{p_w\}$ cycle length for e_i in the above equation, the relationship would be exact. Going one step further, we assume that the p_w themselves are all equal - specifically

$$p_w = \rho_{i-1} \tag{210}$$

so that each line service time subslot has the same probability of packet arrival.

Alternative 2: - Cyclical Line Process

The cyclical output distribution can be fed in. This introduces a certain amount of bookkeeping complexity. It is expected that in the steady state the divergence from Alternative 1 is minimal.

Finally we make the important assumption of the independence of $U_{k-1}^{(r)}$ and the behavior of device k; we can then easily compute the joint probabilities of the events required for computation of $U_k^{(r)}$.

4.7.7 Approximations

Because of the complexity involved in solving the exact model, one may seek applicable standard approximations from the literature. The usual simplification requires assumption of a Poisson process for the input or an exponential service time distribution or both. We have already demonstrated for the single link case that these approximations are rather weak - too conservative; however, their relative tractability may produce other analytical benefits.

4.7.7.1 Chain of M|M|1 Queues

An analysis can be adoped from [GROSS, 1974]. The output process of an M|M|1 queue with infinite buffer capacity can be shown to have the same distribution as the input process. Thus, with infinite buffers, for a series of queues with Poisson input and exponential service at each server, each queue operates as an M|M|1 queue. A few models with finite storage and resultant blocking are considered in [GROSS, 1974]. It is noted, however, that if the number of buffers are reasonably large, the effect of buffer overflow or blocking is small.

4.7.7.2 Chain of M|D|1 Queues

Several works by [RUBIN,1974a], [RUBIN, 1974b] address the chain of queues problem. The models considered are representative of data networks rather than voice networks. Distributions of individual channel waiting times and analysis of random message lengths are investigated. The basic result is the following:

> THEOREM: (Capacity ordering invariance property). The overall waiting time over an N-channel path with capacities $(C_1, C_2, \ldots C_N)$ is the same as that over an N-channel path with capacities $(C_{i_1}, C_{i_2}, \ldots, C_{i_N})$, where the latter sequence is an arbitrary ordering of $(C_1, \ldots, C_N)$. The overall waiting time depends only on the minimal capacity, $\min(C_1, \ldots, C_N)$.

4.7.8 Extension to Network of Queues

The same techniques and approximations employed in the tandem link model can be applied to a network of queues. However, some restrictions may need to be imposed (e.g., fixed routing). With the possibility of several incoming and outgoing links, a complex bookkeeping procedure arises.

We hope to report at a later date, results obtained with such an approach to analyzing a network of queues carrying packet voice traffic.

4.8 CONCLUSION

A methodology is needed to be able to design packet voice networks. Such networks differ from the packet data case in the following fundamental respects:

1. Regularity of input traffic.

2. More complex performance criteria (smoothness, error tolerant).

3. Performance criteria will be imposed on a worst-case end-to-end basis.

4. Different set of applicable protocols.

We have modeled a single link situation and implemented the model with a computer program. The link model can accept a wide variety of speech models and system parameters and yields the complete steady state or transient distribution of delay. Section 4.5 presented the results of a large number of studies on this model and Section 4.6 compared some standard approximations to the model.

Some of the important facts learned from the studies are:

1. Standard approximations are overly conservative in that they predict poorer performance than can be actually attained.

2. The single link delay distribution is approximately exponential.

3. Percentile delay performance criteria track very well with equivalent performance criteria placed on the mean delay.

4. A closed form experssion for the mean delay was obtained by numerical fit.

5. A closed form expression for optimal packet length, obtained from an approximation, agreed very closely with results from our detailed model.

6. Only a small number of buffers was found necessary to sustain adequate performance. A small number of buffers also reduced the transient excursions and durations.

The single link model was incorporated into the tandem link model described in Section 4.7. Implementation and study of this model will give results for the situation when a link must service an incoming line as well as local terminals. It will also be a testbed for methods of combining delay distributions for more than one link to get end-to-end results. The tandem link situation is considerably simpler than the general case because of simple topology and lack of routing alternatives.

Based on the results already obtained a foreseeable methodology for general network design for packet voice can be outlined as follows:

1. Interactively set topology, link parameters, and system parameters.

2. Using closed form analytical approximations for link delay behavior, obtain an optimal routing pattern.

3. Evaluate performance results using the detailed link model.

4. Return to Step 1 until designer is satisfied that the network is the cheapest one that satisfies all the design criteria.

REFERENCES

[BRADY, 1967] Brady, P. T., "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., 47, No. 1 (January 1968), pp. 73-91.

[BRADY, 1969] Brady, P. T., "A Model for Generating On-Off Patterns in Two-Way Conversations," B.S.T.J., 48, No. 7 (September 1969), pp. 2445-2472.

[BRADY, 1971] Brady, P. T., "Effects of Transmission Delay on Conversational Behavior on Echo-Free Telephone Circuits," B.S.T.J., 51, No. 1 (January 1971).

[COHEN, 1976] Cohen, D., "Issues in Transnet Packetized Voice Communication," NSC Note No. 85

[COOPER, 1972] Cooper, R. B., Introduction to Queueing Theory, The MacMillan Company, 1972.

[COVIELLO, 1976] Coviello, G. J., "System Implications of Packetized Voice," ACM-NBS Symposium, June 1976.

[DAILEY, 1968] Daley, D.J., "The Correlation Structure of the Output Process of Some Single Server Queueing Systems," Ann. Math. Statist., 39, (1968) pp. 1007-1019.

[EMLING, 1963] Emling, J. W., "The Effects of Time Delay and Echoes on Telephone Conversations," B.S.T.J. (1963).

REFERENCES (Cont'd)

[FINCH, 1959] Finch, P. D., "The Output Process of the Queueing System M/G/1., J.R. Statist. Soc B, 21, (1959), pp. 375-380.

[FORGIE, 1975] Forgie, J. W., "Speech Transmission in Packet-Switched Store-and-Forward/Networks," Proc. of NCC (1975)

[FORGIE, 1976] Forgie, J. W., "Some Comments on NSC Note No. 78," NSC Note No. 82, (1976)

[GROSS, 1974] Gross, D., Harris, Fundamentals of Queueing Theory, John Wiley, 1974.

[HUGGINS, 1976] Huggins, A. W. F., "Effect of Lost Packets on Speech Intelligibility," NSC Note No. 78, (1976).

[JAFFE, 1964] Jaffe, J., "Markovian Model of Time Patterns of Speech," SCIENCE, 144 (May 15, 1964), pp. 884-886.

[JENKINS, 1966] Jenkins, J. H., "On the Correlation Structure of the Departure Process of the $M/E_n/1$ Queue," J. R. Statist, Soc., B, 28, (1966), pp. 336-344.

[KING, 1971] King, R. A., "The Covariance Structure of the Departure Process from M/G/1 Queues with Finite Waiting Lines,"

REFERENCES (Cont'd)

[KLEMMER, 1967] Klemmer, E. T., "Subjective Evaluation of Transmission Delay in Telephone Conversations," B.S.T.J., (July-August 1967), p. 1141-1147.

[KLEINROCK, 1976] Kleinrock, L., Queueing Systems Volume II: Computer Applications, Wily-Interscience, 1976.

[KRAUSS, 1967] Krauss, R. M., "Effects of Transmission Delay and Access Delay on the Efficiency of Verbal Communications," J. of the Acoustical Society of America, 41, 2 (1967), pp. 286-292.

[NORWINE, 1938] Norwine, A. C., "Characteristic Time Intervals in Telephone Conversation," B.S.T.J., 17, (April 1938), pp. 281-331.

[NSC, 1976] National Speech Conference, Lincoln Laboratories (August 1976).

[RUBIN, 1974a] Rubin, I., "Tandem Queues with Constant Channel Service Times and Group Arrivals," University of California, Los Angeles Tech. Rep., UCLA-ENG-7417, March 1974.

[RUBIN, 1974b] Rubin, I., "Communication Networks: Message Path Delays," IEEE Trans. on Information Theory, Vol. IT-20, No. 6, (November, 1974), pp. 738-745.

REFERENCES (Cont'd)

[SAATY, 1961] Saaty, T. L., Elements of Queueing Theory with Applications, McGraw-Hill, 1961.

[SIEMENS, 1974] Siemens Co., Telephone Traffic Theory Table and Charts Part I, Siemens Aktiengesellschaft, 1974.

[SCHMOOKLER, 1970] Schmookler, M. S., "Limited Capacity Discrete Time Queues with Single or Bulk Arrival," Princeton University, Princeton, New Jersey, June 18, 1970.

CHAPTER 5

A CIRCUIT SWITCH NODE MODEL

CHAPTER 5

TABLE OF CONTENTS

		PAGE
5.1	INTRODUCTION	5.1
5.1.1	Global Objectives	5.1
5.1.2	Performance Measures and Parameters	5.3
5.1.3	Methodology	5.7
5.2	GENERIC CIRCUIT SWITCH MODEL DESCRIPTION	5.9
5.2.1	Switch Architecture	5.9
5.2.2	Switch Operation/Typical Call Flow	5.15
5.2.3	The Control Element	5.19
5.2.4	The Connection Element	5.28
5.2.5	Assumptions/Limitations	5.33
5.3	SWITCH OPERATION IN THE NETWORK ENVIRONMENT	5.35
5.3.1	Routing/Control Strategies	5.35
5.3.2	Signaling Techniques	5.51
5.4	CONCLUSIONS	5.58
	REFERENCES	5.59

CHAPTER 5

TABLE OF CONTENTS: FIGURES

PAGE

FIGURE 1: GENERIC CIRCUIT SWITCH ARCHITECTURE..... 5.2

FIGURE 2: CALL PROCESSING SCENARIO..... 5.4

FIGURE 3: SWITCH ARCHITECTURAL EVOLUTION..... 5.10

FIGURE 4: CALL ESTABLISHMENT FLOW THROUGH GENERIC CIRCUIT SWITCH NODE..... 5.16

FIGURE 5: QUEUEING MODEL USED TO ANALYZE DIAL-TONE DELAY..... 5.23

FIGURE 6: QUEUEING MODEL USED TO ANALYZE CONNECTION DELAY..... 5.25

FIGURE 7: CONNECTION ELEMENT TOPOLOGY..... 5.30

FIGURE 8: ROUTE COMPARISON BETWEEN ORGINATING-OFFICE CONTROL AND PROGRESSIVE CONTROL STRATEGIES... 5.36

FIGURE 9: OPERATIONAL COMPARISON BETWEEN OOC AND PRC... 5.38

FIGURE 10: ARCHITECTURAL COMPARISON BETWEEN SIGNALING TECHNIQUES..... 5.52

FIGURE 11: VARIATIONS OF THE SIGNALING TECHNIQUES..... 5.54

CHAPTER 5

A CIRCUIT SWITCH NODE MODEL

5.1 INTRODUCTION

The generic circuit switch architecture which forms the basis of our analytic investigation is shown in Figure 1. As depicted, the model is fairly general and can incroporate several different types of switching machines.

5.1.1 Global Objectives

There are several objectives of the circuit switch modeling effort; the primary underlying goal has been to quantify the internal blocking and delay caused by the switch itself and to assess circuit switch performance (in isolation) as well as in a network. Specifically, we desire the:

- Ability to determine circuit switch performance as a function of relevant architectural parameters.
- Ability to quantitatively assess the impact of a particular switch architecture/operation on overall network performance.

Of necessity, the realization of the preceding two goals also requires a technology assessment and survey of existing circuit switches. Circuit Switch technology assessment is not the subject of this chapter.

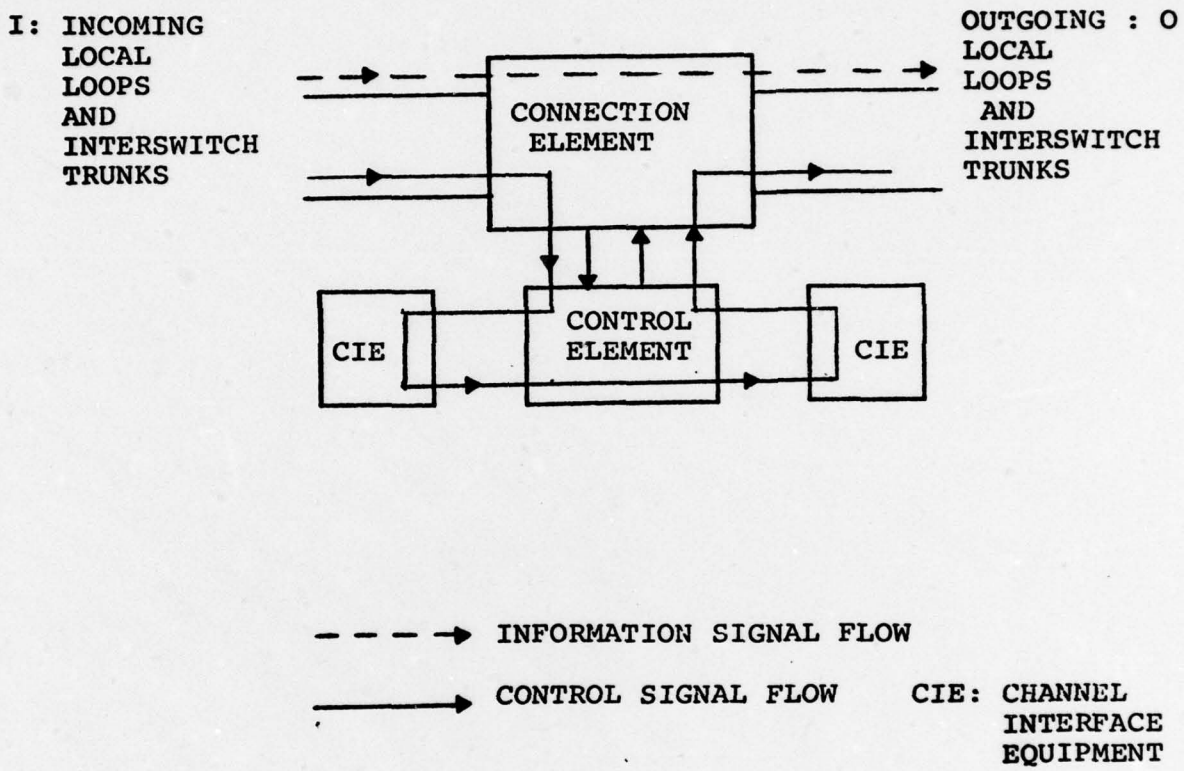


FIGURE 1: GENERIC CIRCUIT SWITCH ARCHITECTURE

It should be emphasized that we are not explicitly advocating or designing a particular switch architecture, although the performance merits of a given approach should become apparent as a result of the use of the analytic models. Finally, the switch model will implicitly provide the capability to evaluate the cost/effectiveness of switch modifications to accommodate new functional capabilities.

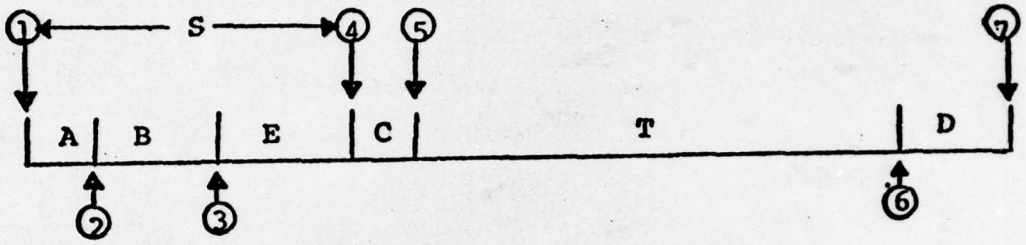
5.1.2 Performance Measures and Parameters

The switch performance measures to be modeled are now discussed relative to the diagram depicted in Figure 2. At the outset, it is imperative to distinguish between those parameters determined by the traffic or subscriber activity and those parameters which are a function of the switch/network architecture and operation.

As shown, the processing of a typical call by a circuit switch results in several delay components. The delays due exclusively to the circuit switch and network mode of operation are:

Dial-Tone Delay: The interval between a caller's indication of intent to transmit and the switch's acknowledgement of such intent. In the voice telephone network, the acknowledgement assumes the form of a dial-tone; in a data network, a clear-to-send message may be returned to the originating terminal.

Connection Delay: The interval between the completion of the entry of the destination address by the originating subscriber and the establishment of the end-to-end path. A given path may consist of several



EVENTS:

1. Phone goes "off-hook"
2. Subscriber receives Dial-Tone
3. Destination address specified
4. Ringing Signal applied
5. Called Party answers (Connection Established)
6. Either party hangs-up
7. Connection broken

TEMPORAL INTERVALS:

A - Dial Tone Delay
B - Dial-in Interval
E - Establish end-to-end connection
C - Subscriber Attention Interval
T - Transaction Transmission Time
D - Disconnection Interval
S - Setup Delay = A + B + E

FIGURE 2: CALL PROCESSING SCENARIO

tandem connections through intermediate switches; hence, the total connection delay can be formally subdivided into the cross-office connection delay (delay through an individual switch) and the cross-network connection delay (end-to-end delay). In telephony, the connection delay is usually defined as the interval that separates the dialing of the last digit and the detection of ringing tone.

Disconnection Delay: The interval between either subscriber hanging up and the complete breakdown of the previously established end-to-end path. Although not directly perceived by the subscribers departing from the system, the disconnection delay is important since if the path requires an excessive period of time to be broken, network resources can be denied to subsequently arriving calls in the interim.

In addition to the above delay components, several others exist, which are primarily a function of the incoming traffic and external subscriber activity. For instance:

Dial-in Interval: The time required to specify the destination address information. The duration of this interval is determined by several factors: the subscriber behavior (interdigit pauses), speed of the signaling equipment (dial-pulse, DTMF, ACU) and capacity of the signaling channels (in-band, dedicated, etc.).

Transmission Time: The time required to deliver the information; determined for voice by the conversation duration, for data by the message length and information channel capacity.

Subscriber Attention Interval: The interval separating the initial application of ringing tone and the called subscriber's response (phone goes off-hook) so that the connection is established. This interval is, of course, highly subscriber dependent and is assumed to be fixed in our analysis.

The duration of the switch-based delays will vary considerably, depending on the actual switch operation, architecture, component technology, and network protocols. As discussed shortly, only a selective subset of the vast number of possible switch configurations will be incorporated into the analytic model.

5.1.2.2 Blocking Components

The establishment of the end-to-end call transmission path from origin to destination assumes the availability of a free path. Otherwise the call setup procedure is unable to secure an outbound connection at the origin or any tandem switch, the call is said to be lost. With respect to the switch, there are two types of blocking which may be encountered:

External Blocking: This is a network blocking component resulting from trunk unavailability. Upon determination of a trunk unavailability condition, the switch will return a "busy" signal to the calling subscriber.

Internal Switch Blocking or "Matching" Loss: Path unavailability caused by the unique topological structure of the connection element (Figure 1), i.e., the current set of input-output connections (I-O) in progress, prohibits a desired I-O connection from being established.

A variety of connection elements, which possess varying degrees of internal blocking exist [LEE, 1955]. For a small number of attached subscribers, so called "non-blocking" connection elements (which exhibit no internal blocking) prove cost-effective. More advanced switch technologies (such as time-division based switching) have reduced the levels of internal blocking frequently encountered in the connection element. The switch operation can also be altered to reduce the impact of internal blocking, at the expense of increased setup delay, by initiating several attempts to obtain a successful connection. This is a very favorable tradeoff provided the switch control element (path determination/establishment hardware or software) is sufficiently fast, and is often used to implement capabilities such as alternate routing, automatic retrial and camp-on. The end-to-end loss probability is composed of several internal and external components and depends on the number of switches which comprise the source-destination path and the network routing plan.

5.1.3 Methodology

The analytic model development will proceed according to the following natural decomposition: connection element architecture and control element architecture/operation. The switch delay

performance components are determined by the control element's operation, whereas the internal blocking components are a function of the connection element structure. A simplified queueing-based performance model is derived in the next section which attempts to account for relevant switch, subscriber and traffic statistics. Existing deficiencies in the current version of the model are emphasized. Furthermore, no attempt is made to account for detailed aspects of a particular switch's operation. The class of switch architectures that are adequately characterized (for our ultimate purpose) by the model will also be tangentially discussed; finally, modifications or additions to the basic model, which are required in order to expand the class of switches whose performance can be quantified, are outlined.

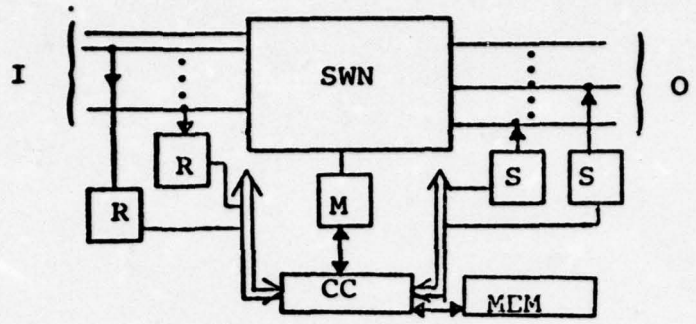
5.2 GENERIC CIRCUIT SWITCH MODEL DESCRIPTION

We now describe the individual components and the generic circuit switch which are modeled and form the basis of the network performance evaluation. The sequence of operations and switch resources used to set up a typical call are also discussed.

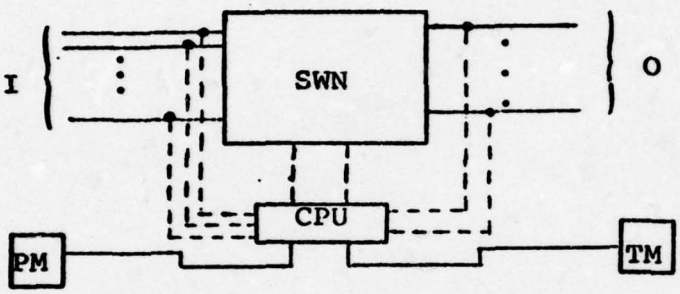
As implied by the event transition diagram of Figure 2, the tasks required to set up a call under a circuit switching discipline are identical regardless of the particular switch architecture employed. However, differences in system operating efficiency/speed arise depending on the type of switch configuration which executes these tasks. For example, many processes can be handled either by hardware or software. Certain system resources can be dedicated on a per line basis or shared among several incoming calls. The switch CPU task scheduling mechanism may be cyclical, or event-driven. In the analysis which follows, we restrict the range of configuration possibilities by defining a specific set of resources and outlining the manner in which they are controlled. Since we are presently concerned with overall system behavior and performance, the current methodology will suffice. The model would have to be considerably refined in order to account for minute facets of a specific hardware and software architecture.

5.2.1 Switch Architecture

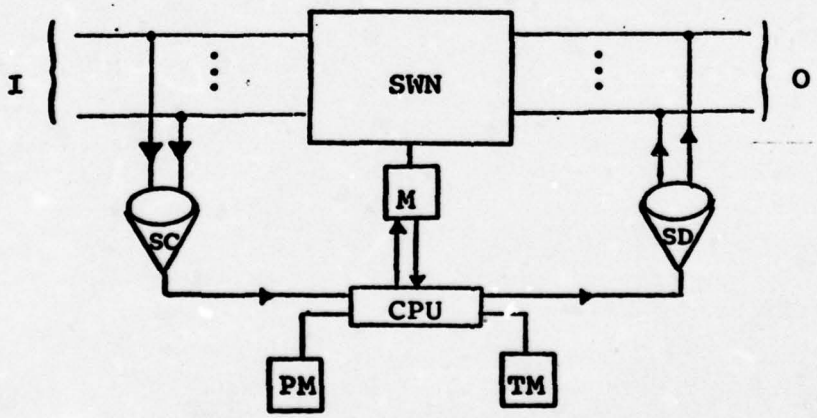
Consider the simplified diagram of Figure 3 which depicts several categories of switch architecture employing partial or full use of electronic components (in either the connection or control element).



(a) DISTRIBUTED



(b) CENTRALIZED



(c) COMBINED

FIGURE 3: SWITCH ARCHITECTURAL EVOLUTION

LEGEND:

CC	- Common Control
CPU	- Central Processing Unit
M	- Marker
PM	- Permanent Memory (Program Store)
TM	- Temporary Memory (Call Store
MEM	- Memory
R	- Call Register
S	- Call Sender
SC	- Scanner
SD	- Signal Distribution
SWN	- Switching Netwo

5.2.1.1 Component Parts

It is evident that there exist a number of specialized units which are integral to the switch operation depending on the distribution of function and intelligence in the overall architecture.

The primary units are:

Marker: The marker's function is the physical establishment of a path through the switch connection element (SWN). The marker can vary in complexity from a voltage actuating mechanism for "setting" relays to a small special purpose computer depending on the type of switch architecture in which it is used. The marker typically is required to perform busy/idle trunk tests, path continuity checking and input-output connection setup. In addition, the marker is often needed to establish a path between incoming local loops and trunks to other specialized equipment such as call registers, senders and tone generators. Occasionally, if the size of the internal switching network is large, several markers may be required in order to provide an adequate speed of service, although the complex scheduling/coordination of multiple marker operation by the common control unit is usually too complex to support this capability. Finally, the marker occasionally performs the number translation functions used to route calls.

Register: A buffer used to store dialed digits generated by an attached local subscriber or signaling address information forwarded by a previous switch during a tandem connection. Depending on the component technology, registers may be hardwired devices or storage locations in a computer memory module.

Senders: A buffer/transmitter used to distribute signaling information (destination address) to tandem or destination switches in an end-to-end connection. In a certain sense, the sender performs the complementary function of the register. Several types of senders may have to be used if a diversity of signaling methods exist in the network (e.g., dial pulse, tones, etc.). Many systems employ devices which serve a dual purpose that both receive and transmit address information, which are called "register-senders."

Once the address information is received, translation of the destination number is often required to conform with a certain routing plan. This can be executed by a dedicated device known as the "translator," under computer control via table look up, or by hybrid devices which combine a subset of the previous functions with translation; the latter are referred to as "register-translators," "sender-translators" or "register-sender-translator."

In every electronic circuit switch, there exists a unit which controls and coordinates the operation and interaction of all peripheral equipment, as well as exercising supervisory/test functions. Referred to as the "common-control" unit, this can range in sophistication from a fully hardwired logic device to a stored program control computer. If a computer (CPU) is employed as the common control unit, then additional memory is required to store the operational program (PM) as well as the status of calls in progress and translation information (TM). Scanning units (SC) and signal distribution equipment (SD) are used to offload the common control CPU's processing burden in modern circuit switches, by performing the repetitive tasks of line scanning, digit collection and assembly, signal generation, etc.

5.2.1.2 Evolution

For the purpose of perspective, we briefly summarize the various forms, switches employing common control have assumed over the last twenty-five years (see Figure 3). A primary distinction between various switch architectures has been the level of intelligence placed in the peripheral equipment. In all cases depicted, there exists a functional unit (common-control or CPU) which bears responsibility for the proper execution of call processing tasks. The individual tasks can however be executed using hardware, software, or a combined technique. In the earliest circuit switches using common control techniques (Figure 3a), the cost of a high capacity memory was prohibitive, hence a small number of shared devices known as registers were used to store incoming address information. Still, due to the high register cost, their number was limited and access to them was controlled through the switching network. Dedicated signal interface units were also required to attach local loops to the switch once a subscriber goes "off-hook," an attempt to seize a register will be initiated by the control unit. This type of architecture is herein dubbed "distributed" due to the existence of several devices performing individual well-defined functions. An example of this type of switch is the Crossbar No. 5.

As computers matured in capability and decreased in cost, switch designers argued that all common control tasks could be adequately performed in software. The computer and accompanying memory units were envisioned to replace the common control unit and its associated peripheral equipment. The extreme form of centralization, shown in (Figure 3b), was also an economic necessity since the CPU at the inception of stored program control was expensive, and its cost-effective inclusion in any switch configuration could only be justified by replacing much of the earlier dedicated peripheral equipment. Clearly, a centralized architecture cannot function

totally independent of any peripheral equipment, however, the presence of such equipment is minimal compared to earlier configurations. Even as the CPU cost declined, the majority of tasks were still relegated to the computer, since designers felt that the switching functions could be performed entirely in software.

As a result of the removal of dedicated units (such as registers, senders, etc.), the operational burden on the CPU increased; the control unit (CPU) was now required to perform the comparatively trivial tasks of digit collection, outpulsing, etc., in addition to coordination of the system work flow, call and route control, etc. This required a complex real-time operating system which often supported multiprogramming in order to provide timely service. One operational innovation that emerged as an outgrowth of CPU control, was the "scanning" of input lines for the "off-hook" condition. This in contrast to the earlier event-driven mode of operation, enabled the CPU to more effectively schedule and accept work. The flexibility needed in order to provide a greater variety of subscriber services, such as abbreviated dialing, call forwarding, etc., can only be economically realized via computer control, but the operational overhead which accompanies a fully centralized implementation limits the realization of such capabilities.

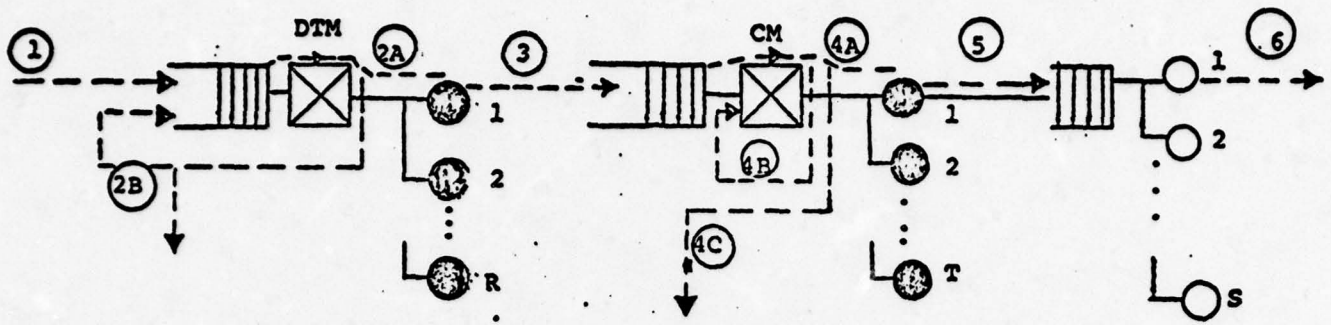
Due to the preceding deficiencies, dedicated special-purpose units have reappeared in the circuit switch architecture such as scanners and signal processing hardware (Figure 3c). With the increased prevalence and decreased cost of microprocessors and LSI circuitry, an extreme distribution of intelligence can be expected to emerge in new switch designs. Such previously flagrant hardware "waste" as dedicated registers per input line may become a cost-effective reality in the future. In short, the global switch architecture may come "full cycle" from distributed to centralized to distributed.

The switching network (SWN) or connection element has changed less than other components. This device has used a variety of crosspoint elements to physically connect inputs to outputs (I/O), and is usually configured in a multi-stage matrix or grid arrangement. Under so-called space-division switching, crossbar relays, reed relays, gas tubes and solid state devices (e.g., thyristors, etc.) have assumed the role of crosspoints. Although solid state crosspoints switch inherently faster than their electromechanical counterparts, their use has been limited due to severe problems associated with crosstalk, high attenuation and electrical compatibility with the signaling and test equipment. Under time-division switching, core memory, high speed electronic memory as well as digital circuitry (shift registers) have been successfully used as "crosspoints" in many systems.

5.2.2 Switch Operation/Typical Call Flow

In the previous sections, we have described various types of switch architecture. Because of the large number of interrelated processes (devices) which are deployed in the establishment of a single call, any analytic attempt to model the switch operation must of necessity be approximate. We now describe a generic switch architecture and call processing mode of operation, which will enable us to conduct a quantitative evaluation of performance. Deficiencies, which exist in the model, will be outlined subsequently.

The processing of a typical call is discussed in reference to the diagram of Figure 4. It should be emphasized that the elements depicted in the generic architecture may represent physical devices such as a register, or a corresponding subroutine used in a stored-program realization, such as the digit collection and analysis program module.



LEGEND:

COMPONENTS

DTM - Dial-Tone Marker
CM - Completion Marker
R - Dial Register/Receiver
S - Signalling Transmitter/Sender
CU - Control Unit (not shown possibly a CPU)
T - Trunks

NOTE:

R, S, T as shown in the above diagram also denote the number of registers, senders and trunks respectively.

CALL FLOW

1. Incoming calls requiring path establishment (both tandem and originating) arrive to the system.
2A. Calls receive a dial register (dial tone returned). Registers operate on a delay or loss basis.
2B. Marker cannot find path to dial register (call either flushed from system or returned to queue and awaits path availability).
3. Dialing-in completed, or signal address received; (call awaits path establishment by the completion marker).
4A. Path established; call attached to trunk.
4B. Path internally blocked, call either flushed from system, tries another path to idle trunk or chooses another trunk.
4C. Trunk group busy, call blocked.
5. Queue for access to signaling equipment (sender), if necessary.
6. Signaling information transmitted to next switch, or ringing applied to destination subscriber.

FIGURE 4: CALL ESTABLISHMENT FLOW THROUGH GENERIC CIRCUIT SWITCH NODE

NETWORK ANALYSIS CORPORATION

For convenience, separate markers are assumed to exist which provide access to the signaling equipment; register access is controlled by the dial-tone marker and trunk access is controlled by the completing marker. Incoming calls, including those originated by the subscribers or due to tandem connections, are assumed to be placed in a queue (software or hardware), where they await the availability of a register (R) so that the switch can receive address information. Registers are shared by both tandem and originating calls. If a dedicated register for each incoming line existed, no marker (DTM) would be required to establish access to the dial registers. Once the marker becomes available, it attempts to seize a register for the requesting subscriber (set up a path to a register, reserve a time slot, etc.). If no register is currently free, the call may be returned to the queue or could be flushed from the system. If an idle register exists, but no access path to it can be setup by the marker, several possible actions can be taken: the call request can be reattempted using a different access path to the idle register; a second idle register might be chosen and an associated access path established; the call could be returned to the request queue for a later attempt; or the call is discarded from the system. Once an idle register is seized, dial tone is returned to the originating subscriber who then enters the destination address; for a tandem connection, the previous switch in the end-to-end path will forward the relevant portion of the destination address. After seizing the register, the marker releases from the processing of the particular call and again becomes free to service other calls.

Once the destination address has been received, the switch attempts to establish a connection to a particular outbound trunk chosen in accordance to a certain routing plan. If the destination subscriber is attached to the same switch as the originating subscriber, then a connection to the appropriate local loop is attempted; if the called subscriber is attached to a different switch, route control and possibly address translation functions are invoked

by the switch control unit to determine a trunk over which the signaling information should be transmitted. Depending on the switch operation, the dial register contents may be transferred to an interim buffer during the path setup phase, so that the register can be released for use by additional calls. Alternatively, the register can be held while the switch is attempting to establish a connection. It should be noted that the latter strategy exhibits increased performance degradation during heavy loads since, if the register is held during the connection phase, an increase in the connection delay will directly influence the length of the dial-tone delay. The former technique (use of an interim buffer when there exists a scarcity of registers), serves to "decouple" the two delays. However, the use of an interim buffer may be less desirable, if registers transmit as well as receive the destination address (register-senders), since the time required by the control unit to perform the buffer transfers could be longer than the connection delay.

Upon reception of the destination address (and possible release of the call register), the call request queues for access to the completing marker (CM). The completing marker performs several functions. Based on the destination address and after suitable address translation (which could be executed apart from the marker), the marker chooses an available outbound trunk to the next switch determined by the routing plan, and attempts to establish a path through the connection element to it. If no idle trunk exists, the call is blocked unless alternate routing is provided, in which case, a connection reattempt using a different idle trunk is performed. If no path through the SWN exists to the idle trunk, a second path may be attempted. It is assumed that the switch memory stores the status of the trunk group so that determination of an idle trunk is rapid, and no hunting is required as in a step-by-step switching system. After seizure of an idle trunk, the switch senders transmit the appropriate signaling information to the next switch along the path; if the connection is intra-switch, a tone generator applies a ringing signal to an attached local loop. Once an idle trunk is seized, if

a sender is not currently free, the call is assumed to queue for sender access. Note, that for interswitch communication to proceed, a request-to-send and an accompanying acknowledgement signal from the tandem switch must be received, prior to transmission of the address information (i.e., the far-end switch must allocate a register for reception of the destination address to be forwarded.) The latter protocol may of course be modified in the event that common-channel interswitch signaling (CCIS) is used with a store-forward dedicated subchannel.

5.2.3 The Control Element

The control element structure and operation (previously described in Figure 4) is now used to derive the delay components of switch performance. In all instances, steady-state operation is assumed.

5.2.3.1 Dial Tone Delay

Let the call arrival stream incident to a given switch be governed by a Poisson process (mean arrival rate, λ calls/second); the call input stream is composed of originating calls (intensity λ_o) and tandem-switched calls (intensity λ_t). Traffic parcels belonging to both classes (originating and tandem) can be directed to local subscribers attached to the switch (destination calls). The input call stream enters the DTM queue and awaits access to a call register. Little information regarding the register holding time distribution is available; for the purposes of tractability it is assumed to be exponential in duration. This, of necessity, represents a simplification, in that the subscriber dialing process is clearly non-Markovian.

An originating caller's register holding time differs from that of a tandem switched call. The latter is generated by the previous switch in the path, whereas the former is entered by a human operator. Denote the originating subscriber's mean register

holding time as d_o and the tandem call's mean register holding time as d_t; the tandem register holding time is determined by the speed of the signaling equipment. For example, if the capacity of the signaling channel is C bps and the size of the signaling message is M, d_t = M/C. The number of registers is given as R. Thus, the average register holding time can be expressed as:

$$d = \frac{\lambda_o}{\lambda_o+\lambda_t} d_o + \frac{\lambda_t}{\lambda_o+\lambda_t} d_t \qquad (1)$$

Assume the marker holding time required to gain access by a caller to the registers is also exponentially distributed (mean duration = m_d). This is not an arbitrary proviso in that if register access is blocked due to path congestion through the switching network, an automatic reattempt may be initiated under switch control.

Based on an analysis of the connection element topology a probability of internal path blocking p can be obtained. If the switch control unit initiated access reattempts indefinitely, then the probability of success after exactly K attempts becomes (with independent trials):

$$\Pr(\text{success in exactly K attempts}) = p^{K-1}(1-p) \qquad (2)$$

If each attempt requires a fixed period of time (L+X) to establish register access, the total time required by the marker to seize a register is geometrically distributed. This represents a discrete version of the exponential distribution. Based on the above limiting argument, the average dial tone marker holding time (with indefinite retrials) is given as:

$$m_d = (L+X)(1+p+p^2+\ldots) = \frac{L+X}{1-p} \qquad (3)$$

The time required to establish register access is comprised of two elements: L, the table look-up period to determine a free register and X, the interval required to connect the calling line to the chosen register (a function of the crosspoint operating speed). Several variations in the marker operation are possible, if a path is blocked; for example, a new register may not be chosen, but instead a second path to the first register may be established. In general, a certain percentage of calls will reattempt access by seizing a second register (q_1), while the other calls retry by using the same register but a different path (q_2). Thus, the mean dial-tone marker holding time becomes:

$$m_d = q_1 \left[\frac{L+X}{1-p}\right] + q_2\left[L+\frac{X}{1-p}\right] \qquad (4)$$

Note that we have implicitly assumed that q_1 and q_2 are inherently determined by the incoming traffic; in actuality, they depend on the switch operating discipline and loading conditions. In addition, m_d depends on the manner in which register access is controlled (on a delay or loss basis). If the marker cannot find an idle register and simply waits until a register becomes available the analysis is significantly complicated and, in fact, becomes intractable with the exception of a few special cases; we therefore postulate that registers are managed on a loss basis, with possible automatic reattempt in the event that all registers are busy. Once an upper limit on total reattempts is exceeded due to register blocking, the call may be returned to the tail of the marker queue or departs from the system.

We will subsequently define hypothetical switch/network operating modes which depend both on the routing and signaling strategies; these strategies determine the manner in which the marker controls the registers for tandem and originating calls. Automatic seizure reattempts when all registers become busy are not as fruitful as reattempts initiated during periods of internal path blocking,

since calls encountering the latter situation can still obtain dial tone by use of a different register or path. Hence, in our analysis, no seizure reattempt, when all registers are occupied, is modeled; instead, call requests are assumed to be flushed from the system (tandem) or returned to the tail of the marker queue (originating), if the "all register busy" condition is encountered.

Assuming that blocked originating call attempts return to the queue and that retrials are governed by a Poisson process, the cumulative originating call arrival rate input to the dial-tone marker queue λ_o^* can be obtained as the solution of the following equation:

$$\lambda_o = \lambda_o^* \, (1-B(a^*,R)) \tag{5}$$

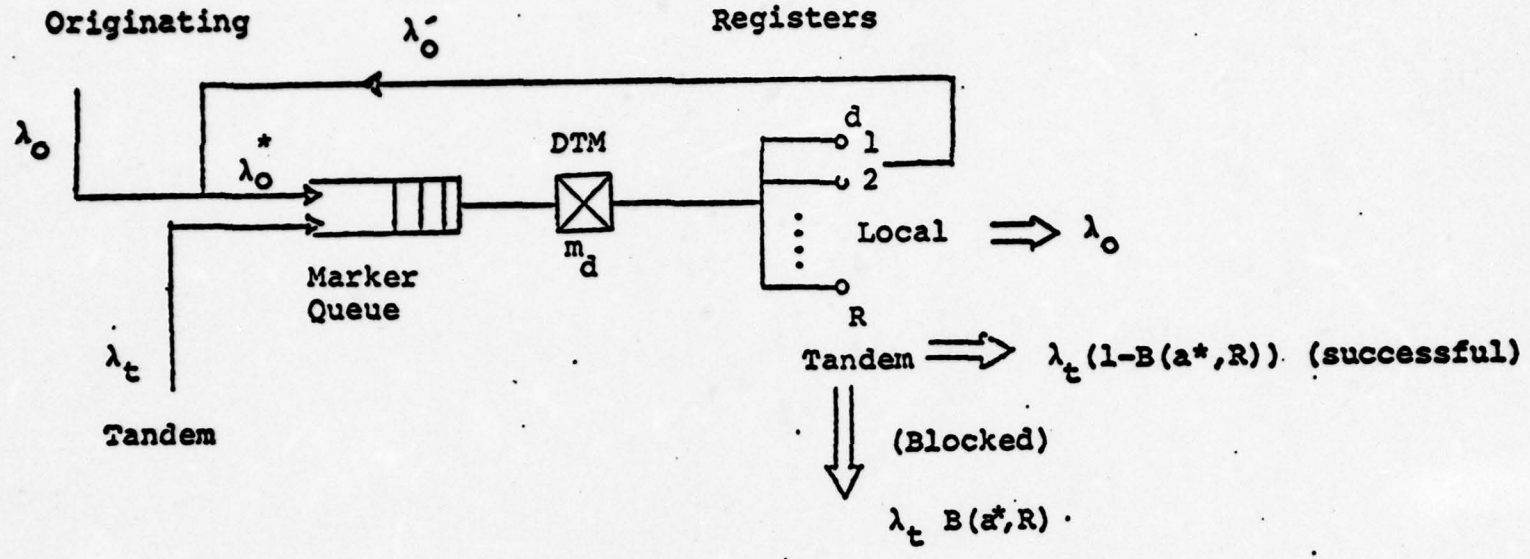
where a^* is the cumulative originating call intensity $a^* = \lambda_o^* d_o + \lambda_t d_t$, R is the number of originating call registers, and $B(a^*,R)$ is the Erlang B blocking equation:

$$B(a,R) = \frac{\frac{a^R}{R!}}{\sum_{i=0}^{N} \frac{a^i}{i!}} \tag{6}$$

Alternatively, if originating calls are allowed to be blocked and are not returned to the queue, $\lambda_o^* = \lambda_o$. Based on the preceding statistical assumptions, the average switch dial-tone delay can be obtained as:

$$D = \frac{1}{\frac{1}{m_d} - (\lambda_o^* + \lambda_t)} \tag{7}$$

Tandem calls are blocked with probability $B(a^*,R)$. The preceding analysis is summarized in Figure 5. We now make the conservative assumption that the output call stream from the dial registers constitutes a Poisson process of intensity, $(\lambda_o^* + \lambda_t)[1-B(a^*,R)]$. If



LEGEND:

m_d	=	average marker holding time = $\frac{L+X}{1-p}$
p	=	internal blocking probability
L+X	=	fixed period of time to establish register access
L	=	table look-up period
X	=	connection establishment
λ_o	=	originating call arrival rate (mean)
λ_t	=	tandem call arrival rate (mean)
λ_o'	=	average arrival rate for retrials
λ_o^*	=	augmented originating call arrival rate (including retrials)
d_o	=	originating call's register holding time
d_t	=	tandem call's register holding time
d	=	average register holding time
a	=	Erlang call register load
R	=	# of registers

FIGURE 5: QUEUEING MODEL USED TO ANALYZE DIAL-TONE DELAY

NETWORK ANALYSIS CORPORATION

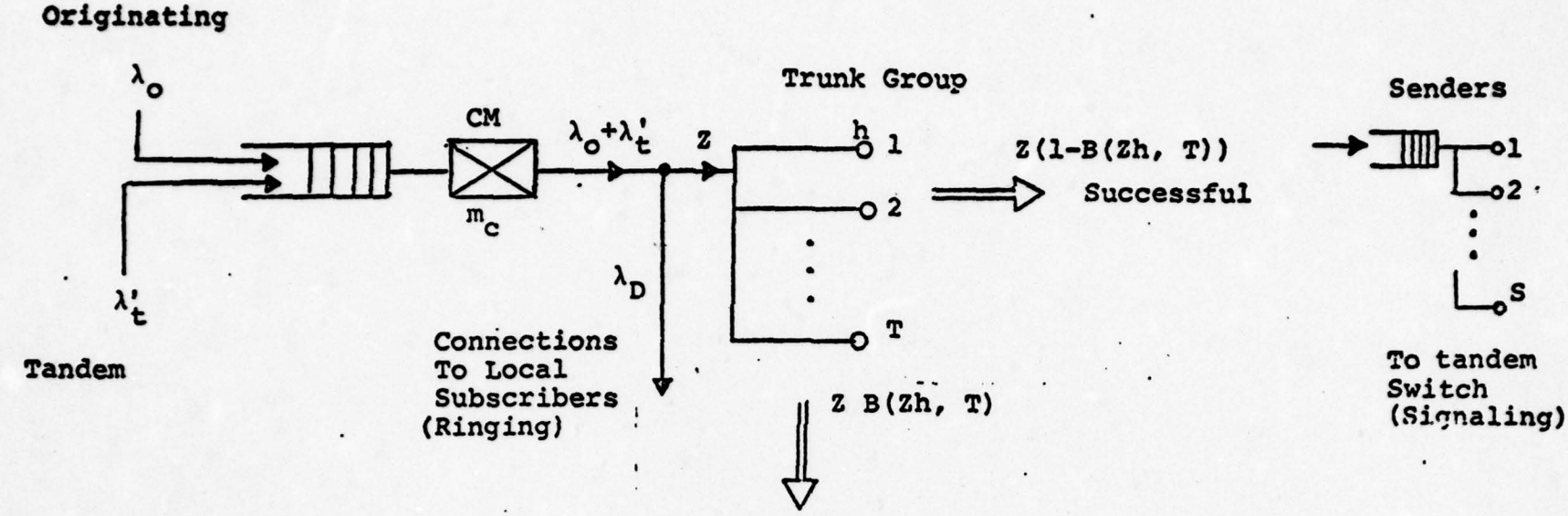
originating calls are allowed to be blocked, the mean call intensity of the output process must be modified to $(\lambda_o+\lambda_t)(1-B(a^*,R))$. In actuality, the successful tandem call stream is smoother than Poisson (variance smaller than mean) due to the statistical "smoothing" imposed by the loss mode of operation; however, this smoothing is currently ignored. For convenience let:

$$\lambda_t' = \lambda_t(1-B(a^*,R)). \qquad (8)$$

5.2.3.2 Connection Delay

Once the destination address is specified, the register contents are assumed to be stored in an interim buffer, and the call advances to the connection phase of the setup process. The register thus becomes available to serve other incoming calls.

The analysis of the switch connection delay is similar to the previous dial-tone delay. Calls desiring a connection enter a queue to await service by the completing marker (CM). The CM must perform address translation, determine an idle trunk and attempt to establish a path through the connection element to that trunk. If no trunk is available the call is blocked. If no free path through the connection element to an idle trunk exists, the connection is retried by either choosing a second idle trunk and attempting to connect to it or by searching for a second path to the original trunk. The relevant address information (signaling) is queued for transmission to the next switch. If CCIS is employed, control information is transmitted over a single signaling channel; if conventional signaling is used, a "sender" device must first be attached to the trunk before transmission of the required information can take place. By invoking the simplifying (yet conservative) assumption of a Poisson arrival stream between devices within the switch, the relevant performance parameters can be obtained. Refer to Figure 6 for a schematic summary.



LEGEND

m_c = average completion marker holding time = $\frac{NX}{1-p} + L'$
p = internal connection element blocking probability
N = number of stages in the connection element
X = crosspoint activation time (space-division or time-division)
L'= translation address execution
h = trunk holding time
e = average sender holding time (varies as a function of signalling strategy)
T = number of outbound trunks
S = number of senders
Z = $\lambda_o + \lambda'_t - \lambda_D$ = call flow offered to switch trunk group

FIGURE 6: QUEUEING MODEL USED TO ANALYZE CONNECTION DELAY

Define the average marker holding time m_c as,

$$m_c = L' + \frac{NX}{1-p} \tag{9}$$

where L' is the fixed time required by the marker to identify an idle trunk (translation and routing via table look-up). The second term accounts for the time necessary to establish a "path" through the connection element to the desired trunk. This clearly depends on the number of stages N which comprise the connection element as well as X, the time required to activate an individual crosspoint within a stage. If an internal path is blocked, calls may reattempt the connection to a trunk in several ways. In general, however, a certain percentage (α_1) will choose a new trunk and attempt to connect to it, whereas the remaining calls (α_2) attempt to establish additional paths to the same trunk. For the general case, therefore, the average marker holding time, m_c becomes:

$$m_c = \alpha_1\left(\frac{L'+NX}{1-p}\right) + \alpha_2\ \left(L'+\frac{NX}{1-p}\right) \tag{10}$$

The average delay due to the completion marker is therefore given by:

$$C_1 = \frac{1}{\frac{1}{m_c} - (\lambda_o+\lambda'_t)} \tag{11}$$

After the path to an idle trunk or local loop is established, the call request attempts to seize signaling equipment so that ringing tone to the destination subscriber or "hopping" to the next switch can take place. Only that portion of the call flow, not specifically destined for subscribers attached to the given switch, is offered to the trunk group. Assuming that λ_D is the call intensity of destination traffic, then the offered tandem flow is $Z = \lambda_o+\lambda'_t-\lambda_D$. Assuming the trunk mean holding time is h and the

trunk access is controlled on a loss basis, the probability that a call request is blocked becomes: B(Zh, T), where T is the number of trunks. In addition, to the length of the message transmission interval, the mean trunk holding time depends in part on the position of the particular node along a path relative to the destination node for certain parcels of traffic. The mean call arrival rate incident to the sender (signaling equipment) queue is therefore:

$$\gamma = Z(1-B(Zh, T)) \tag{12}$$

Finally, queueing for access to the senders or the signaling channel takes place. The sender occupancy time is a direct function of the signaling and control strategies used in the network. For example, the sender could be released as soon as signaling information is forwarded to the next switch in the end-to-end path (progressive route control) or could be held until the entire end-to-end path is established (originating office control). Assuming that the sender holding time is exponential with mean r, and S senders at the switch, the total delay due to sender access becomes:

$$C_2 = \frac{(\gamma r)^S \frac{1}{r}}{(S-1)!\left(\frac{S}{r} - \gamma\right)^2} P_o \tag{13}$$

With

$$P_o = \left[\sum_{m=0}^{S-1} \frac{(\gamma r)^m}{m!} + \frac{(\gamma r)^S}{S!} \left(\frac{S}{S-r\gamma}\right) \right]^{-1} \tag{14}$$

which is the average delay encountered in an M/M/S queue. Thus the total connection delay at a single circuit switch for a successful call becomes (from Equations 11 and 13):

$$C = C_1 + C_2 \qquad (15)$$

Under originating office control, the sender occupancy can exert a feedback influence on total cross-network setup delay; simply, the increased duration for cross-network connection will result in a longer sender holding time at the originating switch and consequently, a greater cross-office delay through this switch for tandem calls, thereby further increasing the setup delay. This, therefore, represents a case where local congestion can "back-up" into the network. Finally, the average cross-office delay through a given switch is obtained (from Equations 7 and 15) as:

$$XOF = \begin{cases} C + d_o, & \text{at the origination switch} \\ C + D + d_o + \tau, & \text{at the destination switch (intra-switch)} \\ C + D + d_t + \tau, & \text{at the destination switch (inter-switch)} \\ C + D + d_t, & \text{at tandem switches} \end{cases}$$

where τ is the subscriber attention interval (time required by destination to pick-up headset). Note that the implicit assumption of independence of the intraswitch Poisson stream has been invoked. The cross-network setup delay is not as readily derived due to its dependence on the system signaling and routing strategies. We will attempt to formally incorporate this dependency in the following sections.

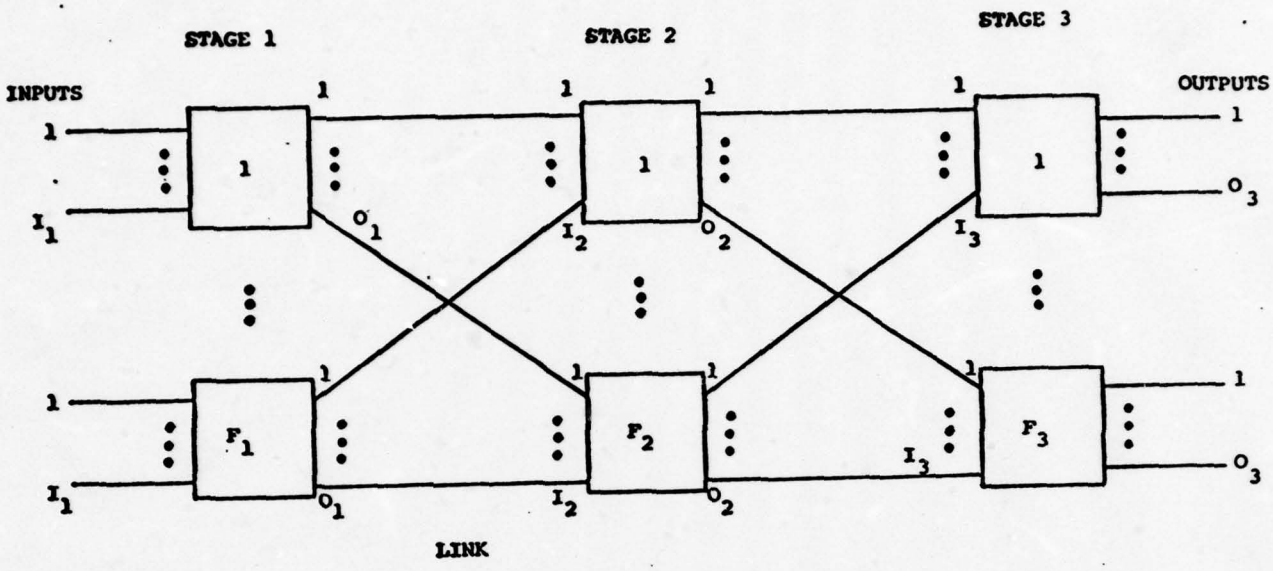
5.2.4 The Connection Element

As discussed in Section 5.1.2.2, the second major component of performance degradation imposed by the switch itself on the incoming call flow is the internal blocking probability, p. The level of internal blocking ("matching" loss) is primarily a function of the connection element topology; additional factors that influence the connection element performance are the inlet-outlet selection procedure and the trunk/link hunting discipline both of which are

determined by the marker operation (i.e., how a path between a chosen inlet-outlet pair is established). Independent of the particular crosspoint type used in the connection element (space-division or time-division), the topology generally consists of a series of crosspoint matrices arranged in a grid or mesh configuration as shown for example in Figure 7. In any grid topology, there exists a sequence of stages; all matrices within a stage possess the identical dimensional capacity, while matrices in different stages usually possess different capacity. A series of "links" interconnect matrices within different stages; matrices within the same stage are referred to as frames. Any path which must be setup through the connection element proceeds sequentially from stage to stage using an appropriate set of crosspoints and links.

Clearly, given a set of input-output pairs, there exists a vast number of potential connection element topologies dependent on the matrix capacity (number of inlet-outlets), number of stages, number of frames within a stage and the interstage wiring pattern. However, several practical considerations limit the number of possible configurations. Matrices are usually manufactured in a set of standard sizes, so that it is not cost-effective to employ frames which require customized design. Depending on the crosspoint element used, the number of stages cannot exceed a certain threshold value due to possible signal attenuation; in addition, the control element operation is significantly more complex for a large number of stages, and the time to establish an internal connection grows at least linearly in the number of stages.

A general conclusion which can be drawn regarding the cost-complexity tradeoff among various connection element topologies is that a configuration composed of many stages of small matrices is economical in terms of number of crosspoints required but complex to control, whereas a configuration using relatively few stages of



N = NUMBER OF STAGES = 3

$I_j \times O_j$ = MATRIX SIZE IN STAGE j

F_j = NUMBER OF FRAMES (MATRICES) IN STAGE j

SYMMETRY CONDITIONS

$O_{j-1} = F_j \quad \forall j$

$F_{j-1} = I_j$

MATRIX TYPES

$I_j = O_j$; DISTRIBUTION

$I_j < O_j$; EXPANSION

$I_j > O_j$; CONCENTRATION

FIGURE 7: CONNECTION ELEMENT TOPOLOGY

large matrices possesses simplicity of operation but is more expensive due to the greater number of crosspoints required. In the limiting case, a single matrix (of dimensions I x O crosspoints) could provide interconnection for I inputs and O outputs. As derived in [CLOS, 1953], a three stage connection element which provides non-blocking access from N inputs to N outputs requires $6N^{3/2} - 3N$ crosspoints compared to N^2 crosspoints for a single stage connection element; for $N \geq 36$, the three stage configuration uses fewer crosspoints. Finally, connection element stages can be classfied according to the function they perform as indicated in Figure 7: distribution, connection or expansion.

Occasionally, separate connection elements are provided for interswitch and intraswitch traffic; in the following analysis, we assume a single connection element is used to service both types of calls. Despite the reduced number of potential connection element topologies which are of practical interest due to the above constraints, the performance analysis of these structures is quite complicated and in most cases intractable. The difficulty stems from the interstage dependence; the effect of this dependence tends to reduce the internal blocking probability because of the overlapping occurrence of congestion events in different stages. Hence, the simultaneity of blocking in different stages is higher in actual operation than if this congestion occurred independently. Since calls can be lost during periods of heavy link occupancy in a single stage, the overlapping congestion in different stages tends to reduce the total period of time during which calls can be internally blocked. Thus, the independence assumption will overestimate the internal blocking, and our analysis will be conservative if it is invoked. We will employ the method of Lee [LEE, 1955], which is relatively simple and possesses wide applicability for most structures of practical interest (series-parallel graphs); it cannot, however, be used for an arbitrary topology without significant

modification. It also cannot be readily generalized to topologies performing concentration. Therefore, for many of the complex structures encountered, the only viable analytic recourse is simulation.

Given a series-parallel connection element topology and with the link occupancy distribution assumed to be Bernoullian in each stage, the internal blocking probability (for a particular inlet-outlet pair) can be expressed as:

$$p = (1 - \prod_{i=1}^{N-1} (1 - \frac{y_i}{O_1}))^{O_1} \tag{17}$$

where N is the number of stages, O_1 is the number of output links from a matrix in the first stage, and y_i is the average link occupancy at stage i. If the average link occupancy is identical at each stage ($y_i = y, \forall_i$), Equation 17 reduces to the well-known Kitteridge-Molina expression:

$$p = (1 - (1 - u)^{N-1})^{O_1} \tag{18}$$

with $u = y/O_1$. The above equation simply accounts for the occurrence of events in which all links output from a frame in the first stage are busy. The computation of the average link occupancy (y) is far from trivial since that portion of the traffic which is lost due to internal path mismatch at each stage must be accounted for. Clearly, the link occupancies are themselves a function of the internal blocking probability. For a total offered load of A erlangs to the connection element (assumed Poisson call arrivals), and balanced subscriber activity (equal load per inlet to the first stage matrix), the link occupancy can be found as (using Equation 18):

$$y = \frac{A(1-p)}{F_1 I_1} \tag{19}$$

Most practical connection element architectures typically exhibit extremely low values of internal loss ($p \leq .001$); hence as a first order approximation to the link occupancy y, the internal blocking p in Equation 19 can be safely neglected.

5.2.5 Assumptions/Limitations

Apart from the obvious statistical simplifications embodied in our current model, there are several important phenomena which have not been accounted for; this, in part, represents a deficiency.

We had earlier indicated that switch architectural variations may modify certain aspects of the generic structure from that depicted in Figure 4. Specifically, only a single marker could be used for both connection and dial-tone. Registers and senders may be combined into a single functional unit. Hardwired registers may be replaced by buffer memory which is accessed not under marker control, but program control. The use of a common-channel signaling strategy may eliminate entirely the presence of the peripheral signaling equipment (register/senders). To a high degree, our model is sufficiently general so that the manner in which a certain function is performed (whether in hardware, software or both), can be incorporated.

There are however several features which presently exist in several switches that are not modeled. These are now described; attempts will be made to incorporate certain facets of these operations into the existing model in subsequent development.

- Time-outs: The model assumes that a switch resource once seized, is held indefinitely until a path is set up or blocking occurs. Under conditions of heavy load, however, a circuit switch can time-out the occupancy of a particular resource (e.g., register) and subsequently "block" the call.

This often results in an improved grade of service to other subscribers; timeouts are also employed to control the incidence of false switch access due to "hits" or malicious users.

- Operating System: No formal attempt has been made to account for the actual mode of switch CPU operation. However, several switches have a cyclically-based task structure, which introduces additional levels of overhead into the call setup procedure. For example, certain operations may not be able to be performed as soon as an event occurs, but must wait until the control program schedules their execution.

- Scanning: Most modern switches do not interface with the local loop on an event-driven basis, but provide a scanning mechanism by which the CPU periodically checks for subscriber activity. In this manner, work flow can be more effectively managed.

- Load Leveling: Many switches provide flow control so that if the activity within the switch exceeds a certain threshold (which can be varied), the calls that subsequently arrive for processing are automatically blocked and a busy signal is returned.

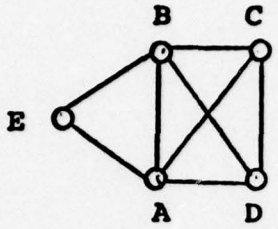
5.3 SWITCH OPERATION IN THE NETWORK ENVIRONMENT

The operation of the circuit switch in a network environment is now examined. Among several tradeoffs investigated, of primary interest is the impact of the switch's operating mode on network performance: Emphasis will be placed on the routing plan and signaling strategy. In addition, of course, the switch technology, operating speed and capacity will also impact network performance. The relevant network performance measures to be derived are, end-to-end blocking probability (loss) and cross-network setup delay. The derivation of end-to-end blocking probability has been the province of the circuit switching network design effort and will not be further discussed here. The topological structure and operation of the switch connection element contributes a certain degree of internal blocking to the end-to-end path loss and will be accounted for in the network design blocking analysis.

5.3.1 Routing/Control Strategies

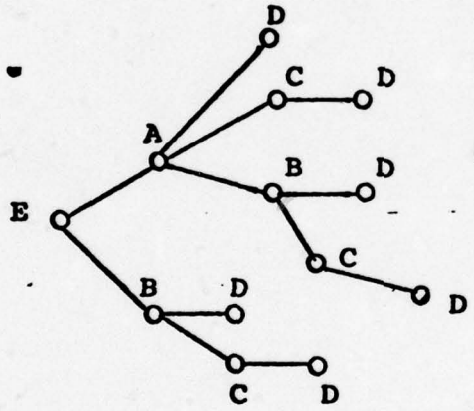
Among the multitude of routing strategies which can potentially be employed in a circuit switched network, we will restrict the scope of the setup delay analysis to two control techniques: Originating Office Control and Progressive or Sequential Control. The relevant differences between both strategies is now described.

Consider the exemplary network topology depicted in Figure 8a. Under the indicated routing plan, the set of available end-to-end paths for traffic originating at node E and destined for node D is illustrated in Figures 8b and 8c for Progressive Route Control (PRC) and Originating Office Control (OOC), respectively. The routing plan entry (Figure 8a) indicates the next node in a path to be chosen by a call received at a given node; for example, a call transmitted from node A (the last hop) and ultimately destined for node D, would

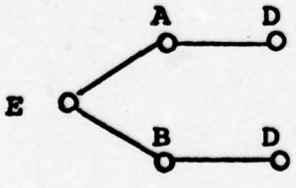


From \ To	A	B	C	D	E
A	-	B	CB	DCB	EB
B	A	-	CD	DC	E
C	AB	BA	-	D	AB
D	ABC	BC	C	-	BA
E	AB	BA	AB	AB	-

(a) Exemplary Network Topology and Associated Routing Plan



(b) E-D Path Traversal Under Progressive Route Control (PRC)



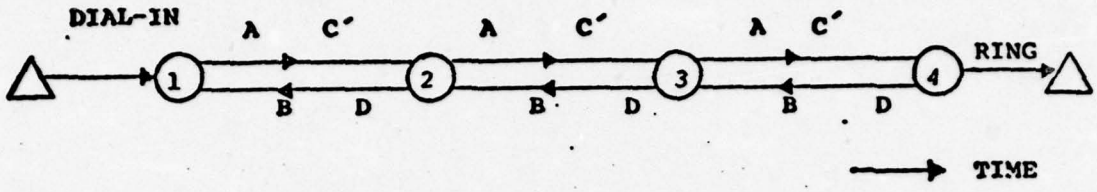
(c) E-D Path Traversal Under Originating Office Control (OOC)

FIGURE 8: ROUTE COMPARISON BETWEEN ORIGINATING-OFFICE CONTROL AND PROGRESSIVE CONTROL STRATEGIES

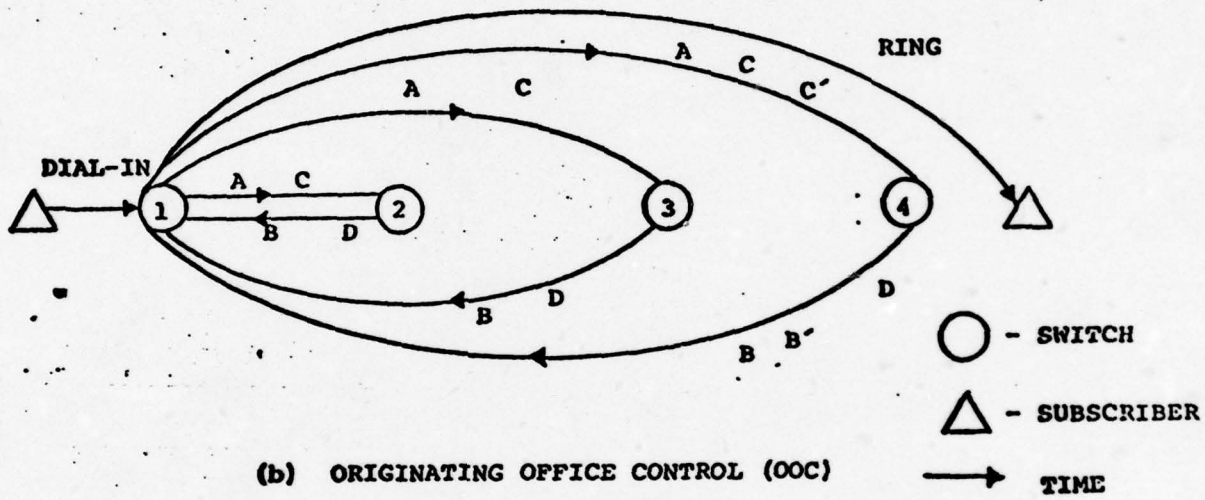
pick A-D as its first choice (primary) route, A-C as the first link in its first alternate route and A-B as the first link in its second alternate route (hence, the A-D entry in the routing plan table is DCB). Clearly, for the routing plan illustrated, the set of allowable paths which can be traversed for E-D traffic (Figures 8b and 8c) differs dramatically as a function of the control strategy employed.

In progressive route control, tandem nodes are capable of performing alternate routing of calls; in the extreme version of originating office control, tandem nodes are not able to perform alternate routing of calls; all route decision-making is delegated to the originating office. Hence, as shown in Figure 8 calls forwarded to node B from node E under OOC can only proceed directly to node D (single choice) for the given routing plan. A "spill-forward" feature is often employed in conjunction with OOC, at certain designated nodes (that are allowed to perform alternate routing), to compensate for the relatively limited freedom of path choice characteristic of OOC. There exists a spectrum of possible control techniques that can be used in a circuit switching environment; however, OOC and PRC are analyzed since they, in effect, represent extreme variants of the procedures to be used. It must be emphasized that the set of paths traversed by a call between a particular origin-destination node pair can be identical under either control strategy (OOC or PRC) and is primarily a function of the routing choices, not the intelligence/control distribution in the network. In addition, the assumption that tandem nodes under OOC possess no decision-making capability is often relaxed. The major distinction between either control strategy that will be repeatedly stressed is the manner in which the route evolves in response to a connection request and not the set of paths traversed.

The sequence of operations performed by the switch for each control strategy and the associated setup delay analysis is presented in the following sections. Figure 9 depicts the essential operational differences between OOC and PRC.



(a) PROGRESSIVE ROUTE CONTROL (PRC)



(b) ORIGINATING OFFICE CONTROL (OOC)

SIGNALS:	**A**	-	Request for connection
	B,B'	-	Ready to receive destination address
	C	-	Transmission of destination address (abbreviated)
	D	-	Connection established
	C'	-	Transmission of destination address (full)
OCCUPANCY:	**PRC**	-	Senders at each switch are held for a portion of the end-to-end path setup interval
	OOC	-	Only sender at the source switch is employed during the end-to-end path setup interval

FIGURE 9: OPERATIONAL COMPARISON BETWEEN OOC AND PRC

5.3.1.1 Progressive Route Control

In this section, we describe the manner in which an end-to-end network connection is established with respect to the switch interaction under progressive route control (refer to Figures 4 and 9). For the sequence of operations described, the signaling strategy is currently restricted to be in-band.

An originating caller at the source switch, awaits system access, seizes a register (dial-tone returned), enters the destination address and obtains a connection to a tandem switch. The sender in the source switch must transmit the entire destination address to the tandem switch's call register, so that the tandem switch can perform the appropriate routing function (see Figure 9). The register in the tandem switch is generally occupied for a shorter period of time than the corresponding register at the origin, due to the inherently faster tandem signaling (accomplished by tone regeneration and eliminating lengthy interdigit pauses). Once the destination address has been received by the tandem switch, the sender (or in certain switch architectures, register/sender) at the source switch is released from the connection and becomes available to service other waiting calls. The physical (space-division) or temporal (time-division) path that has been established up to this point in the source switch is, of course, retained. Once the destination address is received at the tandem switch, it exercises total control over subsequent route establishment. Generally, if a call becomes blocked at a tandem switch, no "traversal reversal" along the existing portion of the connection is allowed, although some variations permit "backing-up" along the path for distances not exceeding a few hops. The major difference between the route control strategies with respect to switch resource utilization, is the length of time a given switch's sender is occupied by the processing of an individual call. Clearly, under PRC, senders at each switch in a given path are used in the establishment of the end-to-end connection during a portion of the setup interval.

We now indicate the analytic details which characterize PRC and are incorporated into the generic node model. The following network-based notation will be used to analyze both the originating office control (OOC) and progressive route control (PRC) techniques.

Given a network topology (set of nodes and links), routing plan (set of paths over which calls are allowed to flow) and link capacity (the number of trunks in a group), the following quantities can be readily determined:

λ_{ij} = Offered traffic between origin-destination node pair i-j (traffic requirement).

λ = Total network offered traffic = $\sum_i \sum_j \lambda_{ij}$

λ_{ij} = Carried traffic between origin-destination node pair i-j.

η_{ij} = Offered traffic on link (i,j), (assuming no blocking takes place).

σ_{ij} = Carried traffic on link (i,j).

σ = Total network carried traffic = $\sum_i \sum_j \sigma_{ij}$

N_{ij} = Number of routes between origin-destination node pair i-j.

V_{ij}^m = Set of switches which comprise the m^{th} route between origin-destination node pair i-j, (including source and sink).

To preserve consistency with our previous development (Equation 3), the mean originating and tandem call intensities at a particular switch K can be expressed as:

$$\lambda_o(K) = \sum_j \lambda_{Kj} \qquad (20)$$

$$\lambda_t(K) = \sum_i \sigma_{iK} \qquad (21)$$

where the first summation is taken over all destination nodes, and the second is performed over all nodes directly connected to K. Note that both originating and tandem calls may contain destination traffic.

With the number of registers, connection element structure, marker speed and input call intensity for switch K specified, the average dial-tone delay D(K) is uniquely determined (see Equation 7). Denoting the next switch along an i-j end-to-end path following switch K as K', the average sender holding time in switch K due to i-j traffic can be computed. For convenience, we assume a uniform type of interswitch signaling throughout the network (e.g., dial-pulse, touchtone MF, ACU, CCIS, etc.), although this assumption is not restrictive and can be readily generalized to account for individual switch signaling hardware. We furthermore postulate the exact manner in which the interswitch signaling (inband or common-channel) is performed (see Figure 9a).

A request for register signal is transmitted from K to K' (signal A in Figure 9), after waiting for the appropriate period of time in the dial-tone marker queue, access to an available register in switch K' is granted and an acknowledgement message, indicating that switch K' is ready to accept the destination address, is transmitted back to K (signal B in Figure 9). The sender hardware (or software if SPC is employed) in switch K then transmits the desired address information to K' (signal C′ in Figure 9). The total average interval during which the sender in K is occupied due to interaction with K' therefore becomes (see Equation 10a):

$$r(K,K') = RR + D(K') + ACK + XMIT(1) \qquad (22)$$

where RR is the time to transmit the request for register (connection), ACK is the time to receive the acknowledgement, D(K') is the average dial-tone delay encountered at node K', and XMIT(1) is the time required by switch K to transmit the address to K'. A distinction is made between the amount of addressing information which

must be sent for the purposes of routing under PRC vs. OOC. In the latter case, an abbreviated address can be forwarded to tandem switches, because the originating office continues to exercise total route control; hence, the time required for signal transmission under OOC (XMIT(2)) is significantly less than that encountered under PRC (XMIT(1)).

Based on the aforementioned routing plan, the percentage call intensity output from node K to each adjacent node N can also be obtained. Thus, the weighted average sender occupancy at switch K is given as:

$$r_K = \sum_N \frac{\sigma_{KN}}{\sigma_K} r(K,N), \quad \sigma_K = \sum_j \sigma_{Kj} \tag{23}$$

In addition, the register occupancy at each switch adjacent to K is increased due to the transmission of the acknowledgement prior to reception of the destination address. This latency can be incorporated into the average register holding time for tandem calls, d_t, and is a function of the speed of the signaling channel and/or equipment. Therefore, the register and sender utilization derived in this section is determined by the speed/capacity of the switches, the speed of the signaling equipment and/or channel and the call intensity input to each node along the end-to-end path. Thus, under progressive route control, the average setup delay ε for unblocked calls originating at i and destined for j using the m^{th} i-j route is given as

$$\varepsilon(i,j,m) = \sum_{L \in V_{ij}^m} XOF(L) \tag{24}$$

where V_{ij}^m is the set of nodes comprising route m, and the summation is taken at each node along the path. Similarly, the average setup delay encountered by all i-j traffic is given by:

$$\varepsilon_p(i,j) = \sum_{m=1}^{N_{ij}} \Omega_m \varepsilon(i,j,m) \qquad (25)$$

where Ω_m is the percentage of i-j traffic carried along route m. Concluding in a similar fashion, the average setup delay for the entire network becomes

$$\varepsilon_p = \frac{1}{\sigma} \sum_i \sum_j \gamma_{ij} \varepsilon_p(i,j) \qquad (26)$$

and is the total level of traffic carried by the network.

It is apparent that the preceding analytical development represents a considerable oversimplification of the actual switch interactions during end-to-end call setup. In addition to the operational limitations discussed earlier in Section 5.2.5, two stochastic deficiencies exist. The assumption concerning the existence of independent Poissonian traffic throughout the network, although common, is inaccurate due to the aforementioned statistical "smoothing" imposed on the successful call flow. However, since the "smoothed" traffic (which possess greater statistical regularity than a Poisson stream of the identical intensity) generates less of a fluctuating demand on the switch, the "pure random" proviso is conservative.

The second deficiency is contained in Equation 13, which implicitly assumes that the sender holding time is exponentially distributed. This is somewhat unrealistic (see Equation 23), since the actual sender service time distribution is determined by the dial tone delay encountered at adjacent switches in the network. The individual dial tone delay based on the assumptions regarding call flow and marker operation is characterized by an exponential density. However, because the dial-tone delay at each adjacent switch will, in general, possess a different mean value, the sender occupancy is described by a weighted sum of exponentially distributed random variables (as in Equation 23); this service time distribution is usually referred to in the literature as hyperexponential. Hence, the M/H/S sender queue

is replaced by an M/M/S queue with an equivalent average weighted service time. As demonstrated in [KOTIAH, 1970], the M/M/S queue, in general, provides superior performance to M/H/S, since the hyperexponential density possess a standard deviation greater than its mean. Thus, the assumption regarding exponentially distributed sender occupancy is not conservative, but must be invoked to preserve tractability. An indication concerning the accuracy of the preceding approximation can be obtained by examining the service distribution's coefficient of variation β (ratio of standard deviation to mean); the hyperexponential distribution always possesses $\beta>1$. For $\beta\approx1$, the M/M/S queue provides an excellent approximation to the M/H/S queue's performance. Note, that a high value of β (the region where the approximation is inapplicable) implies that a certain proportion of calls encounter widely differing dial tone delay at adjacent switches, which indicates a poor routing assignment or insufficient tandem switch capacity.

5.3.1.2 Originating Office Control

The sequence of operations required to set up a call under originating office control (OOC) differ significantly from those associated with progressive control (PRC). In direct contrast to PRC, only the originating switch is responsible for making route decisions during the establishment of the call path. However, since the entire destination address is stored at the source node, only the minimum amount of information required by a tandem switch to route the call is transmitted from the origin. Hence, the registers at each switch are occupied for a shorter period of time due to tandem call information. Depending on the network topology, OOC may offer greater flexibility in route choice, because of its inherent ability to "backup" an evolving connection which becomes blocked at an interim switch; however, the route flexibility under OOC as exemplified in Figure 8 is not uniformly superior.

A second major distinction between OOC and PRC is the occupancy of the sender at tandem nodes. Under OOC, switches do not use their sender equipment for tandem traffic call setup since all signaling is performed by the origin; hence, no queueing for sender access is required at a tandem switch. Once a connection is established at a particular intermediate node, the tandem switch transmits an acknowledgment of this condition back to the originating switch (signal D in Figure 9b). The origin, which still retains the use of its sender during the processing of the current call, receives the acknowledgment and then transmits a request for register access to the next switch along the path via the already established connection (signal A). Thus, as shown in Figure 9, once the connection to switch 2 is established, switch 1 (not switch 2) forwards the destination address to switch 3. The register/sender or sender at the originating switch is therefore occupied during the entire setup procedure for the call, but the senders at the intermediate switches along a path are never used. As discussed, due to the fact that route control is completely relegated to the source switch, the duration of the interswitch signaling interval is significantly reduced, except at the destination switch. Once the abbreviated address is received by the destination switch, and it determines that the called subscriber is attached to it (not to a different switch), the destination node requests the full subscriber address from the originating switch (signal B' in Figure 9b). After transmission of this information (signal C'), the connection is established.

The average sender occupancy interval at the originating switch can now be formulated. Given the network routing plan, the sequence of switches, which a given call using the m^{th} route between node i and node j, traverses in an end-to-end connection can be determined. Now if the time required to establish a connection at switch K (excluding the tandem sender queueing delay, which is not encountered in OOC) is denoted as $C'(K)=C_1(K)$ (See Equation 11), the sender occupancy at switch i due to (i,j) origin-destination

traffic can be expressed as:

$$r_i^j = \sum_{m=1}^{N_{ij}} \sum_{K \epsilon V_{ij}^m} (RR+D(K)+ACK+XMIT(2)+C'(K)) + ACK+XMIT(1) \quad (27)$$

The double summation accounts for the interswitch protocol delay required to establish the ongoing connection at tandem switches along a given route as follows: RR is the time to transmit a request for access from the originating switch to the tandem switch, D(K) is the average dial-tone delay experienced at switch K, ACK is the time required to transmit the acknowledgment of a tandem switch's availability back to the originating switch, XMIT(2) is the time required to send the abbreviated destination address from the origin to the destination, and C'(K) is the connection establishment interval at switch K. We further assume that the notification of a connection's success by a tandem switch back to the origin (signal D in Figure 9b) is overlapped with the actual connection establishment. The last term in Equation 27 (not included in the double summation), assumes that the final switch in the path, after determining it is the destination switch, transmits a request for the full destination address back to the origin (ACK), which then supplies the desired information (XMIT(1)); once the total address is received, the connection is established.

Note that Equation 27 for OOC accounts for the sender occupancy due to traffic over a given end-to-end path (i,j) and is not solely influenced by the effect of adjacent switches (as in Equation 22 for PRC); hence, a different notation is employed. Given the parameter γ_{ij} (mean carried call flow originating at i destined to j), the total average sender delay at switch i can be obtained as:

$$r_i = \frac{1}{\gamma_i} \sum_j \gamma_{ij} r_i^j \quad (28)$$

Thus, the setup delay along path (i,j) under originating office control is simply (see equations 16 through 19):

$$\varepsilon_o(i,j) = XOF(i) \tag{29}$$

where XOF(i) is the cross office delay encountered at switch i, obtained by substituting the sender occupancy r_i into the expression for mean connection delay (Equation 13). The call flow input to the sender queue of Figure 6 must be modified from γ of Equation 12 to:

$$\gamma = (\lambda_o - \lambda_D)(1 - B(Zh, T)) \tag{15d}$$

since under OOC only calls originated at a given switch require use of its signaling equipment; tandem calls do not use senders as assumed in Equation 12. Finally, the average network setup delay under originating office control becomes:

$$\varepsilon_o = \frac{1}{\sigma} \sum_i \sum_j \gamma_{ij} \varepsilon_o(i,j) \tag{31}$$

Note that Equation 31 employs the identical implicit assumption used in Equation 24, i.e., the sender occupancy is exponentially distributed. However, in the case of Equation 29, the simplification is taken with respect to end-to-end network paths not simply at adjacent switches. Hence, similar care must be exercised regarding the applicability of the above formulation, as was described for Equation 24.

In the previous sections, we have endeavored to effect a performance comparison between originating office control and progressive control routing strategies. Despite the advantages of one strategy over another, the choice is often based on more mundane considerations. For example, inherent capabilities of the switching equipment could argue heavily in favor of a particular strategy; a case in point occurs in a network which possesses switches of widely differing capacity, thereby necessitating connection setup to be controlled from a select subset of powerful switches (hierarchical topology, originating office control with spill).

5.3.1.3 Disconnection Delay

In the preceding formulation, the cross network disconnection delay has not been explicitly taken into account. The mechanism by which a circuit switch performs connection breakdown does not appreciably vary as a function of the route control strategy. Several possible techniques exist for effecting the connection breakdown; they primarily differ in the manner in which the connection is supervised and the propagation of the disconnect signal.

In certain implementations, either the origination or destination switch can initiate the disconnection procedures, while other variations require that the originating switch be responsible for controlling the path cleardown, independent of which subscriber (called or caller) first "hangs-up". If a per-circuit signaling technique is used in the network, the disconnect signal can be allowed to propagate along the already established connection to each switch. When all the switches in the path receive the disconnect signal and their connection memories are cleared with the appropriate interswitch trunks released, the connection can be regarded as "broken." Under common channel signaling (CCIS), however, the disconnect signal is transmitted in a store-forward fashion and the calling path proper cannot be used to convey disconnect information. Despite queueing delays, CCIS should be significantly faster due to the higher capacity signaling channel.

Note that the above characterization of the disconnection delay is somewhat pragmatic in that several important operations have been ignored, notably logging of call elapsed time for billing purposes and statistics collection. Hence, depending on the perspective adopted (user, switch or network), the processing of a particular call can be considered to be terminated at various times. For the purposes of the current analysis and in the absence of detailed disconnection procedures, we will assume that the disconnection interval is significantly shorter than the setup delay, and will hence be ignored.

5.3.1.4 Subscriber Attention Interval

The length of time required by the destination subscriber to acknowledge ringing (phone goes "off-hook") has been previously defined as the subscriber attention interval, τ. Under either routing strategy, all trunks which comprise an end-to-end connection are occupied during the subscriber attention interval; hence, the additional factor τ must be incorporated into the trunk holding time, i.e.

$$h = v + \tau \qquad (32)$$

where v is the conversation duration or transaction transmission interval depicted in Figure 2. In subsequent experimentation, we will fix the value of the subscriber attention interval.

5.3.1.5 Channel Seizure Policy

A more subtle phenomenon which also augments the trunk holding time is the channel seizure policy used in the network. Under either OOC or PRC, the actual allocation of transmission bandwidth is performed in a sequential manner; hence, prior to the advancement of a connection request from a given switch to the next switch in the path determined by the routing plan, the appropriate interswitch trunk must first be seized. Note that the sequentiality of trunk seizure is independent of the network signaling strategy (common channel signaling or per channel signaling); however, variations to existing signaling methods have been proposed in which trunks are only "reserved" but not formally allocated until the status of the destination subscriber (busy or free) has been ascertained. The sequential seizure of channels implies that for a portion of the connection setup interval, trunks comprising an end-to-end path are occupied but not constructively used for the transmission of information. Clearly, those trunks which are attached to switches en-

countered earliest on the end-to-end path will be occupied in this unproductive manner for the longest period of time. This mode of operation thus indirectly increases the end-to-end loss probability, since, as the cross network setup delay grows, the individual trunk holding time increases and thus, the network blocking probability also increases.

The impact of sequential channel seizure is not easily incorporated into the preceding analysis due to an intrinsic dependency which is shortly described. The trunk holding time is made up of several components: transmission interval v, subscriber attention interval τ and the channel seizure overhead (CSO). A more accurate expression for the holding time can be explicitly formulated due to the m^{th} routed (i,j) origin-destination traffic at node K as (see Equation 32):

$$h(i,j,m,K) = v + \tau + \sum_{L\varepsilon\, V_{ij}^{m}(K)} XOF(L) \qquad (33)$$

where $V_{ij}^{m}(K)$ is the partial path from switch K to the destination switch j (i.e., the set of nodes on the m^{th} i-j route which are encountered once the path has been established successfully up to node K). For example, if $V_{ij}^{m} = \{i,b,c,K,d,e,j\}$, then $V_{ij}^{m}(K) = \{d,e,j\}$. Then in a straightforward fashion, one can obtain the average trunk holding time at switch K due to all (i,j) traffic (see Equations 25 and 33):

$$h(i,j,K) = \sum_{m=1}^{N_{ij}} \Omega_m\, h(i,j,m,K) \qquad (34)$$

where Ω_m is the fraction of i-j traffic carried along route m. Finally, the mean trunk holding time at node K due to all traffic parcels is given by:

$$h_K = \frac{1}{\sigma} \sum_i \sum_j \gamma_{ij}\, h(i,j,K) \qquad (35)$$

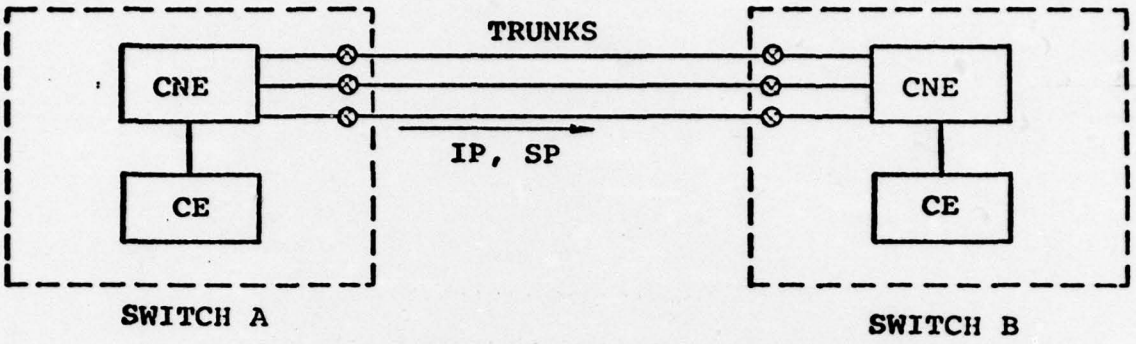
The dependency which precludes incorporating the above equations directly into our analysis arises in Equation 33. The cross-office delay at successive switches along an end-to-end path is determined by the mean trunk holding time at each switch, which is in turn dependent on the cross office delays at other switches comprising various partial network paths. Equation 33 therefore, defines an implicit relationship among the mean trunk holding times at every node in the network. Unfortunately, the functional dependency of the cross office delay on mean trunk holding time cannot be expressed in closed form as evidenced by Equations 12 through 16; however, the relationship can be quantified via numerical methods.

5.3.2 Signaling Techniques

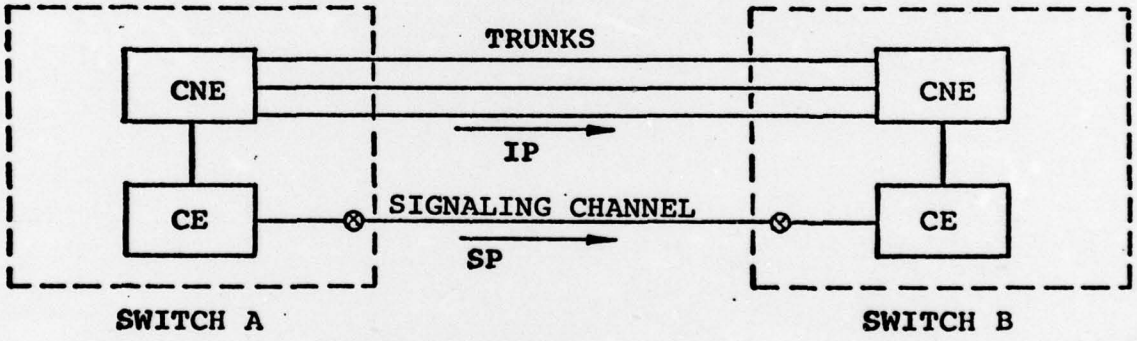
In the preceding section, we have analyzed the network performance as a function of the routing/control strategies. However, the cross-network setup delay is also influenced by the signaling technique employed. The network signaling technique determines the speed with which interswitch protocol messages can be exchanged, the duration of the destination address transmission interval and thus, the cross-network setup delay. We now investigate two broad categories of signaling strategy: per circuit signalling and common channel signaling. The major architectural differences between the two techniques are summarized in Figure 10.

5.3.2.1 Per-Circuit Signaling

The per-circuit signaling technique (shown in Figure 10a) utilizes the basic communication path (trunks) for both the transission of information as well as the exchange of interswitch address



(a) PER CIRCUIT SIGNALING



(b) COMMON CHANNEL SIGNALING

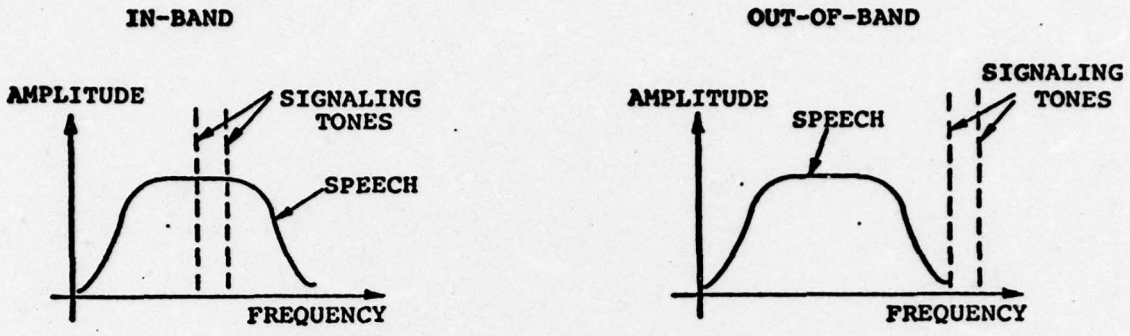
LEGEND: ⊗ - CHANNEL INTERFACE EQUIPMENT
CE - CONTROL ELEMENT
CNE - CONNECTION ELEMENT
IP - INFORMATION PATH
SP - SIGNALING PATH

FIGURE 10: ARCHITECTURAL COMPARISON BETWEEN SIGNALING TECHNIQUES

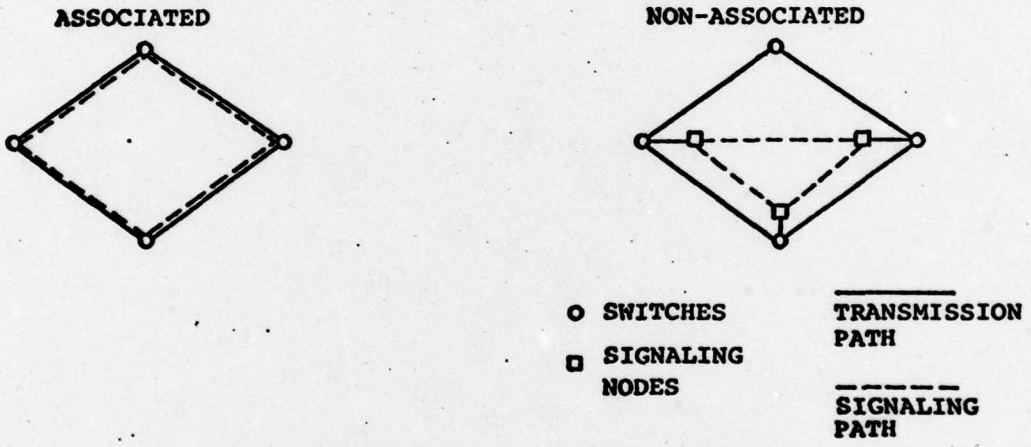
signals. Hence, under per-circuit signaling, a channel which is seized to ultimately service a call (provide transmission bandwidth) also acts as the medium by which the calling path is established. At present, per-circuit signaling is the most prevalent form of signaling strategy.

Many variations of the basic technique exist such as DC signaling, in-band signaling and out-of-band signaling (see Figure 11). All variations use the same physical channel for both transmission and signaling, but not the identical portion of the channel bandwidth. In-band signaling is more expensive to implement than out-of-band signaling due to the elaborate channel filters which must be used to distinguish between speech and signaling tones. Out-of-band signaling is often introduced to reduce channel filter cost and compensate for the problem of "talkdown," in which speech energy is misinterpreted as signaling or supervisory information causing a premature disconnection of the circuit. Under an out-of-band technique, signaling information is transmitted in a band where the speech energy is minimal. All per-circuit signaling techniques possess the advantage of being able to use existing transmission facilities; hence, the creation of a dedicated higher capacity signaling channel is not required.

For a per-circuit strategy, the maximum signaling transmission rate is determined by the operating speed of the signaling equipment (registers and senders) and not the channel capacity. The speed of the signaling equipment varies depending on the type of end instrument employed (dial pulse, DTMF, ACU, etc.). The length of the signaling message coupled with the operating speed therefore deterimines the address transmission interval. Denoting the speed of the signaling equipment as F (digits/second) and the number of address digits transmitted between switches as d_1 (d_2) for the full (abbreviated) address, then the address transmission interval for per-circuit signaling is given as:



(a) PER CIRCUIT SIGNALING VARIATIONS



(b) COMMON CHANNEL SIGNALING VARIATIONS

FIGURE 11: VARIATIONS OF THE SIGNALING TECHNIQUES

$$XMIT(1) = F\ d_1 \quad (PRC) \tag{36}$$

$$XMIT(2) = F\ d_2 \quad (OOC) \tag{37}$$

For end instruments which employ dial pulse signaling, F must be regarded as an average value, since in addition to the signaling speed and length of the address information (number of digits), the numerical value of the dialed digits also influences the duration of the address transmission interval. For example, a value of "9" requires the generation of nine make-break pulses. Based on the above equations, the average sender occupancy (r) and tandem register holding time (d_t) can be obtained.

5.3.2.2 Common Channel Signaling

As shown in Figure 10b, common channel interswitch signaling (CCIS) differs fundamentally from per-circuit signaling in that a dedicated channel servicing several trunks is provided for the exchange of signaling information. Several advantages accrue as a result of this approach. The expensive channel filters required for per-circuit signaling are eliminated and can be replaced with a modem or an all digital interface. Thus, the speed at which signaling information can be transmitted depends on the signaling channel. The problems of "talkdown" and signaling information attenuation are also dramatically reduced.

There are several variations of CCIS exemplified in Figure 11b by two extreme cases: Associated mode and Non-Associated mode. In an associated CCIS, a dedicated signaling channel is provided for every trunk group; under non-associated CCIS, signal distribution points (indicated by squares in Figure 11b) are introduced which provide in effect a separate signaling subnetwork. The transmission

of signaling information under CCIS is typically performed in a store-forward manner; in this regard, packet switching has emerged as one of the more attractive candidate techniques. Clearly, a spectrum of strategies exist within the above two extremes. CCIS requires the existence of stored program control at switches acting as signal interchange points. Both associated and non-associated modes of CCIS are supported in the proposed CCITT No. 6 signaling system.

Since CCIS inherently executes a link-by-link method of connection establishment, the previous distinction between OOC and PRC may be rendered moot. In fact, under the most advanced form of CCIS, the circuit switch model proposed in Figure 4 is not directly applicable. No senders are used for the transmission/reception of destination address information; instead buffers act as the repository for the address signals. Queueing for access to one of several "senders" is replaced by queueing for a single signaling channel. In addition the interswitch protocols are considerably more complex. However, due to the lack of detailed information regarding the interswitch protocol under CCIS, similar assumptions are used as before concerning the connection establishment (Section 5.3.1.1). The major operational differences under CCIS concerning the switch model is a substitution of the multi-server sender queue of Figure 4 by a single server queue for the signaling channel.

Address information is assumed to be encoded according to a fixed format. If the speed of the signaling channel is assumed to be C bps, and the length of the signaling information is denoted as b_1 (b_2) bits for the full (abbreviated) address, the destination address transmission interval becomes (compare with Equations 36 and 37):

$$XMIT(1) = C\ b_1 \tag{38}$$

$$XMIT(2) = C\ b_2 \tag{39}$$

Thus the tandem register occupancy and signal channel utilization can be obtained in a straightforward fashion as follows:

The length of time the signaling channel is held at a particular switch K should be modified to:

$$r_K = \text{XMIT(1) or XMIT(2)} \tag{40}$$

depending on the type of address signaling employed.

The delay encountered at a switch in order to perform tandem signaling over the common signaling channel must be modified from Equation 13 to an equivalent single server formulation (M/D/1) as:

$$C_2(K) = \frac{r_K}{2} \frac{(2 - r_K)}{(1 - r_K)} \tag{41}$$

Similarly, the tandem register holding time becomes:

$$d_t(K) = \begin{cases} \text{XMIT(1)} + C_2(K) + \text{ACK} \\ \text{XMIT(2)} + C_2(K) + \text{ACK} \end{cases} \tag{42}$$

Note that no access request signal (signal A in Figure 9) has to be issued prior to transmission of the destination address information (yielding decreased operational overhead), and the buffer which holds the dialed digits is occupied for the entire switch connection interval pending acknowledgement (ACK) of the successful reception of the address information. The signaling protocol implied by Equation 42 is inherently link-by-link (PRC); a corresponding formulation for an end-to-end control strategy (OOC) can be derived taking into account only the originating traffic (input to the signaling channel queue) and the call routes (as in Equation 31), but is omitted for the sake of brevity. Finally, note that the overlap between the tandem register holding time and cross-office connection setup also require a modification to Equation 15 due to the concurrent execution of both tasks.

5.4 CONCLUSIONS

This memorandum has postulated a generic circuit switch architecture which forms the basis of an analytic model used to obtain various switch-related and network-related performance measures. Due to the vast diversity which exists in both circuit switch architecture and operation, the formulation has been considerably simplified and of necessity has ignored several facets of a given switch's operation. In pursuit of closed form solutions, certain statistical assumptions were made in order to preserve analytic tractability.

The global issues which were addressed in conjunction with switch architecture/operation include:

- Network Routing/Control Strategies
  - Progressive Route Control
  - Originating Office Control
- Network Signaling Techniques
  - Per-Circuit Signaling
  - Common Channel Signaling

For each of the above issues, a closed form solution for the cross network setup delay has been obtained. In the forthcoming experimentation, the model derived herein will be used to investigate switch cost/performance tradeoffs and will also supplement the ongoing circuit switching network design effort.

REFERENCES

[KOTIAH, 1976] Kotiah, T.C., "On Two-Server Poisson Queues with Two Types of Customers", Operations Research, 1970, pp. 597 - 603.

[LEE, 1955] Lee, C.Y., "Analysis of Switching Networks", Bell System Technical Journal, 1955, pp. 1287 - 1315.

[CLOS, 1953] Clos, C. "A Study of Non-Blocking Switching Networks", BSTJ, Vol. 32, 1953, pp. 406 - 424.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
NETWORK ANALYSIS CORPORATION
BEECHWOOD, OLD TAPPAN ROAD
GLEN COVE, NEW YORK 11542

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
None

3. REPORT TITLE
~~EIGHTH SEMIANNUAL TECHNICAL REPORT, MARCH 1977, FOR THE PROJECT~~ INTEGRATED DOD VOICE AND DATA NETWORKS AND GROUND PACKET RADIO TECHNOLOGY ~~(VOLUMES 1 THROUGH 4)~~ Volume I. Integrated DoD Voice and Data Networks. Part 2.

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
EIGHTH SEMIANNUAL REPORT, MARCH 1977

5. AUTHOR(S) (Last name, first name, initial)
NETWORK ANALYSIS CORPORATION (9) Semiannual Technical rept. no. 8,
(10) Howard/Frank Israel/Gitman

6. REPORT DATE
(11) MARCH 1977

7a. TOTAL NO. OF PAGES
746

7b. NO. OF REFS
151

8a. CONTRACT OR GRANT NO.
(15) DAHC 15-73-C-0135,

b. ~~PROJECT NO.~~

c. ARPA ORDER ~~NO.~~ 2286

d.

(12) 2540,

9a. ORIGINATOR'S REPORT NUMBER(S)
SEMIANNUAL REPORT 8 (4 VOLUMES)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES
None

12. SPONSORING MILITARY ACTIVITY
Advanced Research Projects Agency, Department of Defense

13. ABSTRACT New research results on the following major questions are reported: Results on integrated DOD Voice and Data Networks include: analytical models for determining blocking and delay on an integrated link and numerical investigation as a function of traffic and design variables; algorithms for integrated network design were developed and programmed. The program is capable of designing networks for voice traffic, signaling and data traffic. A circuit switch model for determining switch and network transit delays for circuit connection set-up was developed. A methodology for classification of telecommunications routing algorithms was developed. Results on topological gateway placement include an algorithm and program for interconnecting packet switched networks, studies of cost/performance tradeoffs, and an application to interconnect the ARPANET and AUTODIN II. In the packet radio area models were developed to estimate network initialization as a function of number of repeaters, transmission rates of repeaters and station, and operation disciplines. Finally, cost trends for large volume packet switched data networks are derived which incorporate switching and transmission costs, satellite and terrestrial channels and local distribution.

14. KEYWORDS - Computer networks, communication networks, terrestrial and satellite networks, packet radio networks, throughput, cost, delay, blocking, ARPA Computer Network, store-and-forward, packet switching, circuit switching, integrated switching, gateways.

389 161 HB

UNCLASSIFIED
Security Classification