LEVEL II

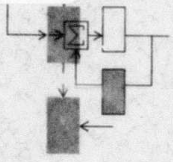
12

November, 1981

LIDS-TH-1161

Research Supported By:

ARPA Contract N00014-75-C-1183
OSP Number 82933



QUEUING ANALYSIS OF A SHARED VOICE/DATA LINK

Daniel Uri Friedman

Laboratory for Information and Decision Systems
MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A108336
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Queuing Analysis of a Shared Voice/Data Link	5. TYPE OF REPORT & PERIOD COVERED Paper (9) Doctoral
	6. PERFORMING ORG. REPORT NUMBER LIDS-TH-1161
7. AUTHOR(s) Daniel Uri Friedman	8. CONTRACT OR GRANT NUMBER(s) ARPA Order No. 3045/5-7-75 ONR/N00014-75-C-1183
9. PERFORMING ORGANIZATION NAME AND ADDRESS Massachusetts Institute of Technology Laboratory for Information and Decision Systems Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Code No. 5T10 ONR Identifying No. 049-383
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209	12. REPORT DATE November, 1981
	13. NUMBER OF PAGES 165 pages (12) 165
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Information Systems Program Code 437 Arlington, Virginia 22217	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider a link shared by some number of off-hook phone callers and a data queue. The allocation of capacity to voice and data depends on the level of speaker activity (number of talkspurt). A Markov chain model is adopted for this activity, and the resulting data queue performance is analyzed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

November, 1981

LIDS-TH-1161

QUEUING ANALYSIS OF A SHARED VOICE/DATA LINK

by

Daniel Uri Friedman

This report is based on the unaltered thesis of Daniel Uri Friedman, submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology in November, 1981. The research was conducted at the M.I.T. Laboratory for Information and Decision Systems, with support provided in part by the Advanced Research Project Agency under Contract No. N00014-75-C-1183.

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

QUEUING ANALYSIS OF A SHARED VOICE/DATA LINK

by

Daniel Uri Friedman

B.S., Rice University

(1976)

S.M. Massachusetts Institute of Technology

(1979)

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

November 1981

Signature of Author Daniel Friedman .....
Department of Electrical Engineering and Computer Science,
November, 1981

Certified by .....
Thesis Supervisor

Accepted by .....
Chairman, Departmental Committee on Graduate Students

QUEUING ANALYSIS OF A SHARED VOICE/DATA LINK

by

Daniel Uri Friedman

Submitted to the Department of Electrical Engineering and Computer Science on November 9, 1981, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

We consider a link shared by some number of off-hook phone callers and a data queue. The allocation of capacity to voice and data depends on the level of speaker activity (number in talkspurt). A Markov chain model is adopted for this activity, and the resulting data queue performance is analyzed.

Thesis Supervisor: Robert G. Gallager

Title: Professor of Electrical Engineering

ACKNOWLEDGEMENTS

I take this opportunity to thank my thesis supervisor, Prof. Robert Gallager, and my thesis readers, Profs. Pierre Humblet and Wilbur Davenport, Jr., for their guidance and instruction, both in connection with the thesis, and, more generally, throughout my graduate study. I also thank Prof. Julian Keilson, Dr. Mark Zachmann, and Dr. Ushio Sumita, of the University of Rochester, for many helpful discussions and insights, especially regarding the material in chapters V and VI.

Financial support was provided by a fellowship from the John and Fannie Hertz Foundation and also by some funding from DARPA grant No. 82933 and NSF grant No. 89082; I gratefully acknowledge their aid.

The report was typed by Mrs. Fran Frolik, and I thank her for her patient assistance in this regard.

To my parents - Wittgenstein's dictum (or my interpretation anyway) still applies.

TABLE OF CONTENTS

I. INTRODUCTION

A. Motivation and Perspective

This thesis is concerned with some problems that arise in connection with the statistical multiplexing of voice and data. The motivation derives from the observation that conversational speech actually consists of alternating periods of sound production and silence. The most readily identifiable source of this behavior is the alternation of dominance that naturally occurs as two parties converse. However, even the speech of the party considered the "talker" is not a continuous stream of sound if one looks on a finer time scale. Rather, it too consists of an alternating sequence of talkspurts (which can comprise a few words or phrases) and silences. On even finer time and frequency scales, one can observe that a talkspurt itself does not always occupy the full "long-term" speech bandwidth, which is approximately between 0 and 4000 Hz.

The success of any capacity sharing scheme that exploits these characteristics depends on a variety of physical and statistical considerations and their interplay. For the sake of discussion, consider a model system in which N speakers share C_v "units" of capacity. Each speaker alternates between an active state, in which one unit of capacity is required, and a silent state in which no capacity is

required. The system operates as follows. Each speaker is connected to an activity detector and a switch. When the detector observes a transition from silence to activity, the switch connects the speaker to a unit of capacity, if it is available. The connection is maintained until silence occurs. If capacity is not available, the speaker is simply locked out and joins a pool of other active, waiting speakers. This pool is served on a first come-first serve (FCFS) basis as capacity becomes available. During the lock out period, speech is lost, and a speaker who becomes silent while waiting, departs the pool.

One useful performance measure of this system is the average loss or cut-out fraction ϕ, defined as

$$\frac{\text{Time Averaged Loss}}{\text{Time Averaged Offered Load}}$$

The value of ϕ is determined by what are effectively the physical and statistical "response times". The physical response time is characterized by the switching time, τ_s, and τ_d, which is the "window" that the detector needs to accurately track activity. For successful operation, it is clear that $\tau_d + \tau_s$ must be much smaller than the average active time and average silence time (τ_{AC} and τ_{SL} respectively). If this condition is met, one must look at the "statistical" loss incurred in waiting for capacity. This depends on N, C_v, and the statistical behavior of the speakers. In the ideal case of

$\tau_s = \tau_d = 0$, and with suitable assumptions about the activity process, one can show

$$\phi = \frac{\sum_{\ell = C_v+1}^{N} (\ell - C_v) p_\ell}{\sum_{\ell=1}^{N} \ell \, p_\ell}$$

where p_ℓ is the "equilibrium" probability that ℓ speakers are active. See Weinstein [1]. We will elaborate on this later.

In the above model, a unit of capacity depends on the context. On the talkspurt-silence level, it is the capacity to handle all the "bit rate" of an encoded talkspurt. On a finer scale, it may refer to some sub-band of the full speech bandwidth, in which case, the loss is only for that sub-band. The physical problems of frequency sub-band multiplexing are formidable because of the small times involved. That is, the short-term bandwidth of a talkspurt moves around rapidly within the long term spectrum so that the detection problem is hard. However, activity detection on the talkspurt-silence level is quite feasible and has been implemented. For example, in the late 1950's, Bell Telephone built the TASI (an acronym for time assigned speech interpolation) system for use on transoceanic cables. Thus, the ratio N/C_v is referred to as the <u>TASI advantage</u>. Although the original system was in an analog environment, we will use TASI as a generic term for any such statistical multiplexing scheme. All future discussion

will pertain only to the talkspurt-silence level.

TASI is successful because little speech is lost during detection and switching. (Typically, τ_s and τ_d are on the order of 10-20 ms whereas τ_{SL} and τ_{AC} are on the order of 1 s.) In the next section, we will see that digital switching between voice and data on these time scales is no harder, and one can consider transmitting data during the silences as well. This is the main topic of the thesis. Because voice traffic must meet certain rather stringent delay requirements, data must have a lower priority to some degree. Thus, the data queue effectively sees a server whose rate is strongly governed by speaker activity. The main question will be whether voice activity returns from "high" levels to its "mean" level sufficiently rapidly, so that data backlogs which accumulate while voice activity is high, can be emptied in "reasonable" time. Before delving into this, we will spend the next few sections discussing some implementation considerations and voice activity models.

B. Network and Implementation Considerations

In a network setting, the previous analysis applies only to an end-to-end or, conceptually, a single-link version of TASI. That is, if N callers at a node A share C_v "channels" that transmit their speech to B, and each channel can only accommodate one active speaker, then the cutout fraction ,ϕ, is the fraction of output that does not reach B. Since all allocation decisions are made at A, the physical realization of these channels is not important in the analysis; they can be viewed as a single link of equivalent capacity. In a real network, connections between users often comprise multihop paths, and a given link is usually a part of paths between many sources and destinations. Thus, one can conceive of "network TASI" in which the output of a caller can be preempted at any node on the path it follows, and all nodes cooperate in globally allocating capacity.

With analog transmission and electromechanical switching (i.e. relays) or even digital transmission and switching with semiconductor logic gates, the nodes have neither the time nor the processing power to make the necessary decisions. Therefore, early TASI systems were indeed single-link operations used to increase the "virtual" voice capacity of relatively expensive backbone trunks (such as transoceanic cables). With digitized speech and the current "software" switching technology, network TASI is possible to implement. In fact,

digitized speech, with activity detection, may be viewed as another type of bursty "data" traffic whose arrival statistics and delay requirements are different from those of conventional types. Thus, the network TASI problem is only part of an integrated voice/data network's overall allocation problem.

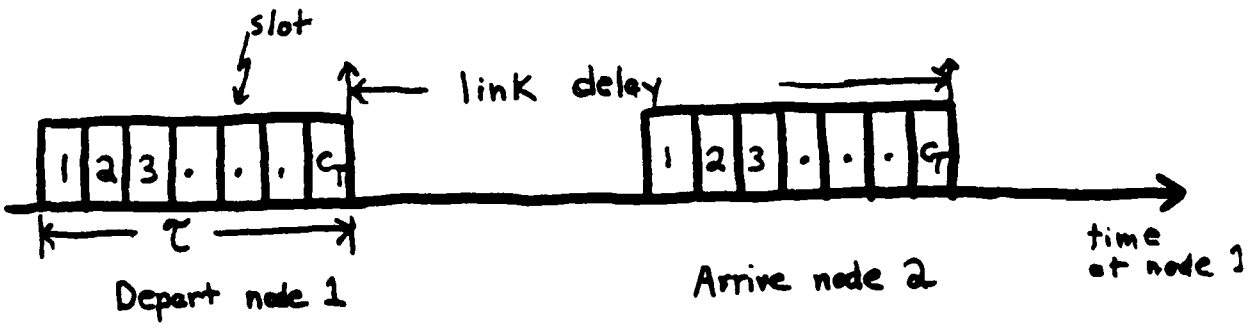
Network resource allocation problems are difficult, and one usually cannot conduct a detailed queuing or loss analysis. Instead, one often attempts to separate the "probabilistic" from the "networking" issues. For example, one might explore the "networking" aspects of minimum delay routing problems by assuming that the average queuing delay (at node i) of traffic using link (i,j) depends only on the link capacity and average flow on the link. Here, a "networking" question is, for example, how should the nodes cooperate to find best routes given that each initially knows only the flows on its links? A single link or tandem links queuing analysis can be used to explore the validity of the assumption, i.e. can higher order moments of the flows be neglected in computing average delay, can statistical dependencies between queues be neglected, etc.

In this approach, prior knowledge of the particular network architecture or quantity of interest can sometimes be used to tailor the single link model so that one can focus on specific issues. This is not done in the thesis. That is, we will use a general model of a single voice/data link and analyze a variety of quantities that might later be used in network approximations. Nevertheless, it is helpful to first

have a qualitative understanding of the transport requirements of voice and data and of various switching disciplines.

Our discussion of these issues will be conducted in the context of the following time-division multiplexed (TDM) switch architecture and transmission format. Where we give values for certain parameters, these values reflect our understanding of the capabilities of current technology. This has come primarily through "private discussion", so we do not provide specific references. At the end, we do indicate some tradeoffs affecting the choices of values for these parameters. A more detailed technical survey and bibliography can be found in [2].

In the TDM architecture, time is divided into units of length τ called <u>frames</u>. (Typically, $\tau \approx 10\text{-}50$ ms). The "atomic" unit of transmission is the <u>block,</u> which consists of b bits. We say that a link has capacity C_T blocks/frame if the associated node→link→node combination has the processing and transmission capacity to handle C_T blocks/frame on a pipelined basis. The meaning of this can be understood with the aid of the following diagram.



During every frame, node 1 "enters" $C_T b$ bits into the channel. Consecutive groups of b bits are viewed as blocks occupying slots in time. Block departures at node 1 are taken as point events,occurring when the last bit of a block enters the channel. The arrival process at node 2 is viewed simiarly, and an event occurs when the last bit of a block is in node 2's memory and ready for further transmission. The time between block departure and arrival is called the link delay. (We assume link delays do not change in time.) In practice, this might consist of physical propagation time and some processing time to get the bits into memory. If the link delay is larger than τ , there will be at least one complete frame of blocks in the pipeline at any instant. This will usually be the case on a geosynchronous satellite link because the round trip propagation time is about .25 sec. (The altitude of the geosynchronous orbit is about 23,000 miles.) By "pipelined on a block basis", we mean that the first bit of a block cannot leave a node until the last bit of that block has arrived. Thus, the first bit incurs a delay which is the sum of the link delay and the duration of a slot.

In a network of these switches, this slotted frame format is used on every (directed) link, though we do not assume that frame boundaries are globally aligned. τ and b are the only global constants. We do assume for now that links are noiseless.

Sources are connected to the network through a host node,

and the host and source communicate through a source buffer. At the destination node there is a decoder which uses the arriving bits to reproduce the source's messages for some "final" user. We consider the decoder to be outside the network. The time that a source is connected to the network is called the session. Two types of users, "voice" and "data", are considered.

Note: The callers of a two-way conversation are treated as independent sources, and each speaker's output is viewed as a sequence of "one-way" messages. To allow them to sustain normal conversation, the network must meet certain delay requirements. Other than this it does not "recognize" them as interacting users. These requirements will be discussed later. All data messages are "one-way".

A data source places bits into the source buffer in an arbitrary manner. These bits are viewed as a sequence of messages as prescribed by the user.The network can transport individual bits as it chooses, as long as the following requirements are met.

1) No loss - All bits must be delivered to the decoder in correct order.
2) The network must separate messages for the decoder.
3) End-to-end delays of messages must meet some (possibly statistical) requirements. The delay of a message is the time between the entry of its last bit into the

source buffer and the delivery of this bit to the decoder.

Speech is digitized at a rate b/τ bits per sec. The encoder also performs activity detection. That is, a block is placed into the buffer at the end of a frame only if the encoder determines that the speaker was active during that frame. (We view this placement as a point event. Also, frame boundaries at the encoder and host node need not be aligned.) We assume that speech loss caused by incorrect activity decisions is negligible. The first block of a talkspurt is marked and contains a number indicating the duration (in frames) of the preceding silence. (Obviously the actual speech digitization rate must be reduced slightly to accommodate such overhead. We neglect this.)

Characterizing the delay requirements of voice is difficult because the "message" is not clearly defined. Theoretically, speakers could communicate via a sequence of "voice telegrams", where each telegram contains "a thought" and can be delayed a few seconds. This is not the same as "normal" conversation, in which speakers implicitly use silences to separate "messages". Approximately 250 ms is usually given as the maximum acceptable delay (the time between the beginning of a talkspurt at the source and the beginning of its reproduction by the decoder). With larger delays, speakers "collide" and must resort to explicit phrases to separate "messages".

In the single-link TASI system first described, all loss

occurs in some initial segment of a talkspurt. This results in clipping that, apparently, is not noticed if the average loss is below .5%. (See Weinstein [1].) We will see that with "network TASI", loss can occur at various places in a talkspurt, and speakers might tolerate larger average loss provided it is scattered. Although this is an "experimental question", we note that if the "loss vs maximum TASI advantage" curve is relatively flat, small increases in acceptable loss result in relatively large increases in the maximum TASI advantage. The single link "loss vs TASI advantage" curve will be discussed in I.D.

We now discuss the implementations of three "standard" switching disciplines within the TDM architecture and their uses with voice and data sources.

1. Circuit Switching - Conceptually, a circuit is a guarantee of a path of specified capacity from the source to the destination for the entire session. The network attempts to establish or set up a path when the user arrives. If it cannot do so within "reasonable" time, the user is rejected. Therefore, circuit-switcher networks are designed to meet a rejection probability requirement.

Historically, circuits were implemented using analog transmission and "hardwired" connections at switches. In the TDM architecture, a "unit" capacity circuit is a guarantee of a slot in every frame at every node along some path between the

source and destination. At set-up time, the nodes establish tables with entries of the form "slot i on incoming link x corresponds to slot j on departing link y". The particular choices for i and j are unimportant, but they must remain fixed while that path serves the circuit. Note that no addressing is required because the correspondence between circuit and path implicitly identifies the source and destination. (Intermediate nodes need not even know who the end users are.)

With analog circuits, the end-to-end delay is a sum of the link delays, and the delay is the same for each increment of the input signal. With TDM circuits, there can be buffering delay because a block must wait for the appropriate departing slot at each node. The exact value of this delay at a particular node depends on the particular incoming and departing slots serving the circuit and the relative alignments of frame boundaries. But because the next occurrence of the appropriate slot must be within τ secs after the arrival of a block, this delay is at most τ at any node. Further, this delay is the same for all blocks since the slot correspondences are fixed. As link delays are also constant (with time), it follows that all blocks incur the same end-to-end delay. In this sense, circuit switching offers synchronous service.

If a source is circuit-switched, the host node looks in the source buffer every τ sec and removes a block, if at least b bits are present. Otherwise, the slot remains

"empty". Notice that the host node must transmit "something" during empty slots. That is, the bits that are in the slot will eventually reach the destination node unless the host indicates otherwise. (If the decoder is also operating synchronously - i.e. expecting a new block every frame - the "idle message" must also reach the decoder. If it is operating asynchronously - i.e. it "wakes up" only when the destination node indicates a new block has arrived - the idle message must only reach the destination node). In a simple approach, the network can reserve one of the possible 2^b bit patterns to indicate an empty slot. However, if idles occur frequently, this is an inefficient source code (in an information theoretic sense). One example of a potentially more efficient strategy is the following. The first bit of a slot is used as a flag. When an idle occurs, the host node sets the flag, and it can then fill the remaining bits with other information. Of course, some identifier or address must be provided for the new information. This procedure is repeated at successive nodes on the path. Notice that it is really a "node to node" strategy. Different nodes can use the empty slot in different ways as long as the idle flag reaches the destination. If some node actually has nothing else to send, this idle must also be indicated, but it is of concern only to the next node on the path. The flag only encodes source idles. The cost of this strategy is 1 bit per slot, and the average gain is ab^1,

where α is the average idle fraction and b^1 is the number of bits per slot available for user information (i.e. aside from the bits required for protocol).

Circuit switching can serve either type of user. The bits of a data user do arrive in correct order at the destination though some protocol must be used to separate messages (We do not discuss this.) With a voice source, the decoder receives either a new talkspurt block or an idle every frame. In the latter case, it presumably reproduces a silence. Because the service is synchronous, the decoder does not need to use the silence information included in the first block of each talkspurt.

2. Store and Forward Switching - Conceptually, the user only has a "promise" of future delivery. Capacity is allocated on a link by link basis, and the information can, in principle, be indefinitely buffered at any node.

This definition obviously leaves many things unspecified. For example - should messages be broken into packets; how large should packets be; how should routes be chosen? A complete discussion of these questions is not appropriate here. For us, the important feature of store and forward switching is that different parts of a message can incur different delays. To focus on this, we consider a version of packet switching in which each packet occupies one block. That is, the source output is segmented into blocks, and each block or packet

traverses the network on a store and forward basis. (We do assume that blocks arrive in correct order.) However, we consider this version of packet switching with the following "caveat" in mind.

Caveat: With the TDM architecture, one must be careful to distinguish between the switching discipline and transmission format. The slotted frame structure is not needed for "packet" switching. To the contrary, it is only used to maintain the synchronous service of circuit switching. The bits in the slots that are not reserved for circuits effectively constitute a separate "virtual" binary channel, which is interrupted when reserved slots occur. In principle, the nodes can use this channel to implement a variety of store and forward schemes. We have assumed that the store and forward switched traffic respects the slot structure only for convenience. Note that one "cost " of this is that every idle is at least one slot long, i.e. packet transmission cannot begin in the middle of a slot.

It is evident that the delay variability of packet switching is not a problem for data users since the "message" is in the bits themselves. Thus, the relevant tradeoff is between line utilization (or scheduling flexibility) and overhead (addressing). A bursty user wastes part of the capacity of a circuit, but packet switched blocks require addresses because their transmission is not "prescheduled". Notice that even though the empty slots of a circuit can be

encoded with a 1 bit flag(thereby releasing the remaining bits) the network has no control over the occurrence of these slots. That is, it does not have the scheduling flexibility of packet switching.

Delay variability does cause problems for voice. Once the decoder begins reproduction of a talkspurt, it looks for a new block every frame. If a block is late, the decoder might try several things, e.g. "stretch" the previous block. For simplicity, we assume that a late block is unusable and results in speech loss. Now consider a talkspurt that lasts M frames, and suppose the blocks incur delays $d_1,..,d_M$. If the decoder begins reproduction when the first block arrives, then block i is lost if $d_i > d_1$. With this approach, substantial loss might occur if the first block is "lucky", i.e. if d_1 is much less than the average delay. Therefore, the decoder might want to deliberately postpone reproduction for a time d so that subsequent blocks have more time to arrive. Thus, there is an end-to-end delay vs. loss tradeoff. (A practical problem with this scheme is that the decoder generally cannot know d_1. Since it is the total delay, i.e. $d_1 + d$, that matters to the speaker, choosing d is difficult. If the decoder does have some knowledge of the delay distribution, it might be able to make a "reasonable" guess, e.g. choose d s.t. $d + d_1 < 250$ ms with high probability.)

Delay variability can also cause distortions in the durations of silences since the initial blocks of successive

talkspurts need not incur the same delay. When the first block of a talkspurt arrives, the decoder knows what the duration of the preceding silence is supposed to be since this information is contained in the block. Presumably, the decoder also knows when it last reproduced a talkspurt block. Thus, if the first block of a talkspurt arrives before it is needed (i.e. before the appropriate length silence has occurred) then the decoder can postpone reproduction in order to recreate the appropriate silence. (It might want to add even more delay to give subsequent blocks of the new talkspurt more time to arrive, as we have discussed.)

In summary:

- Data can use either packet or circuit switching. With either discipline, all bits do arrive correctly and in order. Some message separation protocol is required. The basic tradeoff is between line utilization and overhead (addressing).

- Voice can use circuit switching, and end-to-end delay is the only "distortion" relative to face-to-face conversation. Packet switching, or what we have termed "Network TASI", is also possible. In this case, the network is free to discard or delay blocks in any way as long as the end-to-end delay is less than about 250 ms, and the loss (due to outright discard at intermediate

nodes or delay variability at the decoder) is acceptable.

3. Fast Circuit Switching - In this case, the user is provided a circuit only during "active periods". The terms"fast" and "active period" are obviously contextual. To discuss some trade-offs, we adopt the following conventions:

1) The "session" is defined by the user, i.e. it is the time that the user wants access to the source buffer.

2) The source can place at most one block into the buffer in any frame.

3) An active period is a sequence of (consecutive) frames in which the source does enter a block.

4) The duty factor is the percentage of frames during which the source is active.

5) A session contains "many" active periods.

For voice, active periods and talkspurts coincide. For data, we have adopted the view that the source "meters" out a message at a rate of one block/frame. The host node initiates a circuit acquisition at the beginning of each active period.

We have seen that the choice between circuit and packet switching is based on a tradeoff between utilization and overhead. Fast circuit switching is another way of increasing the utilization for bursty (i.e. low duty factor) sources.

From the user's viewpoint, fast circuit switching is acceptable as long as the network can set up the circuit and deliver the first block of the active period within an acceptable time. (We assume that blocks are buffered until the circuit is set-up.) Notice that fast circuit switching does not pose a delay variability problem for talkspurt reproduction because once the first talkspurt block arrives, subsequent blocks arrive at a steady rate of one per frame. However, it can cause silence distortions because the circuit set-up time can differ for successive talkspurts. If both packet and fast circuit switching can provide acceptable service to a bursty user, the choice for the network depends on the relative overhead costs of packet and fast circuit switching and the "nature of the bustiness". To understand this, we first need to examine the costs of circuit set-up.

Circuit set-up algorithms build paths on a link by link basis. The capacity reserved on a partially completed path is not available for other circuits while the algorithm tries to extend the path. In practice, the algorithm might backtrack if it cannot extend some partial path. However, even if some link does not end up in the final path, it is not released until the backtracking actually reaches it. For attempt rates below some "critical range", most circuit requests are successfully completed, and the throughput increases as the attempt rate increases. (Throughput is number of ccess/unit

time.) Beyond this critical range, the network's resources are increasingly consumed by partially completed paths,and the probability of successful completion decreases. In this region, throughput decreases as the attempt rate increases.

Now as a measure of "burstiness", the duty factor is simply the long-term, average activity fraction, and a given duty factor can be obtained by many combinations of activity frequency and average duration of active periods. From the previous discussion, it should be clear that, for fixed duty factor, the overhead of fast circuit switching decreases relative to the overhead of packet switching, as the "burstiness" tends to the "infrequent but long active period" type. For example, a user wishing to make several large file transfers in a session might be more efficiently served by fast circuit switching. (Of course, as the duty factor itself increases, "ordinary" circuit switching, i.e. providing a circuit for the entire session, becomes relatively more attractive than either packet, or "fast circuit" switching.)

Variable Rate Speech Coding

So far, we have assumed that the decoder uses all b bits of a block to reproduce a frame of speech. With variable rate coding [3], speech is digitized so that a full block contains information to reproduce speech at several possible levels of fidelity. That is, a block actually contains, say, b/n sub-blocks, the slot size is reduced by n , and speech quality increases as the number of sub-blocks used in reproduction increases. This type of coding might be useful in congestion control because by sending fewer sub-blocks, the network can reduce delay but maintain the "continuity" of the conversation at a lower quality. Presumably, this is preferable to total loss of blocks or excessive delays.

Speech Digitization Techniques

1) Standard "toll quality" speech uses 64 Kbps PCM with a sampling rate of 8000 Hz and 8 bit quantization.

2) Differential PCM - Transmit the difference between successive samples. We have seen references to 16 Kbps and 32 Kbps DPCM system.

3) Linear Predictive Coding (LPC) - In this approach the vocal tract output during a frame is modelled as the output of a linear time invariant system of some order k.The encoder examines the speech during the

frame and uses its observations to generate values for the k system coefficients that result in a "best fit" of the observed samples to the model. These values and various other parameters are then quantized and transmitted. Using LPC, it is possible to achieve intelligible speech with transmission rates as low as 1 - 2 Kbps.

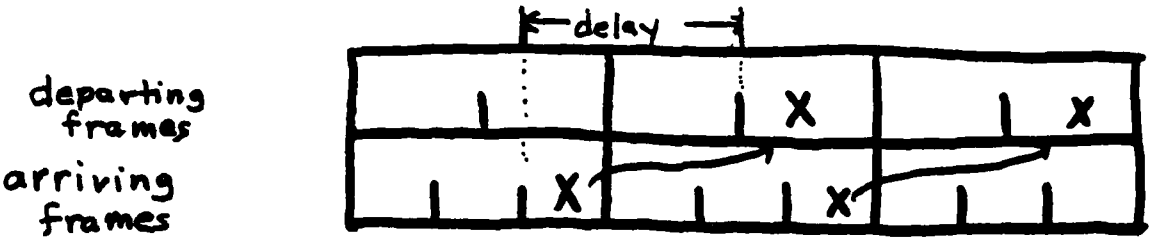
Noisy Links

Channel noise is combatted by error correcting coding or error detection and retransmission. The latter is particularly troublesome for voice traffic because it increases both delay variability and average end-to-end delay. Thus voice traffic must usually accept the noise immunity provided by error correction. This is an important consideration in the choice of the digitization technique. (64 Kbps PCM is relatively insensitive to errors; LPC speech with 1 - 2 Kbps rates is sensitive to errors.)

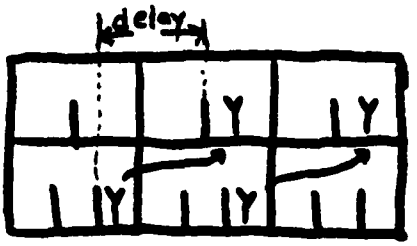
Choice of τ and b and Circuit Delay

Recall that TDM circuit switching introduces buffering delays because a block must wait for its slot on the departing link. This delay at a particular node depends on the choice of slots serving the circuit, but for a given choice, it is proportional to τ and b. For example, one can provide an equivalent capacity circuit by using a block size of b/2 and a frame size of $\tau/2$. For a given block of b bits at

the source, this circuit will impose only half the buffering delay on the leading edge of the block.



In the diagram, we have shown two arriving frames and the corresponding departing frames. (The departing link has a lower raw bit capacity so slots are wider, but all slots contain b bits.)



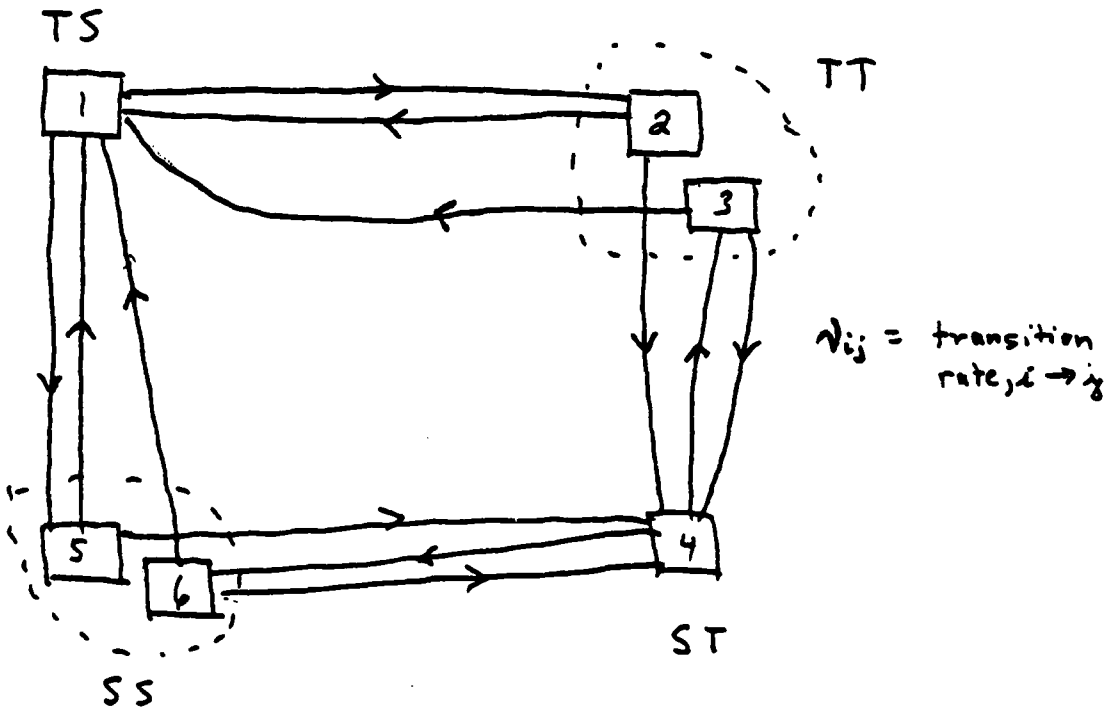
Here the frame time and block size are halved, and the buffering delay of the leading edge is also halved. Essentially,

this is a different interleaving of the b bits in the given source block. This can reduce the overall end-to-end delay as the following example shows.

We are concerned with the time between the beginning of a talkspurt at the source and the beginning of its reproduction at the destination. With PCM speech, the encoder can produce digitized speech on a sample by sample basis, but if activity detection is desired, it cannot release the first sample until it has looked at the output for some more time. The minimum time needed for this depends on the sensitivity of the tracking algorithm, but it is generally greater than the time between samples (i.e. $> \frac{1}{8000} = .125$ ms). Thus the activity detection process requires a frame structure at the encoder with some minimum frame time τ_E. (This adds a delay τ_E.) Suppose that within this time the encoder produces b_E bits. Now the PCM decoder does not need all b_E bits to begin reproduction. It can essentially work on a sample by sample basis, i.e. its minimum block size is 8 bits. Thus, if the network transmits the encoder's block using a network frame time of $\tau_E/2$ and block size $b_E/2$, the first sample of the talkspurt incurs less buffering delay in traversing a given circuit. (Of course, the other components of circuit delay, namely the link propagation delays, are not reduced by this.)

C. A Speaker Activity Model

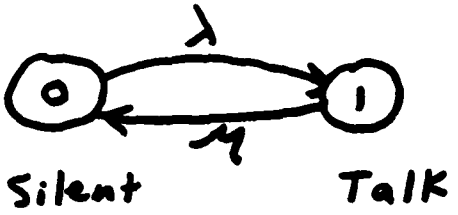
In [4], Brady introduced a continuous time Markov chain model of the activity of a single speaker A engaged in a conversation with another speaker B. The four possible combinations of activity are denoted by TT, TS, ST, SS; where T = talk, S = silence, and the first letter in a pair refers to A's state. His model has the following state diagram.



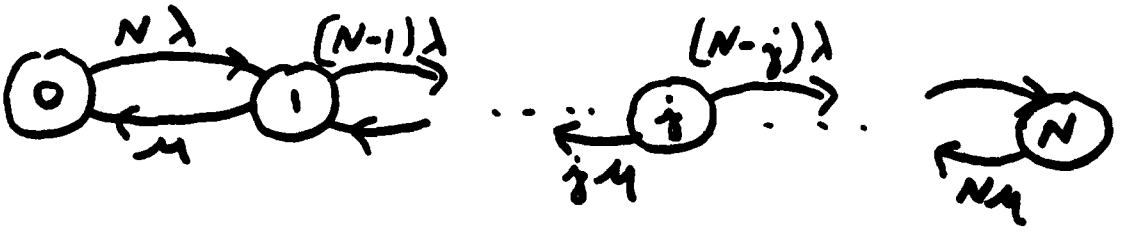
Note that the combinations TT and SS are split to introduce more memory. For example if TT is entered from TS, one might guess that a return to TS is more likely than a passage to ST. Presumably, A is dominant and B briefly interrupts.)Hence,

one would guess $\nu_{21} > \nu_{24}$.

Our objective is to model the activity of N independent speakers at one end of a set of N independent conversations between two sites. At this point, we could extract the marginal behavior of A, say, and extrapolate to N speakers. (By independence, the joint process would be a product process.) This leads to considerable complications. For example, A is in talkspurt whenever the chain is in states 1, 2, or 3. The time that a Markov chain "sojourns" in a subset of its states is generally not easy to characterize in a "closed form". The following model has been proposed by Weinstein [1], and we adopt it. Each speaker is modelled by a two state chain.



From the independence of the speakers and properties of Markov chains, it follows that the process A(t) = number active at time t is characterized by a birth-death chain.



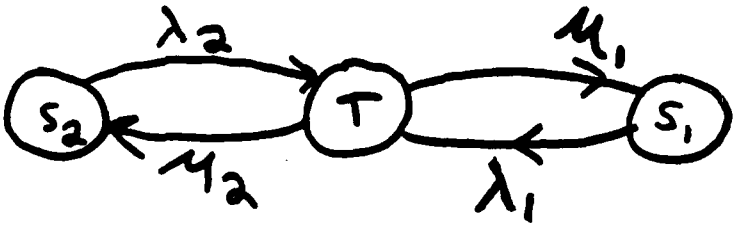
From this, one can see that the ergodic distribution of A is binominal on N trials with "Pr (success)" = $\frac{\lambda}{\lambda+\mu}$. So the mean activity level is $\bar{A} = N \frac{\lambda}{\lambda+\mu}$, and the standard deviation is $\frac{(N\lambda\mu)^{1/2}}{\lambda+\mu}$.

As representations of the capacity demand of blocked, digitized speech, these models have the obvious defect of using continuous time. But given that the frame time τ is much smaller than most active periods and silences, this is not a serious problem. (We could use discrete time with geometric distributions and obtain similar results.) A more substantial question is whether talkspurts and silences can be approximated by exponential distributions. (We adopt the notation $x \sim F(\cdot)$ to mean that F is the distribution function of x. The notation $\exp(\lambda)$ refers to the exponential distribution with parameter λ, i.e. $F(x) = 1 - e^{-\lambda x}$. The mean is then $\frac{1}{\lambda}$.) Brady indicates that the exponential distribution is reasonably good for talkspurts and suggests this is because most talkspurts are what he terms "solitary talkspurts", i.e. those that begin in TS and end in SS without any passage through TT. Now for the Brady chain, given that a sojourn in state 1 ends with a transition directly to 5, the duration of time spent in state 1 $\sim \exp(\nu_{15} + \nu_{12})$. (For any finite Markov chain, the aposteriori distribution of the exit time from a state i, given that the transition was to j, is independent of j and $\sim \exp\left(\sum_{\substack{j \\ j \neq i}} \nu_{ij}\right)$.) Of course, the fact that

most talkspurts in the Brady model are exponentially distributed does not by itself imply that real talkspurts can be so approximated. The latter assertion must be empirically verified. The only point is that if the Brady model is accurate and most talkspurts are solitary, then the exponential approximation will also be good.

Brady also indicates that silences are not well modelled by an exponential distribution and suggests that this is because there are really two types of silences -- long silence occuring when a party is listening and shorter silences punctuating the speech of a dominant talker. But we are really interested only in the behavior of the aggregate process A(t) and not in individual speakers. The following "plausibility arguments" indicate that the birth-death model for A(t) is reasonable to use in the case of "large N", even if the 2-state single speaker Markov model is not good.

First consider the following, more refined, single speaker model.



Here there are two silent states S_1 and S_2 and, to attempt to capture the desired behavior, one might assume $\mu_1, \lambda_1 >> \mu_2, \lambda_2$.

The relative equilibrium probabilities (i.e. unnormalized) are $P(T) = \lambda_1 \lambda_2$, $P(S_1) = \lambda_2 \mu_1$, $P(S_2) = \lambda_1 \mu_2$. The aggregate vector process for N speakers, $[T(t), S_1(t)]$, where $T(t)$ is the number in T and $S_1(t)$ is the number in S_1, is a Markov process. When $T = i$ and $S_1 = j$, T is being driven to $i - 1$ by a Poisson process of rate $i(\mu_1 + \mu_2)$ and to $i + 1$ by a Poisson process of rate $j_1 \lambda_1 + (N-i-j_1)\lambda_2$. Now we make two assumptions for large N.

1) T rarely becomes very large (large is for example $N - 0(1)$ or $N - 0(\sqrt{N})$) so that $N - T$ is also large, i.e. T and N-T are both $0(N)$.

2) Given that $N - T$ is large, the relative populations in S_1 and S_2 can be replaced by their equilibrium relative mean values, i.e. $S_1 = (N - T)\dfrac{p(S_1)}{p(S_1) + p(S_2)}$,

$$S_2 = (N - T)\frac{p(S_2)}{p(S_1) + p(S_2)} .$$

With these assumptions, we are asserting that the (non-Markov) marginal process $T(t)$ can be approximated by a birth-death process of the type used for $A(t)$. That is when $T = i$, it is driven to $i - 1$ by a Poisson process of rate $i(\mu_1 + \mu_2)$ and to $i + 1$ by a Poisson process of rate

$(N - i)\left[\lambda_1 \frac{p(S_1)}{p(S_1)+p(S_2)} + \lambda_2 \frac{p(S_2)}{p(S_2)+p(S_1)}\right]$. Alternatively we are asserting, that the replacement of the three state model by a two state model (for a single speaker) with $\mu \leftarrow \mu_1 + \mu_2$, $\lambda \leftarrow \frac{\lambda_1 p(S_1) + \lambda_2 p(S_2)}{p(S_1) + p(S_2)}$ is valid in the aggregate.

As another approach, we might attempt to extrapolate a limit theorem from renewal theory to alternating renewal processes. This theorem is as follows (See Feller [5] p. 370.) Consider the process obtained by "merging N independent renewal processes, i.e. look at the collective sequence of renewal epochs. Suppose the mean renewal time for the k^{th} process is $\frac{1}{\mu_K}$ and that the individual processes are "rare", i.e. no single process contributes greatly to the merged process. Set $\alpha = \mu_1 + \ldots \mu_N$. Then in the "steady state", the waiting time for the next event in the merged process (which is not a renewal process in general) is approximately distributed as exp(α).

Now consider N identical, alternating renewal processes; say each process has two states 0 and 1. Suppose the renewal time in state 0 has distribution function F_0 with mean $\frac{1}{\lambda}$, and for state 1, these are F_1 and $\frac{1}{\mu}$ respectively. Let M(t) be the merged process, M(t) = number in state 1. We would like to assert that M(t) can be approximated by the birth-death model. More precisely, we would like to show that

when $M = i$, the time until one of the i processes in state 1 goes to 0 is approximately $\sim \exp(i\mu)$, and analogously for the $N-i$ processes in state 0. (Note that this is not the assertion that the time until the next 1→0 transition is $\sim\exp(i\mu)$. The next 1→0 transition, i.e. the next time M decreases, may result from one of the (N-i) processes in 0 going to 1 and back to 0.) The derivation of the limit theorem rests on the fact that, in steady state, the waiting time for the k^{th} renewal process has its so called residual lifetime distribution. (This distribution is defined as follows. Suppose one examines a renewal process at some time t and asks for the distribution of time Y_t until the next renewal. Then one can show that as $t\to\infty$, the density of Y_t approaches $\frac{1 - G(y)}{m}$ where G(y) is the renewal distribution and m is its mean, $m = \int_0^\infty y\, d\, G(y) = \int_0^\infty (1-G(y))dy$.)

To extend this limit theorem we would have to show that in the steady state and given $M = i$, the 1→0 waiting times for the i processes in state 0 are distributed as the residual lifetime distribution associated with $F_1(x)$. (and analogously for the i processes in state 0). Of course, once a transition occurs, say a 1→0 occurs first, the particular process that changed joins the other N-i processes in state 0, and its renewal time distribution is just the original $F_0(x)$ rather than the residual lifetime distribution. However, for those i such that i and $N-i$ are both large, its effect on the total might be negligible. That is, we might always be

able to treat the processes as if their renewal times are distributed as the respective residual lifetimes.

Note that if we count states S_1 and S_2 as a superstate in the three state single speaker model, the resultant two state processes is an alternating renewal process with one renewal time $\sim \exp(\mu_1 + \mu_2)$ and the other $\sim \left[\frac{\mu_1}{\mu_1+\mu_2} \exp(\lambda_1) + \frac{\mu_2}{\mu_1+\mu_2} \exp(\lambda_2)\right]$. Suppose that state T is state 1, and the silent superstate is state 0. If the generalization of the limit theorem is correct, then when M = i, the distribution of time until the next transition of one of the (N-i) silent speakers can be approximated by $\exp(\alpha)$ where $\alpha = (N-i)\left[\frac{1}{\mu_1+\mu_2}\left(\frac{\mu_1}{\lambda_1} + \frac{\mu_2}{\lambda_2}\right)\right]^{-1} =$

$= \frac{(N-i)}{p(S_1)+p(S_2)}\left[\lambda_1 P(S_1) + \lambda_2 P(S_2)\right]$. This is the same as the previous formula. That is, we can view the initial approximation made by replacing the number in S_1 and S_2 by their respective means (given that the number in T = i) as a special case of this limit theorem. (Since F_1 is exp $(\mu_1 + \mu_2)$, the distribution $\exp(i(\mu_1 + \mu_2))$ is exact for the speakers in T. This is because the residual lifetime distribution associated with an exponential distribution is the same exponential distribution.)

Weinstein [1] does indicate that for N > 25-30, the birth-death model for aggregate activity appears to be as good as the Brady model, in the sense that if one compares simulations

of the birth-death model on N speakers and N independent, marginal Brady speakers, the empirical distributions are similar. (By N independent, marginal Brady speakers we mean run simulations of N Brady models and examine the process M(t) = number of "A speakers" in talkspurt at t .)

D. The Loss Fraction

For the birth-death model of A(t), the transition probability $\Pr[A(t) = \ell | A(0) = i]$ approaches a limit p_ℓ that is independent of i. This p_ℓ is called the equilibrium or ergodic probability of state ℓ (See Chapter II) and is given by $\binom{N}{\ell} \left(\frac{\lambda}{\lambda+\mu}\right)^\ell \left(\frac{\mu}{\lambda+\mu}\right)^{N-\ell}$. It admits an interpretation as the limiting fraction of time the chain spends in state ℓ. This expression actually applies to more general models of speaker activity. For example, if each speaker is modelled as a two-state alternating renewal process with mean silence $\frac{1}{\lambda}$ and mean talkspurt $\frac{1}{\mu}$, and if the notion of an equilibrium is well defined for this process, then the binominal distribution for $A(\infty)$ follows from independence. See Weistein [1]. Once expressions for the $\{p_\ell\}$ are obtained, one can substitute into the formula

$$\phi = \frac{\sum_{\ell=C_v+1}^{N} (\ell - C_v) p_\ell}{\sum_{\ell=1}^{N} \ell p_\ell}$$

to obtain the average loss.

Although this formula only depends on the means of the silence and talkspurt distributions, the actual manner in which loss occurs, e.g. 1% of each talkspurt vs. 1 of every 100 talkspurts in its entirety, depends on the complete distri-

butions. To see this, consider the following examples. Suppose $N = 2$, $C_v = 1$, and the activity of each speaker is modelled as an alternating renewal process with silence $\sim \exp(\lambda)$ and a talkspurt that is a mixture of two deterministic times τ_1 and τ_2. There are two cases:

- $\tau_1 >> \frac{1}{\lambda} > \tau_2$, where the mixing probability, α is chosen so that $\alpha\tau_1 + (1-\alpha)\tau_2 = \frac{1}{\lambda}$, i.e. the long talkspurt occurs rarely. Then it is evident that one speaker can occassionally lose several entire talkspurts while the other ties up the circuit with a long talkspurt.

- $\tau_1 = \tau_2 = 1/\lambda$. In this case a speaker will never lose an entire talkspurt since the event of simultaneous completion of silences has zero probability. However, the mean silence and talkspurt times are the same in both cases so the average losses are the same.

From the previous expression for ϕ we can obtain two simple bounds

$$\phi \leq \frac{N\, P_r(A \geq C_v + 1)}{\bar{A}} = \frac{\mu + \lambda}{\lambda} P_r(A \geq C_v + 1)$$

$$\phi \geq \min\{0, \frac{(\bar{A} - C_v)}{\bar{A}} P_r(A \geq \bar{A} + 1)\}$$

From the properties of the binominal distribution, we know that if N and C_v approach infinity with $C_v = \bar{A}(1+\epsilon)$, $\epsilon > 0$,

then $P_r(A \geq C_v + 1)$ approaches 0. (Essentially, this is because the standard deviation is $0(\sqrt{N})$, which approaches 0 as a percentage of the mean, $\bar{A} = N \frac{\lambda}{\lambda+\mu}$, as $N \to \infty$.) Since this works for any $\varepsilon > 0$, it follows that we can approach the TASI advantage $\frac{N}{\bar{A}}$ arbitrarily closely (in percentage terms) with no loss, as $N \to \infty$. Note $\frac{N}{\bar{A}} = \frac{\mu + \lambda}{\mu}$, which is the inverse of the equilibrium activity fraction for one speaker. It also follows from the properties of the binominal distribution that as $N \to \infty$, $P_r(A \geq \bar{A} + 1)$ approaches a positive limit. If we choose $C_v = \bar{A}(1 - \varepsilon)$, $1 > \varepsilon > 0$, then the lower bound is $\frac{\bar{A}\varepsilon}{\bar{A}} P_r(A \geq \bar{A} + 1) = \varepsilon P_r(A \geq \bar{A} + 1)$. By the previous remark, this remains bounded away from 0 as $N \to \infty$. That is, if we attempt TASI with $C_v = \bar{A}(1 - \varepsilon)$, then for any $1 > \varepsilon > 0$, the loss is bounded away from 0 in the limit. For these reasons, the ratio $\frac{N}{\bar{A}} = 1 + \frac{\lambda}{\mu}$ is called the maximum TASI advantage.

The following table is taken from Weinstein [1]. In this table, $C_v = 36$ and $\frac{\lambda}{\lambda+\mu} = .4$

N	$\bar{A}$	TASI Advantage	Utilization = $\bar{A}/C_v$	Loss
60	24	1.66	.66	3.8×10^{-5}
70	28	1.95	.77	1.5×10^{-3}
75*	30	2.08	.83	.005*
80	32	2.22	.88	.014
85	34	2.36	.94	.029
90	36	2.5	1	.05
100	40	2.77	1.11	.11

The recommended operating point is N = 75.

Although the TASI advantage $1 + \frac{\lambda}{\mu}$ can be approached in the limit with no loss, for any finite N and $N > C_v > \bar{A}$, the loss is still nonzero. That is, although the mean activity level is $\bar{A}$, the activity process does exhibit fluctuations above $\bar{A}$, and this results in loss. On the other hand, if infinite buffering of speech is allowed, then for finite N and C_v, we can achieve a stable queue and no loss with any $C_v > \bar{A}$. This follows from the queuing theory " metaprinciple" that stability is present as long as the service rate exceeds the average arrival rate of work. If the buffer is finite, overflow speech is lost, but the loss fraction decreases as the buffer size increases. (Of course, the average delay also increases.) Thus there is a loss vs. delay vs. TASI advantage tradeoff. A formal model of this single-link, buffered TASI multiplexer has been developed and analyzed by Berger [6].

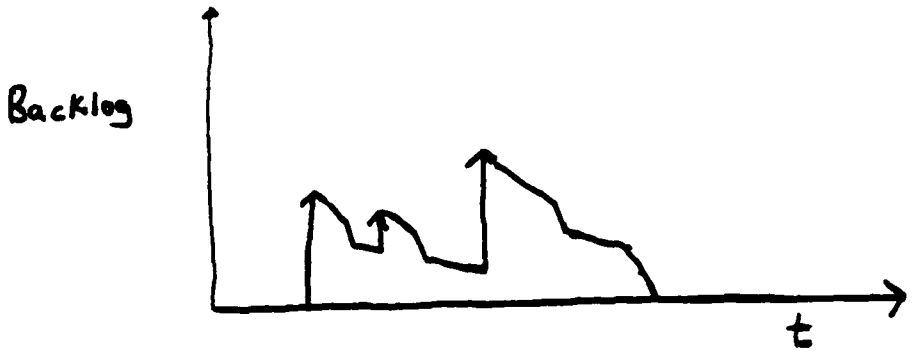
E. A Voice/Data Link Model

A link having capacity C bps is shared by N callers, whose activity is modelled by the birth-death chain, and an infinite data buffer. Data arrives in messages. The message arrival point process is taken as a renewal with mean rate η messages/sec. Message lengths are modelled as i.i.d. random variables with mean length ξ^{-1} bits.

Remarks

- Although the term "bit" is used for a "unit" of message length, we treat these lengths as continuous variables.
- A few of our results will apply to general arrival processes, but, in the main, we will assume Poisson arrivals and exponential length distributions. For a queue fed by many small, independent sources, the assumption of Poisson arrivals is reasonable because of the limit theorem mentioned in I.C.

For now we assume that the allocation of capacity is given and depends only on A(t). That is, when A(t) = i, the data backlog is decreasing at some rate r_i bps, and we are not concerned with how the remaining $(C-r_i)$bps is used to satisfy the spearkers. In a later chapter we will discuss the control problem -- how should the capacity be divided given some cost functions associated with an allocation policy. Now, our main concern is to analyze a given allocation. With this assumption, a sample function of the data backlog has the following general shape.



The jumps indicate message arrivals, and changes in slope reflect changes in speaker activity.

In terms of the TDM architecture, the model is a limiting case in which the frame time and block size are very small compared to speaker talkspurt and silence times, i.e. we ignore the discrete nature of the TDM architecture. As indicated, this is not an unrealistic assumption.

To describe the message arrival process, we adopt the standard "A/B" notation from Queuing Theory. That is, A is the message interarrival time distribution, and B is the message length distribution. M stands for the Markovian or exponential distribution, G is general etc. It should be noted that, although message lengths are i.i.d, service times, i.e. actual times to transmit messages, are not. The dependencies enter through the speaker activity process, which will also be called the phase process.

II. Notation and Background

In this chapter, a brief discussion of finite, time-homogeneous Markov chain is presented. The material is to a large extent an adaptation of Keilson [7]. The purpose is to establish some notation and a body of "quotable results".

Vectors are denoted by single underscoring, and matrices by double underscoring. The symbol $\triangleq$ means "is defined as". If $\underline{x}$ is a vector, the associated diagonal matrix $\underline{\underline{x}}_D \triangleq (\underline{\underline{x}}_D)_{ij} = x_i \delta_{ij}$, $\delta_{ij} \triangleq$ Kronecker delta. The vector $\underline{1}$ is a vector of all ones, and if η is a scalar, $\underline{\eta} \triangleq \eta \underline{1}$. The identity matrix, which would be $\underline{\underline{1}}_D$ in the above notation, is denoted by $\underline{\underline{I}}$. If x and y are vectors or matrices of the same dimension, the statements $x = y$, $x \geq y$, $x > y$ mean that specified relation holds on a componentwise basis. If $\underline{\underline{A}} \geq \underline{\underline{0}}$ is a matrix then $\underline{\underline{A}}$ is called <u>substochastic</u> if $\underline{\underline{A}}\,\underline{1} \leq \underline{1}$, <u>stochastic</u> if $\underline{\underline{A}}\,\underline{1} = \underline{1}$, and <u>strictly substochastic</u> if $\underline{\underline{A}}\,\underline{1} < \underline{1}$.

<u>The Perron-Romanovsky-Froebinius Theorem</u> (PRF)
(See Gantmacher [8]).

Let $\underline{\underline{A}}$ be an nxn matrix s.t. $\underline{\underline{A}} \geq \underline{\underline{0}}$. $\underline{\underline{A}}$ is called <u>reducible</u> iff it can be put into the form $\begin{pmatrix} \underline{\underline{B}} & \underline{\underline{0}} \\ \underline{\underline{C}} & \underline{\underline{D}} \end{pmatrix}$, for some square matrix $\underline{\underline{B}}$, via a sequence of permutations (interchanges of rows and the corresponding columns). Otherwise $\underline{\underline{A}}$ is <u>irreducible.</u>

$\underline{A}$ is called <u>primitive</u> if for some integer $m \geq 1$, $\underline{A}^m > \underline{0}$. Associate an n-node directed graph with $\underline{A}$ by placing an arc from $i \to j$ iff $A_{ij} > 0$. Then one can show that $\underline{A}$ is irreducible iff the graph is strongly connected, and $\underline{A}$ is primitive iff the g.c.d. of all cycle lengths is 1, and $\underline{A}$ is irreducible.

<u>Theorem PRF</u>

Let $\underline{A} \geq \underline{0}$ be irreducible. Then $\underline{A}$ has a real positive eigenvalue r with the following properties.

1) If λ is any other eigenvalue, then $|\lambda| \leq r$.

2) r is of algebraic and geometric multiplicity one. That is r is a simple root of the characteristic polynomial and has a one dimensional eigenspace. The associated right eigenvector $\underline{x}_R$ can be chosen real and positive, i.e. $\underline{x}_R > \underline{0}$. r is the only eigenvalue having such an eigenvector.

3) If m and M are the minimum and maximum row sums then $m \leq r \leq M$. Strict inequality holds in both cases unless $M = m$.

4) If there are h eigenvalues of modulus r (counting r), then the spectrum of $\underline{A}$ is mapped into and onto itself by a rotation of the complex plane of angle $\frac{2\pi}{h}$.

5) $h = 1$ iff $\underline{A}$ is primitive. r is called the spectral radius of $\underline{A}$, $sp(\underline{A})$; the PRF root; or PRF eigenvalue.

<u>Remark</u>: Since $\underline{A}^T$ is irreducible iff $\underline{A}$ is (reverse directions of arcs), we can apply PRF to $\underline{A}^T$ as well. This yields possibly

different bounds on r in terms of column sums and a positive left eigenvector $\underline{x}_L$ for $\underline{A}$.

In the primitive case, the behavior of $\underline{A}^m$ as $m \to \infty$ is determined by r, $\underline{x}_R$ and $\underline{x}_L$. Let $\underline{J} \triangleq \underline{x}_R \underline{x}_L^T$, where $\underline{x}_R$ and $\underline{x}_L$ are normalized so that $\underline{x}_L^T \underline{1} = 1$, $\underline{x}_L^T \underline{x}_R = 1$. Let $\underline{\Omega} = \underline{A} - r \underline{J}$. Then one can show

1) $\underline{x}_L^T \underline{\Omega} = \underline{0}^T$, $\underline{\Omega} \underline{x}_R = \underline{0}$, $\underline{J} \underline{\Omega} = \underline{\Omega} \underline{J} = \underline{0}$.

2) $\underline{J}^2 = \underline{J}$, $\underline{J} \underline{A} = \underline{A} \underline{J} = r \underline{J}$.

3) $\underline{J}$ is a dyad having 1 as an eigenvalue of geometric and algebraic multiplicity one; and 0 as an eigenvalue of algebraic and geometric multiplicity n-1.

4) The eigenvalues of $\underline{\Omega}$ are

   i) 0 with associated eigenvectors $\underline{x}_R$ and $\underline{x}_L$

   ii) the eigenvalues of $\underline{A}$ other than r. If $\underline{A} \underline{z} = \lambda \underline{z}$, $\lambda \neq r$, then $\underline{\Omega}(\underline{\Omega} \underline{z}) = \lambda \underline{\Omega} \underline{z}$.

5) $\underline{A}^m = r^m \underline{J} + \underline{\Omega}^m$, and since r is uniquely maximal, $(\underline{\Omega}/r)^m \to \underline{0}$ as $m \to \infty$. This implies $(\underline{A}/r)^m \to \underline{J}$.

The last statement suggest the following algorithm for finding $\underline{x}_L$ and r.

1. $\underline{x}_o^T \leftarrow \underline{1}$
2. $\underline{x}_{i+1}^T \leftarrow (\underline{x}_i^T \underline{A})$
3. $\underline{x}_{i+1}^T \leftarrow \underline{x}_{i+1}^T / \underline{x}_{i+1}^T \cdot \underline{1}$

The algorithm is essentially a computation of $\frac{\underline{x}_o^T \underline{A}^m}{\underline{x}_o^T \underline{A}^m \underline{1}}$ = $\frac{\underline{x}_o^T[r^m \underline{J} + \underline{\Omega}^m]}{\underline{x}_o^T[r^m \underline{J} + \underline{\Omega}^m]\cdot\underline{1}}$. By definition of $\underline{J}$ this is

$$\frac{(\underline{x}_o^T \underline{x}_R)\ \underline{x}_L^T + \underline{x}_o^T(\underline{\Omega}/r)^m}{\underline{x}_o^T\left[\underline{x}_R\underline{x}_L^T + (\underline{\Omega}/r)^m\right]\cdot\underline{1}}$$

Since $(\underline{\Omega}/r)^m \to \underline{0}$ as $m \to \infty$, this converges to $\underline{x}_L^T$. Since $\underline{x}_L^T \underline{A} = r\ \underline{x}_L^T$, we can then find r by $r = \underline{x}_L^T \underline{A}\cdot\underline{1}$. Note that there is no particular reason to increment by only one power of $\underline{A}$ in step 2. We could precompute $\underline{A}^m$ for some large m and then use it. The limitation is that if $r > 1$ or $r < 1$, then $\underline{A}^m$ (without scaling) may overflow or underflow the machine. One simple way around this is to simultaneously compute large powers of A and rescale. We can do this by successive squarings. (Also note that $\underline{x}_o$ is arbitrary as long as it's positive.)

1. $\underline{x}_o^T \leftarrow \underline{1}$; $\underline{A}_o \leftarrow \underline{A}/(\underline{x}_o^T \underline{A}\cdot\underline{1})$

2. $\underline{x}_{i+1}^T \leftarrow (\underline{x}_i^T \underline{A}_i)$

3. $\underline{A}_{i+1} \leftarrow \underline{A}_i^2$

4. $\underline{A}_{i+1} \leftarrow \underline{A}_{i+1}/(\underline{x}_o^T \underline{A}_{i+1}\cdot\underline{1})$

One may easily check that $\underline{A}_m = \underline{A}^{2^m}/(\underline{x}_o^T \underline{A}^{2^m} \underline{1})$. The convergence

rate is determined by the rate at which $(\underline{\Omega}/r)^k$ approaches 0, this being geometric with rate r_1/r where r_1 is the second largest eigenvalue modulus.

Discrete Time Chains

Suppose a Markov chain has M states {1, ..., M}, and S(k) denotes the state after the k^{th} transition. The evolution of S is characterized by a stochastic matrix $\underline{A}$ called the transition probability matrix. If S(0) is drawn according to a distribution $\underline{\pi}(0)$, then the distribution of S(k) is $\underline{\pi}(k)$, where $\underline{\pi}^T(k) = \underline{\pi}^T(0)\ \underline{A}^k$. The existence of a limit for $\underline{\pi}(k)$ as $k \to \infty$ is a fundamental question. There are three cases.

1) If $\underline{A}$ is primitive then a limiting $\underline{p}$ exists, independent of $\underline{\pi}(0)$. Further $\underline{p}$ is the unique left eigenvector -- $\underline{p}^T\ \underline{A} = \underline{p}^T$, $\underline{p} > \underline{0}$, $\underline{p}^T\ \underline{1} = 1$. This follows from the fact that $\underline{A}^k$ approaches $\underline{J} = \underline{1}\ \underline{p}^T$ as $k \to \infty$. (Because of the primitivity, all eigenvalues other than 1 are strictly inside the unit circle.) $\underline{p}$ is called the ergodic, equilibrium, or steady state distribution, and the chain is called ergodic.

2) $\underline{A}$ irreducible but imprimitive -- The irreducibility guarantees the existence of a unique left eigenvector $\underline{p}$ s.t. $\underline{p} > \underline{0}$ and $\underline{p}^T\ \underline{1} = 1$. But $\underline{\pi}(k)$ exhibits oscillatory behavior, and $\underline{p}$ is not a true limit. Recall that $\underline{A}$ imprimitive implies that there exist i,j s.t. j is reachable

from i at only at periodically spaced transition epochs. (From earlier remarks, it is clear that one can take $j = i$, i.e. there is a node in the graph s. t. every cycle beginning and ending at i has length that is a multiple of some integer ≥ 2.) To see the oscillatory behavior, consider the simple case of $\underline{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which corresponds to the chain  . In this case $\underline{p} = (\frac{1}{2}, \frac{1}{2})$, and S does spend half its time in each state, asymptotically. However $\underline{A}^k = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + (-1)^k \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$. So if $\underline{\pi}(0) = (1,0)$, $\underline{\pi}(k) = (\frac{1}{2} + (-1)^k \frac{1}{2}, \frac{1}{2} - (-1)^k \frac{1}{2})$ which oscillates between (1,0) and (0,1).

Remark: If $\underline{p}$ is a left eigenvector with eigenvalue 1, then $\underline{\pi}(0) = \underline{p}$ implies $\underline{\pi}(k) = \underline{p}$, $\forall$ k. If $\underline{p} \geq \underline{0}$ and $\underline{p}^T \underline{1} = 1$, then $\underline{p}$ is called a stationary distribution for obvious reasons. Irreducibility guarantees the existence and uniqueness of a stationary $\underline{p}$, and primitivity guarantees that $\underline{\pi}(k)$ approaches this $\underline{p}$ in the limit. We will only deal with irreducible, non-negative matrices, so the term "stationary distribution" will always mean "unique stationary distribution".

3. $\underline{A}$ reducible - $\underline{A}$ is stochastic so 1 is trivially an eigenvalue. However, it can have geometric and algebraic multiplicities larger than 1, though there is at least one

left eigenvector which can be chosen nonnegative. Gantmacher [8] discusses the possibilities in detail.

Continuous Time Chains

The transition probabilities of discrete time are replaced by transition rates ν_{ij}, where $\nu_{ij} \geq 0$, $\nu_{ii} = 0$ V i. When S = i, the next state and time of transition are determined by a set of competing Poisson processes with rates ν_{ij}. If an event from process j occurs first, the state goes to j at the time of occurrence (Since $\nu_{ii} = 0$, there are effectively at most M-1 processes. If $\nu_{ij} = 0$ V j, the state i is called absorbing. We only consider those chains for which $\sum_j \nu_{ij} > 0$, V i.) A straightforward calculation shows that no matter which j "wins", the a posteriori distribution on the holding time in i is exp (ν_i) where $\nu_i = \sum_j \nu_{ij}$. That is, the minimum of the exponential random variables with parameters ν_{ij} is distributed as exp(ν_i) regardless of which variable realizes the mimimum. Further, the probability that j "wins" is $\frac{\nu_{ij}}{\nu_i}$. This leads to the following equivalent view of the chain. When S = i, the time of the next transition is drawn from exp(ν_i), and at this time, the next state is drawn according tc the probabilities ν_{ij}/ν_i. That is, there is a discrete time chain with transition probability matrix $\underline{\nu}_D^{-1}\ \underline{\nu}$ (where $\underline{\nu}_D \triangleq$ diagonal (ν_i) and $(\underline{\nu})_{ij} = \nu_{ij}$), and a state dependent clock determining the actual time of

transitions.

The role of the discrete time $\underline{A}$ matrix is played by the infinitesimal generator matrix $\underline{Q} \triangleq \underline{\nu} - \underline{\nu}_D$. The transition probability matrix $\underline{P}(t)$, where $P_{ij}(t) = \Pr[S(t) = j / S(0) = i]$, is the unique solution of the so called forward and backward equations $\dot{\underline{P}}(t) = \underline{Q}\,\underline{P}(t) = \underline{P}(t)\,\underline{Q}$ s.t. $\underline{P}(0) = \underline{I}$. The solution $\underline{P}(t)$ is the well known matrix exponential $\exp(\underline{Q}\,t) = \sum_{n=0}^{\infty} \frac{(\underline{Q}\,t)^n}{n!}$. If $S(0)$ is drawn from a distribution $\underline{\pi}(0)$, then $\underline{\pi}^T(t) = \underline{\pi}^T(0)\underline{P}(t)$.

The existence of limits is determined by the irreducibility of the nonnegative matrix $\underline{\nu}$, or equivalently, by the connectivity of the graph in which an arc $i \to j$ is present iff $\nu_{ij} > 0$. If irreducibility is present, then $\underline{\pi}(t)$ approaches a limit independent of $\underline{\pi}(0)$, and the chain is termed ergodic. We assume irreducibility for the discussion in this chapter. The limit $\underline{p}$ satisfies $\underline{p}^T\,\underline{Q} = \underline{0}^T$, $\underline{p} > \underline{0}$ $\underline{p}^T\,\underline{1} = 1$ and is called the stationary or ergodic distribution. (Note $\underline{Q}\,\underline{1} = \underline{0}$ by construction, so 0 is an eigenvalue.) Since a limit exists, all nonzero eigenvalues of $\underline{Q}$ must have nonpositive real parts, else $\exp(\underline{Q}\,t)$ explodes, and such a solution is not probabilistically meaningful. However, we have not introduced a notion analogous to primitivity to prevent oscillation, i.e. to rule out purely imaginary eigenvalues. Such a notion is not necessary because in continuous time, any state j reachable from i, is reachable within any positive time, i.e. $P_{ij}(t) > 0$ $\forall$ t. We

derive the above properties of the continuous time chain by introducing the following uniformization procedure.

Let $\hat{\nu}$ be chosen $\geq \max\{\nu_i\}$, and set $\underline{A}_{\hat{\nu}} = \underline{I} + \underline{Q}/\hat{\nu}$. Then $\underline{A}_{\hat{\nu}} \geq \underline{0}$, $\underline{A}_{\hat{\nu}}$ is stochastic and

$$\exp(\underline{Q}t) = \exp(-\hat{\nu}t(\underline{I} - \underline{A}_{\hat{\nu}})) = \sum_{n=0}^{\infty} e^{-\hat{\nu}t} \frac{(\hat{\nu}t)^n}{n!} (\underline{A}_{\hat{\nu}})^n$$

This has a simple probabilistic interpretation. A chain is driven by a single, Poisson clock of rate $\hat{\nu}$. At each "bong" a transition occurs according to the matrix $\underline{A}_{\hat{\nu}}$. Such a transition occurs from i to j, $i \neq j$, with probability $\frac{\nu_{ij}}{\hat{\nu}}$ and from i to i with probability $1 - \frac{\nu_i}{\hat{\nu}}$. Notice that if $\hat{\nu} > \nu_i$ this self-loop has positive probability. This reflects the fact that state i is being driven faster than its "natural" rate, ν_i.

The matrix $\underline{A}_{\hat{\nu}}$ is irreducible since $\underline{\nu}$ is, and λ is an eigenvalue of $\underline{A}_{\hat{\nu}}$ iff $\hat{\nu}(\lambda-1)$ is an eigenvalue of $\underline{Q}$. The corresponding eigenvectors are the same. Applying the PRF theorem to $\underline{A}_{\hat{\nu}}$ and using these relations between $\underline{Q}$ and $\underline{A}_{\hat{\nu}}$, we can conclude that

- 0 is an eigenvalue of $\underline{Q}$ with algebraic and geometric multiplicity of 1,
- all other eigenvalues of $\underline{Q}$ have negative real parts,
- $\underline{Q}$ and $\underline{A}_{\hat{\nu}}$ have the same stationary distribution $\underline{p}$.

We now show that $\underline{\pi}^T(0)\ \exp(\underline{Q}\ t)$ approaches $\underline{p}^T$ as $t \to \infty$, for any $\underline{\pi}^T(0)$.

It suffices to show that $\exp(\underline{Q}\ t)$ approaches $\underline{J} = \underline{1}\ \underline{p}^T$. Let $\underline{\Omega} = \underline{A}_{\hat{\nu}} - \underline{J}$. Then $(\underline{A}_{\hat{\nu}})^n = \underline{J} + \underline{\Omega}^n$. This implies

$$\exp\ (\underline{Q}t) = \sum_{n=0}^{\infty} e^{-\hat{\nu}t}\ \frac{(\nu t)^n}{n!}\ (\underline{J} + \underline{\Omega}^n) = \underline{J} + \exp(-\hat{\nu}t(\underline{I} - \underline{\Omega})).$$

By construction, all eigenvalues of $\underline{\Omega}$ are on or inside the unit circle, but 1 itself is not an eigenvalue of $\underline{\Omega}$. Therefore, the eigenvalues of $\underline{I} - \underline{\Omega}$ have positive real parts, which implies that $\exp(-\hat{\nu}t(\underline{I} - \underline{\Omega})) \to 0$ as $t \to \infty$.

One useful consequence of uniformization is an algorithm for finding the stationary distribution of $\underline{Q}$. The previous algorithm does not work directly on $\underline{Q}$ because $\underline{Q}$ contains transition rates, and its powers have no probabilistic meaning. However, the stationary distribution of $\underline{Q}$ is the same as that of $\underline{A}_{\hat{\nu}}$ for any uniformizing rate $\hat{\nu}$. The previous algorithm does work on $\underline{A}_{\hat{\nu}}$ provided that it is primitive. The following argument shows that any choice of $\hat{\nu} > \max\{\nu_i\}$ yields a primitive $\underline{A}_{\hat{\nu}}$. For such a choice of $\hat{\nu}$, $(A_{\hat{\nu}})_{ii} = 1 - \nu_i/\hat{\nu} > 0$, $\forall$ i. This implies that every node in the graph associated with $\underline{A}_{\hat{\nu}}$ has a self-loop. In turn, this implies that the g.c.d. of all cycle lengths is 1. Primitivity then follows from irreducibility.

If a continuous time chain is ergodic, the ergodic probability p_i admits an interpretation as the limiting fraction of time the chain spends in state i. For a discrete time chain,

the ergodic probability p_i is the limiting fraction of transition epochs at which the chain leaves state i. (One could also count the fraction of epochs at which state i is entered. In equilibrium, these are the same, i.e. looking immediately before or after a transition makes no difference.) If $\underline{A}_\nu^\wedge$ is a uniformizing matrix for a $\underline{Q}$ matrix, then $\underline{A}_\nu^\wedge$ is the transition probability matrix for a discrete time chain embedded at "clock bongs". Intuitively, $\underline{A}_\nu^\wedge$ and $\underline{Q}$ have the same ergodic probabilities because (uniform rate) Poisson sampling takes a truly "random" look at the continuous time chain. Now, if the continuous time chain has a transition rate matrix $\underline{\nu}$ (so that $\underline{Q} = \underline{\nu} - \underline{\nu}_D$), the stochastic matrix $\underline{\nu}_D^{-1}\ \underline{\nu}$ is a transition probability matrix for a discrete time chain embedded at transition epochs of the continuous time chain as determined by the state dependent clock. Since $\underline{\nu}$ is irreducible, $\underline{\nu}_D^{-1}\ \underline{\nu}$ has a stationary distribution $\tilde{\underline{p}}$. However, $\tilde{\underline{p}}$ and $\underline{p}$, the ergodic distribution for $\underline{Q}$, need not be the same because the holding times in various states need not be the same. That is, even though the holding time in each state is exponentially distributed, the means are different, so that it is not uniform Poisson sampling. However, there is a simple relationship between $\underline{p}$ and $\tilde{\underline{p}}$ which is obtained by rescaling to correct for clock rate differences. Specifically, $\underline{p}^T\ \underline{Q} = \underline{0}^T$ implies $\underline{p}^T(\underline{\nu} - \underline{\nu}_D) = \underline{0}^T$ which implies $\underline{p}^T\ \underline{\nu}_D(\underline{\nu}_D^{-1}\ \underline{\nu}) = \underline{p}^T\ \underline{\nu}_D$. Since the stationary distribution $\tilde{\underline{p}}$ is unique, it follows that $\underline{p}^T\ \underline{\nu}_D$ is a scalar multiple of $\tilde{\underline{p}}^T$, i.e.

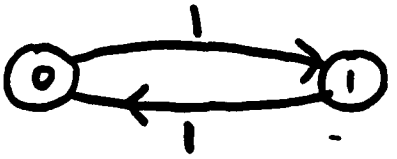
$$\tilde{p}^T = (p^T \underline{\nu}_D)/(p^T \underline{\nu}_D \cdot 1) \text{ or } \tilde{p}_i = \frac{p_i \nu_i}{\Sigma p_i \nu_i} .$$

These issues can be seen in the following example. Consider the continuous time chain  .

$$\underline{\nu} = \begin{pmatrix} 0 & \lambda \\ \mu & 0 \end{pmatrix}, \quad \underline{\nu}_D = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \text{ so}$$

$\underline{Q} = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$. The ergodic probabilities are $p_0 = \frac{\mu}{\lambda+\mu}$, $p_1 = \frac{\lambda}{\lambda+\mu}$.

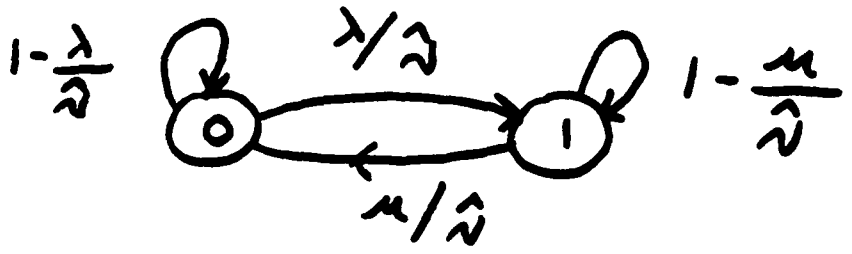
Thus $\underline{\nu}_D^{-1} \underline{\nu} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ which corresponds to the discrete time chain.



This chain is irreducible with $\tilde{p}_0 = \frac{1}{2}$, $\tilde{p}_1 = \frac{1}{2}$. However, it is not ergodic since it is not primitive. Here, $\tilde{p}_i \neq p_i$ if $\lambda \neq \mu$ because this discrete time chain ignores the difference in clock rates. If $\lambda = \mu$, the continuous time chain is already uniformized so that $\tilde{p}_i = p_i = \frac{1}{2}$. But the

matrix $\underline{\nu}_D^{-1}\ \underline{\nu}$ is still imprimitive and its eigenvalues are 1 and -1.

If we pick a uniformizing rate $\hat{\nu} > \lambda,\mu$, the associated discrete time chain is



This chain is irreducible and primitive, i.e. ergodic.

In discrete time, the left eigenvector condition for $\underline{p}$, $\underline{p}^T = \underline{p}^T\ \underline{A}$, indicates a global balance. (Writing out the equation gives $p_i = \sum_j p_j\ A_{ji}$.) In continuous time, the analogous condition is with "probability flow". The condition $\underline{p}^T\ \underline{Q} = \underline{0}^T$ translates into $p_i \sum_j \nu_{ij} = \sum_j p_j\ \nu_{ji}$. Some chains exhibit a stronger form of balance, called <u>detailed balance</u>, in which there is equilibrium between every pair of states -- $p_i\nu_{ij} = p_j\nu_{ji}$. In matrix terms, this is $\underline{p}_D\ \underline{Q} = \underline{Q}^T\ \underline{p}_D$, where $\underline{p}_D$ is the diagonal matrix obtained from $\underline{p}$. This implies $\hat{\underline{Q}} \triangleq \underline{p}_D^{\frac{1}{2}}\ \underline{Q}\ \underline{p}_D^{-\frac{1}{2}}$ is symmetric, which in turn implies that $\hat{\underline{Q}}$ and hence $\underline{Q}$ have real eigenvalues. Further, $\hat{\underline{Q}}$ is diagonalizable via an orthogonal matrix $\underline{x}$ which implies that $\underline{x}\ \underline{p}_D^{\frac{1}{2}}$ diagonalizes $\underline{Q}$.

For a general chain, balance exists between any two sub-

sets N_1 and N_2 which partition the state space, i.e.

$$\sum_{i \in N_1} p_i \sum_{j \in N_2} \nu_{ij} = \sum_{i \in N_2} p_i \sum_{j \in N_1} \nu_{ij} .$$

For birth-death processes, we can choose $N_1 = \{i | i \leq i_o\}$. $N_2 = \{i | i > i_o\}$, and this balance equation is the detailed balance condition for i_o and $i_o + 1$. Since i_o is arbitrary, this shows that birth-death processes always exhibit detailed balance.

Miscellaneous Matrix Theory

1) An irreducible nonnegative matrix is similar to a scaled stochastic matrix. To see this, let $\underline{A} \geq \underline{0}$ be irreducible. Let r be the PRF root and $\underline{D}$ be the diagonal matrix obtained from a positive right eigenvector $\underline{x}_R$. (Note $\underline{x}_R > \underline{0} \Longrightarrow \underline{D}$ invertible.) Then

$$\frac{\underline{D}^{-1} \underline{A} \underline{D}}{r} \underline{1} = \frac{\underline{D}^{-1} \underline{A} \underline{x}_R}{r} = \underline{D}^{-1} \frac{r \underline{x}_R}{r} = \underline{D}^{-1} \underline{x}_R = \underline{1}.$$

2) Girshgorin's Theorem (See [9].)

Let $\underline{M}$ be an n x n matrix (possibly complex). Let $r_i \overset{\Delta}{=} \sum_{\substack{j=1 \\ j \neq i}}^{n} |M_{ij}|$. Then the eigenvalues of M are contained

in the union of the circles centered at $\{M_{ii}\}$ with respective radii $\{r_i\}$. Applying this to a $\underline{Q}$ matrix

(for which $M_{ii} = -\nu_i$, $r_i = \nu_i$) gives another quick proof that nonzero eigenvalues of $\underline{Q}$ have negative real parts. Further, if $\underline{D}$ is a diagonal matrix with positive diagonal terms, then the theorem shows that the eigenvalues of $\underline{D} - \underline{Q}$ have positive real parts.

III. Statistics of the Speaker Process

A. Transition Probabilities

Consider the two state chains of a single speaker. The $\underline{Q}$ matrix for this process is

$$\begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

The equilibrium distribution is binominal $p_0 = \frac{\mu}{\lambda+\mu}$, $p_1 = \frac{\lambda}{\lambda+\mu}$. The nonzero eigenvalue of $\underline{Q}$ is $-(\lambda+\mu)$. Let $f_{ij}(t) = \Pr[j \text{ at } t | i \text{ at } 0]$ $i,j = 0,1$. From that fact that $f_{ij}(t)$ is of the form $a + b\, e^{-(\lambda+\mu)t}$, $f_{ij}(\infty) = p_j$, $f_{ij}(0) = \delta_{ij}$, it follows that

$$f_{ii}(t) = p_i + p_j\, e^{-(\lambda+\mu)t}$$

$$f_{ij}(t) = p_j - p_j\, e^{-(\lambda+\mu)t} \qquad i \neq j.$$

For the N speaker process, the $\underline{Q}$ matrix is tridiagonal with

$$Q_{ii} = -(N-i)\lambda - i\mu$$

$$Q_{i,i+1} = (N-i)\lambda \qquad 0 \leq i \leq N$$

$$Q_{i,i-1} = i\mu$$

(Note that the indexing runs from 0 to N to preserve the connection with number of active speakers.) Given the single speaker transition probabilities, we can derive the N speaker

probabilities using the independence. The expression is

$$p_{ij}(t) = \sum_k \binom{i}{k} \underbrace{[f_{11}(t)]^k [f_{10}(t)]^{i-k}}_{\text{term 1}} \underbrace{\binom{N-i}{j-k} [f_{01}(t)]^{j-k} [f_{00}(t)]^{N-i-j+k}}_{\text{term 2}}$$

where it is understood that $\binom{A}{B} = 0$ if $B > A$ or $B < 0$. This expression comes from the fact that given $A(0) = i$, it is possible to have $A(t) = j$ if some of the k active at 0 are active at t and j-k of the inactive at 0 are active at t, these events occuring with probabilities given by term 1 and term 2 respectively.

By examining the expressions for $f_{ij}(t)$ one can see that $P_{ij}(t)$ is a linear combination of the exponentials $e^{s_k t}$, where $s_k = -k(\lambda+\mu)$, $0 \leq k \leq N$. The $\{s_k\}$ are of course the eigenvalues of $\underline{Q}$. The coefficient of the s_o term is just $p_j = \binom{N}{j} \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\mu+\lambda}\right)^{N-j}$, the equilibrium probability that $A = j$.

From the discussion on detailed balance, we know that $\underline{Q}$ is similar to a symmetric matrix which implies $\underline{Q}$ is diagonalizable. (In a more elementary way, this particular $\underline{Q}$ is diagonalizable, i.e. has a basis of eivenvectors, because it has distinct eigenvalues, and eigenvectors corresponding to distinct eigenvalues are linearly independent.) Let $\underline{S}$ = diagonal $\{s_k\}$. If $\underline{L}$ diagonalizes $\underline{Q}$, i.e. $\underline{Q} = \underline{L}^{-1} \underline{S} \underline{L}$

(the rows of $\underline{L}$ are left eigenvectors of $\underline{Q}$ and the columns of $\underline{L}^{-1}$ are right eigenvectors of $\underline{Q}$) then $\underline{P}(t) = \exp(\underline{Q}\, t) = \underline{L}^{-1} \exp(\underline{S}\, t)\underline{L}$. Since $\underline{S}$ is diagonal, $[\exp(\underline{S}\, t)]_{ij} = e^{s_i t} \delta_{ij} \; e^{-i(\lambda+\mu)t} \delta_{ij}$. So if we can find $\underline{L}$, we can find another representation of $\underline{P}(t)$.

We now develop explicit expressions for the eigenvectors of $\underline{Q}$. The expressions are not original (see Karlin [10]), but the derivation was obtained independently. For this reason, the discussion is not too detailed.

We wish to find a matrix $\underline{L}$ s.t. $\underline{L}\,\underline{Q} = \underline{S}\,\underline{L}$, i.e. the j^{th} row of $\underline{L}$, $\underline{L}(j)$, satisfies $\underline{L}(j)\underline{Q} = s_j \underline{L}(j)$. To complete the diagonalization, we then need to find $\underline{L}^{-1}$. Let $\varepsilon = \frac{\lambda}{\mu}$ and redefine $\underline{Q}$ by factoring out μ from each term, i.e. $\underline{Q} \leftarrow \frac{1}{\mu}\underline{Q}$, so that now

$$Q_{ii} = -(N-i)\varepsilon - i$$

$$Q_{i,i-1} = i$$

$$Q_{i,i+1} = (N-i)\varepsilon$$

This factoring reduces the eigenvalues by μ, i.e. s_k now is $-k(1+\varepsilon)$, and does not affect the eigenvectors. (That is, in the expression for $P_{ij}(t)$, the coefficients of the exponentials only depend on the relative values of λ and μ.)

An eigenvector $\underline{L}(0)$ is already known to be the equilibrium probability vector, $p_i = \binom{N}{i} \frac{\varepsilon^i}{(1+\varepsilon)^N}$. For convenience, we take $\underline{L}_i(0) = \binom{N}{i}\varepsilon^i$. Notice that $\underline{L}(0)$ is an N-fold convolution of

the vector [1 ε] with itself, i.e. $\underline{L}(0) = [1 \quad \varepsilon]^{*N}$. The convolution operation is associative and commutative so $[1 \quad \varepsilon]*\ldots*[1 \quad \varepsilon] = [1 \quad \varepsilon]^{*N}$ is well defined.) One can check that [1 - 1] is an eigenvector for s_1 in the case N = 1. We now show that the vectors [1 - 1] and [1 ε] generate the $\{\underline{L}(j)\}$.

Claim: $\underline{L}(j) = [1 \quad \varepsilon]^{*(N-j)} * [1 \; -1]^{*j}$. Our method of proof is to relate $\underline{Q}$ and $\underline{L}$ for an N speaker process to those for an N-1 speaker process. We use a superscript on matrices and vectors to indicate the number of speakers.

Define $\underline{G}^N$ by

$$\underline{G}^N_{ij} = 1 \qquad j \geq i$$

$$\underline{G}^N_{ij} = 0 \qquad j < i$$

Once can check that $(\underline{G}^N)^{-1}$ has ones on the diagonal, minus ones on the superdiagonal, and zeroes elsewhere. As a change of basis, this transformation is replacing a state j by the sum of all states $\leq$ j. The crucial step is then to show that

$$\tilde{\underline{Q}}^N = (\underline{G}^N)^{-1} \, \underline{Q}^N \, \underline{G}^N = \left(\begin{array}{c|c} \underline{Q}^{N-1} - (1+\varepsilon)\underline{I} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \\ \hline 0 \; 0 \; \ldots \; N_{\mu} & 0 \end{array} \right)$$

This relationship can be verified by a straightforward computation, so we omit proof. We can now draw two conclusions.

- If $\phi_N(s) = \det(s\underline{I} - \underline{Q}^N)$ then $\phi_N(s) = s\phi_{N-1}(s+(1+\varepsilon))$. This follows from the fact that similar matrices have the same characteristic polynomials, and, an expansion of $\det(s\underline{I} - \tilde{\underline{Q}}^N)$ along the last column. The relation between ϕ_N and ϕ_{N-1}, in turn implies that $s_o = 0$ is an eigenvalue of $\underline{Q}^N$, and that if $s_j = -j(1+\varepsilon)$ is an eigenvalue of $\underline{Q}^{N-1}$, then $s_{j+1} = -(j+1)(1+\varepsilon)$ is an eigenvalue of $\underline{Q}^N$. Since the eigenvalues of $\underline{Q}^1$ are 0 and $-(1+\varepsilon)$, it follows that the eigenvalues of $\underline{Q}^N$ are s_j, $0 \leq j \leq N$, as indicated before.

- A direct computation shows that if

  $\underline{L}^{N-1}(j)\ \underline{Q}^{N-1} = s_j\ \underline{L}^{N-1}(j)$ then

  $[\underline{L}^{N-1}(j),\underline{0}]\tilde{\underline{Q}}^N = s_{j+1}[\underline{L}^{N-1}(j),0]$

  or $[\underline{L}^{N-1}(j),0]\ (\underline{G}^N)^{-1}\underline{Q}^N = s_{j+1}[\underline{L}^{N-1}(j),0](\underline{G}^N)^{-1}$

In words, if $\underline{L}^{N-1}(j)$ is an eigenvector of s_j for the (N-1) speaker process, then $[L^{N-1}(j),0](\underline{G}^N)^{-1}$ is an eigenvector of s_{j+1} for the N speaker case. Since $\underline{L}^N(0) = [1\ \varepsilon]^{*N}$ $\forall$ N, $\underline{L}^1(1) = [1\ \ -1]$, and multiplication by $(\underline{G}^N)^{-1}$ is equivalent to convolution with $[1\ \ -1]$, the claim follows by induction.

Remark: $[1 \quad \varepsilon]^{*(N-j)} * [1 \quad -1]^{*j}$ can be expressed in terms of the binominal coefficients and powers of ε. The resulting polynomial is a special case of the Krawtchouk polynomials (see Karlin[10]).

To complete the diagonalization we need to find $\underline{L}^{-1}$. We now show that $\underline{L}^2 = (1 + \varepsilon)^N \underline{I}$.(We drop the superscript for number of speakers so that $\underline{L}^2$ has its usual meaning.) Thus one can take $\underline{L}(1+\varepsilon)^{-N/2}$ for the diagonalization.

Claim: $\underline{L}^2 = (1 + \varepsilon)^N \underline{I}$

The proof is somewhat involved so we "sketch" it. First, one shows that $\underline{L}$ is also a right eigenvector matrix. This implies $\underline{L} = \underline{L}^{-1} \underline{D}$ for some diagonal matrix $\underline{D}$. Then one shows $\underline{D} = a\underline{I}$ for some constant a. These are the involved parts. Determining a is easy. $\underline{L} = a\underline{L}^{-1}$ implies $\underline{L}^2 = a\underline{I}$. The first row of $\underline{L}$ is $[1 \quad \varepsilon]^{*N}$, and the first column of $\underline{L}$ is $\underline{1}$. Now $([1 \quad \varepsilon]^{*N})\underline{1} = (1 + \varepsilon)^N$ by construction. Thus $a = (1+\varepsilon)^N$.

The matrix $\underline{L}$ has several other interesting properties which derive from the rich structure of the chain. They seem to have little probabilistic significance, so discussion is omitted.

B. Mean First Passage Times

We mentioned that the crucial question is one of time scales. One useful characterization of this is the mean first passage time between states. Let $T_{ij} \triangleq$ mean first passage time from i to j, $T_i^+ \triangleq T_{i,i+1}$; $T_i^- \triangleq T_{i,i-1}$. For the case N=1, $T_{01} = T_0^+ = \frac{1}{\lambda}$ and $T_{10} = T_1^- = \frac{1}{\mu}$. We can in fact derive recursive relations for these quantities for a general birth-death process. Let λ_n and μ_n denote the birth and death rates respectively. Then

$$T_n^+ = \frac{1}{\lambda_n+\mu_n} + \frac{\mu_n}{\lambda_n+\mu_n}[T_{n-1}^+ + T_n^*].$$

This equation says that we must

a) wait until the first exit from n, the mean of this time is $\frac{1}{\lambda_n+\mu_n}$ and

b) if this transition is to n-1, which occurs with probability $\frac{\mu_n}{\lambda_n+\mu_n}$, we must wait another $T_{n-1}^* + T_n^+$ to first reach n+1.

Solving for T_n^+ yields

$$T_n^+ = \frac{1}{\lambda_n} + \frac{\mu_n}{\lambda_n} T_{n-1}^+ .$$

The following induction argument shows that a solution to this recursion is

$$T_n^+ = \frac{1}{\lambda_n p_n} \sum_{j=0}^{n} p_j, \text{ where the } \{p_j\} \text{ are the ergodic probabilities.}$$

Proof: Detailed balance shows $\lambda_n p_n = \mu_{n+1} p_{n+1}$. The basis, $T_o^+ = \frac{1}{\lambda_o}$ is trivial. Assuming the formula true for n-1, we can substitute into the recursion to obtain

$$T_n^+ = \frac{1}{\lambda_n} + \frac{\mu_n}{\lambda_n} \left(\frac{1}{\lambda_{n-1} p_{n-1}} \sum_{j=0}^{n-1} p_j\right)$$

$$= \frac{1}{\lambda_n} \left(1 + \frac{1}{p_n} \sum_{k=0}^{n-1} p_j\right) \quad \text{(using detailed balance)}$$

$$= \frac{1}{\lambda_n p_n} \sum_{j=0}^{n} p_j$$

By analogous reasoning, $T_n^- = \frac{1}{\mu_n p_n} \sum_{j=n}^{N} p_j$. (See Keilson [7] for more on the actual first passage time distributions.) The general mean first passage time between two states is now seen to be

$$T_{ij} = \sum_{\ell=i}^{j-1} T_\ell^+ \qquad j > i$$

$$T_{ij} = \sum_{\ell=j+1}^{i} T_\ell^- \qquad j < i$$

For the speaker chain, we can say more. First observe that in the speaker chain, λ_n is a decreasing sequence, and μ_n is an increasing sequence. For such a chain, one suspects that $T_n^- > T_{n+1}^-$ and $T_n^+ < T_{n+1}^+$. This is indeed the case. A simple inductive argument goes as follows. (We only prove the inductive step. The basis is a simple computation). From

the expressions derived

$$T_n^- = \frac{1}{\mu_n} + \frac{\lambda_n}{\mu_n} T_{n+1}^-$$

$$T_{n+1}^- = \frac{1}{\mu_{n+1}} + \frac{\lambda_{n+1}}{\mu_{n+1}} T^-_{n+2}$$

From monotonicity, $\frac{1}{\mu_n} \geq \frac{1}{\mu_{+1}}$ and $\frac{\lambda_n}{\mu_n} \geq \frac{\lambda_{n+1}}{\mu_{n+1}}$. By the induction hypothesis, $T_{n+1}^- > T_{n+2}^-$ which implies $T_n^- > T_{n+1}^-$, thus completing the induction. (Similar reasoning works for T_n^+.)

For the speaker chain, $\lambda_n = (N-n)\lambda$ and $\mu_n = n\,\mu$. The birth and death rates become equal at $n = N\frac{\lambda}{\lambda+\mu} = \bar{A}$. For $n > \bar{A}$, a transition to n+1 becomes an "uphill battle" of increasing difficulty. (Similarly for T_n^-, $n < \bar{A}$.) Thus one suspects that the time to go from $\bar{A}$ to $\bar{A} \pm 0(N)$ becomes quite large as $N \to \infty$. An analysis of this has been done by Bellman and Harris [11]. They show that the actual distribution approaches an exponential with a mean that grows very quickly with N. Of more interest to us are mean passage times towards $\bar{A}$, especially from above.

From the previous results we can derive some simple bounds. Recall

$$T_n^- = \left(\sum_{j=n}^{N} p_j\right) \frac{1}{p_n\,\mu_n}$$

$$T_n^- = \frac{1}{n} + \frac{\lambda_n}{\mu_n} T_{n+1}^-$$

The first equation shows $T_n^- \geq \frac{1}{\mu_n}$. Combining the second one

with the inequality $T^-_{n+1} \leq T^-_n$ (shown above) shows $T^-_n \leq \frac{1}{\mu_n - \lambda_n}$, for n s.t. $\mu_n > \lambda_n$ i.e. for $n > \bar{A}$. Let $n_1 > n_2 > \bar{A}$ be given. Then the lower and upper bounds show

$$T_{n_1,n_2} \geq \sum_{j=n_2+1}^{n_1} \frac{1}{j\mu} = \frac{1}{\mu} \sum_{j=n_2+1}^{n_1} \frac{1}{j}$$

$$T_{n_1,n_2} \leq \sum_{j=n_2+1}^{n_1} \frac{1}{j\mu - (N-j)\lambda}$$

$$= \frac{1}{\mu(1+\varepsilon)} \sum_{j=n_2+1}^{n_1} \frac{1}{j - \bar{A}} \quad , \text{ where } \varepsilon = \frac{\lambda}{\mu}$$

These harmonic sums can easily be bounded in terms of the logarithm. Specifically, if x and y are positive integers, with $x < y$, then

$$\log\left(\frac{y+1}{x}\right) \leq \sum_{j=x}^{y} \frac{1}{j} \leq \log\left(\frac{y}{x-1}\right)$$

(Logs are base e.) Applying these bounds to the previous bounds on passage times we obtain

$$\frac{1}{\mu} \log\left(\frac{n_1+1}{n_2+1}\right) \leq T_{n_1,n_2} \leq \frac{1}{\mu(1+\varepsilon)} \log \frac{n_1 - \lceil \bar{A} \rceil}{n_2 - \lceil \bar{A} \rceil} \quad ,$$

where $\lceil\ \rceil$ is the ceiling function.

We now investigate these bounds as N, n_1, n_2 approach infinity, and n_1 and n_2 have some specified growth relative to the mean. That is, we assume n_1 and n_2 approach infinity as

$$n_1 = \lceil \bar{A} \rceil + f_1(N)$$

$$n_2 = \lceil \bar{A} \rceil + f_2)n)$$

where $f_1(N) \geq f_1(N) \geq 1$.

Remarks:

- To avoid technicalities, we only consider "asymptotically commensurate" functions, i.e. for any functions g and h that that we compare, we assume the ratio $g(N)/h(N)$ goes to ∞, to 0, or to a positive limit, as $N \to \infty$. In the last case, we write $g(N) = 0(h(N))$. If $g(N)/h(N) \to 1$, we write $g(N) \sim h(N)$. Clearly, $f_1(N)$ and $f_2(N)$ are at most $0(N)$, and $f_2/f_1 \to \gamma$ where $0 \leq \gamma \leq 1$.

- We neglect all "integer roundoff" errors, e.g. $\bar{A} \sim \lceil \bar{A} \rceil$ and $\frac{n_1+1}{n_2+1} \sim \frac{n_1}{n_2}$ since $\bar{A} = 0(N)$.

With this behavior for n_1 and n_2 specified, the bounds become

$$\frac{1}{\mu} \log \frac{\bar{A} + f_1}{\bar{A} + f_2} \leq T_{n_1,n_2} \leq \frac{1}{\mu(1+\varepsilon)} \log (f_1/f_2)$$

The asymptotic behavior of these bounds is as follows:

Lower Bound

- If either $f_1 \sim f_2$ or $f_1(N)/N \to 0$, the bound approaches 0 since the argument of the log $\to 1$. In the latter case, this is because $\bar{A}$ dominates f_1 and f_2.

- In the other cases, the bound approaches the positive constant

$$\frac{1}{\mu} \log \left[\frac{1 + f_1/\bar{A}}{1 + f_2/\bar{A}}\right]$$

Upper Bound

- This is constant only if $f_1 = 0(f_2)$. If $f_1 \sim f_2$, the constant is zero, otherwise it is positive.

- Since $f_1 \geq f_2$, f_1/f_2 cannot vanish. Thus the only other possibility is $f_1/f_2 \to \infty$, in which case the bound grows as log (f_1/f_2).

Thus:

- Both bounds approach 0 if $f_1 \sim f_2$.

- If $f_1 = 0(N)$, $f_2 = 0(N)$, but $f_1 \not\sim f_2$, the bounds approach (different) positive constants.

- Otherwise, they disagree in asymptotic behavior.

The upper bound is somwehat "closer to the truth" in the following sense. (The proof of these results is appendix A.) Suppose n tends to infinity as $\bar{A} \to f(N)$. Then

Region I: $f(N)/\sqrt{N} < \infty$

Then $T_n^- \to 0(\frac{1}{\sqrt{N}})$

Region II: $f(N) = x_N\sqrt{N}$ where $x_N \to \infty$ but $x_N/\sqrt{N} \to 0$

Then $T_n^- \to 0(\frac{1}{x_N\sqrt{N}})$

Region III: $f(N) = O(N)$

Then $T_n^- \rightarrow O(\frac{1}{N})$.

Now recall that the bounds on the one-step passage time T_n^- are

$$\frac{1}{n\mu} = \frac{1}{\mu_n} \leq T_n^- < \frac{1}{\mu_n - \lambda_n} = \frac{1}{\mu(1+\epsilon)} \frac{1}{n-\bar{A}}$$

Thus the one-step upper bound is correct in regions II and III, whereas the one-step lower bound is correct only in region III. (Since the two bounds do agree in region III, $O(\frac{1}{N})$ must be the correct behavior. We say more about the constant in the appendix.)

IV. Effective Service Times

Given a data service rate vector, $\underline{r} = (r_o,\ldots,r_N)$, a message length x, and an initial phase state i, we would like to know the distribution of time it takes to complete service and the phase state at completion (jointly). We first consider the simplest form of this problem in which N=1, $r_o = 1$, and $r_1 = 0$, i.e. we consider a "Markovian server" who is either "on" or "off". In this case (because $r_1 = 0$), the problem is equivalent to finding the total time T that must elapse until the time accumulated in state 0 is equal to x, given a start in i. Let

$$H_i(t,x) \triangleq \Pr[T \leq t | i \text{ and } x],\ i = 0,1$$

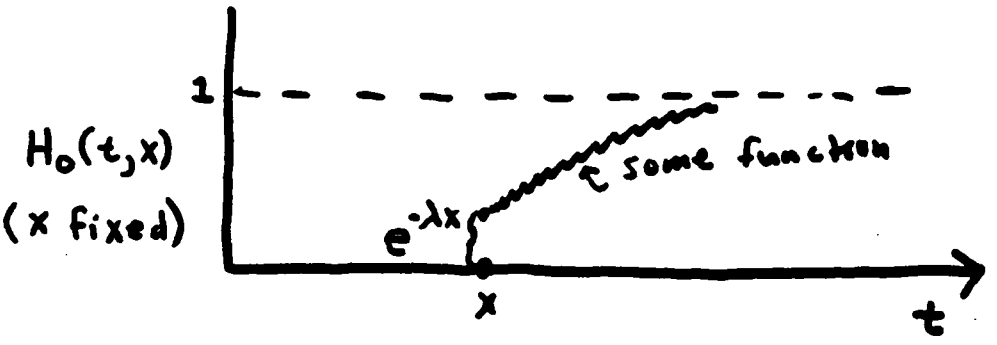
To find H_i, we first compute a related quantity. If the chain starts in state i, and we observe it for a time t, what is the amount of time w spent in state 0. Let

$$F_i(t,x) \triangleq \Pr[w \leq x | i \text{ and observation for } t].$$

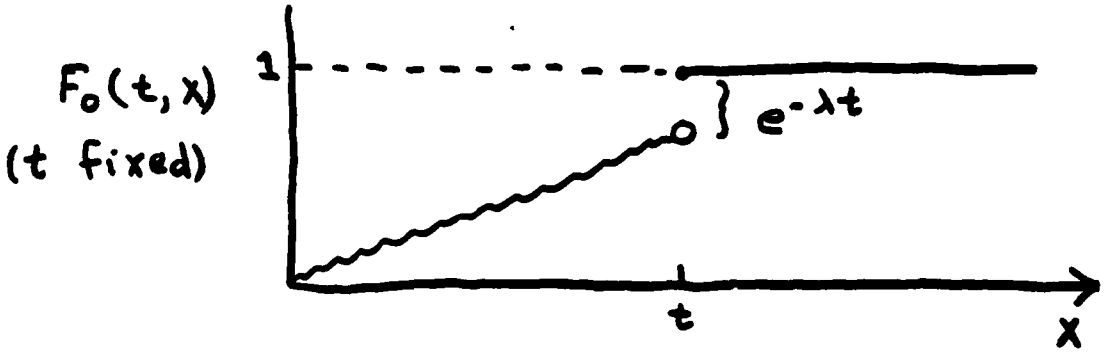
F and H are related as follows. The event $(T \leq t | x,i)$ occurs iff the event $(w > x | i,t)$ occurs. Therefore $H_i(t,x) =$ $= 1 - F_i(t,x) + \Pr[w = x | i,t]$. (There may be impulses so we have to worry about the "=" part of "$\geq$".)

Before computing F_i, we make some observations about H_i. First, if the chain starts in i=1, the amount of time needed to finish x is the amount of time needed to reach i=0

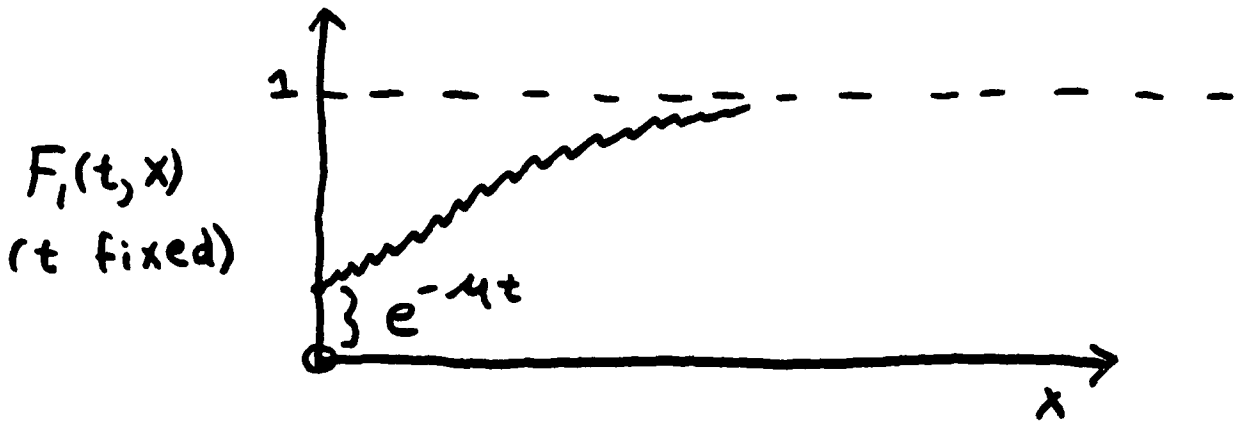
(exponentially distributed with mean μ^{-1}) plus the amount of time needed to finish x given a start in i=0. Thus $H_1 = H_0 * \mu_e^{-\mu t}$. Second, the event $(T < x|0)$ is impossible while the event $(T = x|0,x)$ occurs with probability $e^{-\lambda x}$ (the probability that the chain first leaves i=0 after time x). Thus the function $H_0(t,x)$ has the following general shape



By similar reasoning, $F_0(t,x) = 1$ for $x \geq t$ and has a jump of height $e^{-\lambda t}$ at $x = t^-$.



And $F_1(t,0) = e^{-\mu t}$, the probability that the chain leaves state 1 after time t so that no time in 0 is accumulated.



F_0 and F_1 are related by the following integral equations which can be obtained by conditioning arguments.

$$F_0(t,x) = \int_0^t \lambda e^{-\lambda v} F_1(t-v,x-v)dv + e^{-\lambda t} I(x \geq t)$$

$$F_1(t,x) = \int_0^t \mu e^{-\mu v} F_0(t-v,x)dv + e^{-\mu t} I(x \geq 0)$$

where $I(\ )$ is the indicator function.

Taking the Laplace-Stieltjes transform on x and the Laplace transform on t ($t \leftrightarrow z$, $x \leftrightarrow s$) gives

$$\hat{\hat{F}}_0(z,s) = \hat{\hat{F}}_1(z,s) \frac{\lambda}{\lambda + s + z} + \frac{1}{\lambda + s + z}$$

$$\hat{\hat{F}}_1(z,s) = \hat{\hat{F}}_0(z,s) \frac{\mu}{\mu + z} + \frac{1}{\mu + z}$$

which imply $\hat{\hat{F}}_0(z,s) = \frac{\lambda + \mu + z}{(\lambda+s+z)\cdot(\mu+z) - \lambda\mu}$.

As one check on correctness, we can compute the expected amount of time spent in state 0 as a function of the observation time t, and then take its Laplace transform. This should equal

$$- \frac{d}{ds} \hat{\hat{F}}_0(z,s)\Big|_{s=0} .$$

The expected time spent in 0 given a start in 0 is

$E \int_0^t I(\text{state}(v) = 0 | \text{state } (0) = 0)dv = \int_0^t p_{oo}(v)dv$, where $p_{oo}(v)$ is the transition probability, $p_{oo}(v) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)v}$. (See Chap. III .) Thus the expected time is $\frac{\mu t}{\lambda+\mu} + \frac{\lambda}{(\lambda+\mu)^2}[1-e^{-(\lambda+\mu)t}]$. This expression can be interpreted as $p_o t$ (p_o = equilibrium fraction of time spent in 0) plus a "bias" term reflecting the start in 0. The asymptotic effect of this bias is to add a constant $\frac{\lambda}{(\lambda+\mu)^2}$. The Laplace transform of this is

$$\frac{\mu}{\lambda+\mu} \frac{1}{z^2} + \frac{\lambda}{(\lambda+\mu)^2} [\frac{1}{z} - \frac{1}{\lambda + \mu + z}],$$

and one can verify equality.

Inverting the double transform $\hat{\hat{F}}_0(z,s)$ on the s variable yields

$$\hat{F}_0(z,x) = \frac{\lambda + \mu + z}{\mu + z} \exp(-z \frac{(\lambda + \mu + z}{z + \mu} x).$$

The inversion on the z variable is considerably more tedious.

The answer is

$$f_0(t,x) = \begin{cases} e^{-\lambda x}\delta(t-x) + e^{-x\lambda}e^{-(\lambda-x)}[\phi(t,x)], & 0 \le x \le t \\ 0 & x > t \end{cases}$$

where $\phi(t,x) = \lambda\, I_0(2\sqrt{u})) + \lambda\mu \quad x \dfrac{I_1(2\sqrt{u})}{\sqrt{u}}$, $u = \lambda\mu x(t-x)$, I_0 and I_1 are modified Bessel functions. Because we took Laplace-Stieltjes transform on the x variable, $f_0(t,x)$ is a probability density as a function of x for t fixed.

From the previous discussion then, $H_0(t,x) = \int_{x^-}^{\infty} f_0(t,v)dv$. There appears to be no simple form for this integral. However, one can compute the mean of T, i.e. the mean total time needed to complete the amount of work x given a start in 0. From probability theory this is $\int_0^{\infty} [1 - H_0(t,x)dt$. The evaluation is tedious, so we only give the answer -- $(1 + \frac{\lambda}{\mu})x$. This may be somewhat surprising at first in that there is no "bias" term. The following argument shows that this lack of bias and linearity occur because $r_1 = 0$ and because of the memoryless property of the exponential distribution. Consider the time to complete an amount of work 2x, given a start in state 0. Because $r_1 = 0$, work is only done in state 0, so that when the first x is completed, the chain must be in state 0. Now suppose the first x is completed after some time u has elapsed since the chain entered state 0 on the visit of completion. (This need not be the initial visit.) Because of the memoryless property,

the time to complete the second x is independent of the past, i.e. is independent of u and the number of the visits. Thus the time to complete the second x is a probabilistic replica of the time to complete the first x. Since x is arbitrary, this shows that the mean completion time is linear in the amount of work, given a start in state 0. For large x, we would expect that the mean time to complete x is approximately $\frac{x}{p_0} = (1+\frac{\lambda}{\mu})x$. Combining this with linearity, it follows that $(1+\frac{\lambda}{\mu})x$ must be the exact expression.

We can in fact obtain the Laplace-Stieltjes transform of $H_0(t,x)$. Let $H_0(z,x) = \int_0^\infty e^{-zt}\, d_t\, H_0(t,x)$

Recall that $H_0(t,x) = \int_x^\infty f_0(t,v)dv$. For t fixed, $f_0(t,v)$ is a pdf having an impulse of wight $e^{-\lambda t}$ at $v = t$, and is 0 for $v > t$. For $v < t$, it has a term (the Bessel function part) which we call $g_0(t,v)$, defined for $v < t$. Then

$$H_0(t,x) = \begin{cases} e^{-\lambda x}\, I(t \geq x) + \int_x^t g_0(t,v)dv & x \leq t \\ 0 & x > t \end{cases}$$

Again the calculation is laborious but straightforward so details are omitted. The answer is

$$\hat{H}_0(z,x) = \exp(-\frac{z(z+\lambda+\mu)}{z+\mu}\, x)$$

Now we may consider x itself to be random with distribution B(x). In terms of a queue, the time needed to complete x is the effective "service time" of a message which has a random length distributed at B(x). (If the message arrives when the system is empty, it may find the server in state 1. In this case, the service time will have an additional component -- the time for the server to return to state 0.) The Laplace-Stieltjes transform of the effective service time starting in state 0 is

$$\int_0^\infty d\,B(x)dx \int_0^\infty e^{-zt} H_o(t,x)dt = \int_0^\infty \hat{H}_o(z,x)d\,B(x) = \hat{B}\left(\frac{z(z+\lambda+\mu)}{z+\mu}\right)$$

where $\hat{B}(z) = \int_{-0}^\infty e^{-zx} d\,B(x)dx$. The mean effective service time is then $-\frac{d}{dz}\,\hat{B}\left(\frac{z(z+\lambda+\mu)}{z+\mu}\right)\Big|_{z=0} = (1+\frac{\lambda}{\mu})\bar{x}$, where $\bar{x} = \int_0^\infty x\,d\,B(x)$.

This is consistent with previous results. If $B(x) \sim \exp(\xi)$ the formula for the transform is $\frac{\xi(\mu+z)}{(\mu+z)\cdot(\xi+z)+\lambda z}$.

Because of the memoryless property of the exponential distribution, we can derive the last expression more directly. Let $P_i(t) = Pr[T \le t | i,\ x \sim B(x)]$, $i = 0,1$.
Then

$$P_o(t) = \int_0^t \lambda e^{-\lambda v} [Pr(x \le v) + Pr(x > v)Pr(T \le t-v | 1, x>v)]dv$$

$$+ \int_t^\infty \lambda e^{-\lambda v} Pr[x \le t]dv$$

The crucial simplification is that the underscored term is just $P_1(t-v)$, because of memorylessness.

$$P_o(t) = \int_0^t \lambda e^{-\lambda v}[1 - e^{-\xi v} + e^{-\xi v} P_1(t-v)]dv + (1 - e^{-\xi t})e^{-\lambda t}$$

$$P_1(t) = \int_t^t \mu e^{-\mu v} P_o(t-v)dv.$$

Taking Laplace transforms and solving for $\hat{P}_o(z)$ yields $\hat{P}_o(z) = \frac{\xi(z+\mu)}{(z+\mu)\cdot(z+\xi)+\lambda z} \cdot \frac{1}{z}$. As $\hat{P}_o(z)$ is the transform of the distribution, we conclude that the transform of the density is $z\hat{P}_o(z)$ which agrees with the previous expression.

The completion time analysis for exponential length messages can be extended to the general case of N speakers and arbitrary service rates $\{r_i\}$. Let $T_i(\xi)$ denote the completion time of a message whose length is exponentially distributed with mean ξ^{-1} given a start in phase state 1. And let $\underline{\hat{T}}(z,\xi) \triangleq$ $\triangleq (\hat{T}_o(z,\xi),\ldots,\hat{T}_N(z,\xi))$ denote the vector of Laplace-Stieltjes transforms. Then a conditioning argument similar to the one above shows

$$[z\,\underline{I} + \xi\underline{r}_D - \underline{Q}]\,\underline{\hat{T}}(z,\xi) = \xi\,\underline{r}$$

where $\underline{Q}$ is the generator for the speech process, and $\underline{r}_D$ is the diagonal matrix obtained from the service rate vector $\underline{r}$. Using the transform to extract moments we obtain $\underline{\bar{T}}(\xi) = (\bar{T}_o(\xi),\ldots,\bar{T}_N(\xi)) = (\xi\underline{r}_D - \underline{Q})^{-1}\,\underline{1}$, and $\underline{\overline{T^2}}(\xi) = (\overline{T_o^2}(\xi),\ldots,\overline{T_N^2}(\xi)) = 2(\xi\underline{r}_D - \underline{Q})^{-2}\,\underline{1}$. If a message initiates service in a state

chosen according to the speaker equilibrium distribution, $\underline{p}$, then its mean service time is $\underline{p}^T \cdot \underline{\bar{T}}(\xi)$, and the variance is $\underline{p}^T \cdot \underline{T}^2(\xi) - (\underline{p}^T \cdot \underline{\bar{T}}(\xi))^2$. It is possible to show that if $\lambda, \mu \to \infty$, with $\frac{\lambda}{\mu}$ remaining fixed then $\hat{\underline{T}}(z,\xi)$ approaches $\frac{\xi\bar{r}}{z + \xi\bar{r}} \underline{1}$. That is, the service rate effectively becomes deterministic at the average rate $\bar{r} = \underline{r}^T \cdot \underline{p}$, so that for any $i, T_i(\xi)$ approaches an exponential with mean $(\xi\bar{r})^{-1}$ in distribution. This is because the speaker process passes through its states "infinitely" often during a service, and the fraction of time it spends in state i approaches p_i.

The previous characterization immediately generalizes if we seek the joint distribution of completion time and the speaker state at completion. Let $T_{ij}(t,\xi) = \Pr[\text{message is completed within time } t; \text{ the completion state is } j | \text{initial state } i]$ and let $\hat{T}_{ij}(Z,\xi)$ be the Laplace-Stieltjes transform. (Note $T_{ii}(t,\xi)$ is a possibly defective distribution, i.e. $\int_0^\infty d_t\, T_{ij}(t,\xi) < 1$, since completion need not occur at j). Then a conditioning argument shows

$$[Z\,\underline{I} + \xi\underline{r}_D - \underline{Q}]\,\hat{\underline{T}}(Z,\xi) = \xi\underline{r}_D \,.$$

Note that $\hat{\underline{T}}(0,\xi)$ is a transition probability matrix on the speaker state space itself, i.e. $\hat{\underline{T}}_{ij}(0,\xi) = \Pr[\text{message completed in } j | \text{start in } i]$. We will use $\hat{\underline{T}}(Z,\xi)$ to compute waiting times in Chapter VI.

V. The M/M Case

A. Orientation

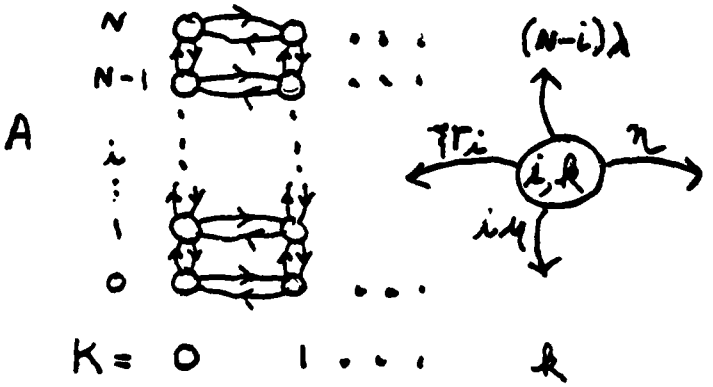
In this chapter, we assume that data arrivals are Poisson with rate η , and message lengths ~exp(ξ). $\underline{r} = (r_0,\ldots,r_N)$ denotes the vector of service rates. With these assumptions, the vector process [A(t), K(t)], where K(t) is the number of messages in the system, is a Markov process. Its state diagram is a two-dimensional grid with a state (i,k) having transition rates

$(i,k) \to (i+1,k)$ at rate $\lambda_i = (N - i)\lambda$

$(i,k) \to (i-1,k)$ at rate $\mu_i = i\mu$

$(i,k) \to (i,k+1)$ at rate η

$(i,k) \to (i,k-1)$ at rate ξr_i if $k > 0$



Notice that the last transition rate $r_i\xi$ has units of messages/sec. (as it must) and depends on r_i and ξ only through their product. For convenience, we take $\xi = 1$.

In looking at the diagram, we notice that if a column is viewed as a superstate, the "column process" is a multidimensional analog of the classical M/M/1 birth-death process in that the transitions into and out of any column have the same structure (except at K=0 where the process is truncated). The voice/data model is in fact a special case of the more general quasi-birth-death (QBD) or column continuous processes. These are time-homogeneous, discrete-state bivariate Markov chains $[X(t),Y(t)]$ for which

1) A transition $(x_1,y_1) \rightarrow (x_2,y_2)$ can occur only if $|y_1 - y_2| \leq 1$, whence the term column continuous. (Also called skip-free left and right.)

2) The rate of a transition within a column, $(x_1,y) \rightarrow (x_2,y)$, is independent of y.

The QBD processes for which the inter-column rates, i.e. $(x_1,y) \rightarrow (x_2,y\pm1)$, are also independent of y (except at boundaries) constitute an important subset. We term these homogeneous QBD processes with the understanding that complete spatial homogeneity might not be present because of boundaries. The voice/data model belongs to this subset and is even more restricted because $(x_1,y) \rightarrow (x_2,y\pm1)$ can occur only if $x_1 = x_2$. This restriction coupled with 2) shows that the marginal process $X(t)$ (in our case $A(t)$) is a Markov process. In this sense, the speaker process is an "independent" phase process "modulating" the transitions of K (though A and K

are generally not statistically independent). Notice that in the general QBD process, the marginal process X(t) is not Markov because the rate of transitions $(x_1,y) \to (x_2,y\pm1)$ can depend on y. An exception to this is the subset of truly homogeneous QBD processes, i.e. those in which the state space of Y is all the integers. For in this case, changes in X are independent of Y.

The homogeneous QBD processes in general can be viewed as multidimensional analogs of the classical M/M/1 birth-death process (i.e. even if $(x_1,y) \to (x_2,y\pm1)$, $x_1 \neq x_2$ is allowed). Thus one might expect ergodic distributions that are matrix analogs of the geometric or truncated geometric ergodic distributions of the M/M/1 queue. This is indeed the case.(Perhaps we should say that, a posteriori,the following makes sense.)

Assume [X,Y] is a homogeneous QBD process in which the state space of X is $\{0,1,\ldots,N\}$ and that of Y is $\{0,1,2,\ldots\}$. Let $\underline{e}(y) = (e_{0y},\ldots,e_{Nj})$ denote the vector of the joint ergodic probabilities for the y^{th} column. Then we will show that

$$\underline{e}^T(y) = \underline{e}^T(0)\ \underline{\theta}^y$$

where $\underline{\theta}$ is similar to a strictly substochastic matrix. (Hence $\underline{\theta}$'s eigenvalues are inside the unit circle, as they must be for this to be meaningful.) This form for the solution is apparently due to Evans and Wallace. (Our historical information was obtained through Neuts [12] and Keilson [13]. For more references and a more complete study of matrix-geometric

methods see Neuts [12].) Independent derivations were subsequently obtained by Keilson and Neuts. The approach was brought to our attention by Keilson who was working on similar Markov models in other contexts when we brought the voice/data problem to him. An exposition of some of his results can be found in [14]. (In this paper, the roles of columns and row are reversed from our usage, so he terms them row continuous.)

In the remainder of this chapter, we explore the application of the matrix geometic method to the voice/data model. As indicated, the general theoretical questions have largely been solved. However, the voice/data model is a restricted case, and we have been able to characterize certain quantities in some detail. To maintain the continuity of development, we do not always explicitly separate those theoretical results that are particular to the voice/data model (and hence new to us). Where possible, we provide references for previously known results. Undoubtedly, we have overlooked some authors and apologize for this. But the queuing theory literature is so vast and diffuse, that a complete literature search is not appropriate unless one is doing a survey paper. This was not our intent.

Our development will use Keilson's approach as a theoretical guideline. This is based on the so called compensation method.(See Keilson [15].) Before proceeding with this, we briefly discuss the z-transform approach.

One first defines the partial z-transforms

$F_i(z) = \sum_{k=0}^{\infty} e_{ik} z^k$, and then using the global balance equations, (which are difference equations of degree 2 in both the i and k coordinates) obtains the relation $\underline{F}^T(z)\underline{A}(z) = \underline{e}^T(0)\underline{B}(z)$, where $\underline{F}(z) = (F_0(z),\ldots,F_N(z))$, and the entries of $\underline{A}(z)$ and $\underline{B}(z)$ are polynomials in z of degree 2 or less. Thus $\underline{F}^T(z) = \underline{e}^T(0)\underline{B}(z)\underline{A}^{-1}(z)$. The matrix $\underline{A}^{-1}(z)$ has poles in the region $|z| < 1$, and one can, in principle, solve for $\underline{e}(0)$ by using the requirement that $\underline{F}(z)$ must be analytic for $|z| < 1$. (The analyticity condition places N constraints on $\underline{e}(0)$. The (N+1)st comes from a "conservation equation.)

This approach shows that the $F_i(z)$ are rational functions and hence that the solution is a sum of geometrically decaying terms. (The decay rates must be eigenvalues of $\underline{\theta}$.) However, one cannot easily extract from the transform solution the manner in which the (possibly complex) roots combine into a real matrix representation $\underline{\theta}$. Also, the numerical inversion of the z-transforms is less attractive than the methods to be presented.

The transform approach was our first approach, and later we found that the same results had been obtained by Yechiali and Naor [16], [17].

B. Stability and Existence of Ergodic Distributions

In the general study of queues and backlog processes, "stability" is usually present iff the "service rate" exceeds the "arrival rate". The precise mathematical formulation of these notions depends on the particular model. Typically, a weak notion of stability is that the emptying of the queue or backlog should be a recurrent event. One can then further require, for example, that the waiting times converge in distribution to a random variable having a certain number of moments etc. For many classes of queuing models, the various requirements are equivalent, but it is not our purpose to discuss this general problem.

For Markov chain queuing models, stability means that the Markov chain is ergodic. For finite Markov chains, ergodicity reduces to the purely structural question of irreducibility, but for infinite chains one must further require some "net return force toward the origin". In the case of Markov queuing models, this translates into the arrival rate < service rate condition. The voice/data model poses no structural barriers since there is a path between any two states provided that at least one $r_i > 0$. The other condition is satisfied if what we term the drift, $\eta - \Sigma p_i r_i = \eta - \bar{r}$, is negative, and we assume this throughout the development. As our interest is in computing various statistics and not in "existence" results, we do not formally prove that the negative drift condition is necessary

and sufficient. At various places, "plausibility arguments" will become apparent, e.g. as the drift approaches 0, the maximal eigenvalue of $\underline{\theta}$ approaches 1. For a formal proof and general discussion of queues with dependent interarrival or service times, the reader is referred to Loynes [18], [19], [20].

C. Derivation of Solution

First consider the process $[A(t), K^H(t)]$ obtained by removing the boundary at zero and extending the columns out to $-\infty$. This process is spatially homogeneous in the k variables because the transition rates depend only on A. Because of homogeneity, the transition probability $\Pr[A(t) = j, K^H(t) = k_1 | A(0) = i, K^H(0) = k_o]$ depends only on $k_1 - k_o$. We take $K^H = 0$ as the origin and define $\underline{\underline{g}}(k,t)$ by

$$g_{ij}(k,t) = \Pr[A(t) = j, K^H(t) = k | A(0) = i, K^H(0) = 0]$$

The asymptotic behavior of $K^H(t)$ as $t \to \infty$ depends on the drift $(\eta - \underline{r})^T \underline{p}$. If the drift is negative, the homogeneous process drifts to $K^H = -\infty$.

Define the k step right first passage time density matrix $\underline{s}^+(k,t)$ by

$$s^+_{ij}(k,t)dt = \Pr[K^H = k \text{ is first reached at } t, t+dt \text{ and } A(t) = j | A(0) = i,\ K^H(0) = 0]$$

for $k \geq 1$. The special case $\underline{s}^+(1,t)$ is denoted by $\underline{s}^+(t)$. Similarly define $\underline{s}^-(k,t)$ for k steps left, i.e. $k \leq -1$. Using conditioning arguments, one can show

$$\underline{s}^+(k,t) = \overbrace{\underline{s}^+(t) * \ldots * \underline{s}^+(t)}^{k \text{ times}}$$

$$\underline{s}^-(k,t) = s^-(t) * \ldots * s^-(t)$$

$$\underline{\underline{g}}(k,t) = \begin{cases} \underline{s}^+(k,t) * \underline{\underline{g}}(t) & k \geq 1 \\ \underline{s}^-(k,t) * \underline{\underline{g}}(t) & k \leq -1 \end{cases}, \quad \underline{\underline{g}}(t) \triangleq \underline{\underline{g}}(0,t)$$

where * denotes the matrix-convolution product,

$$[\underline{a}(t) * \underline{b}(t)]_{ij} = \sum_k \int_0^t a_{ik}(\tau)\, b_{kj}(t-\tau)d\tau \ .$$

Let $\underline{\sigma}^+(k,\omega)$, $\underline{\sigma}^-(k,\omega)$, $\underline{\gamma}(k,\omega)$ be the Laplace transforms of $\underline{s}^+(k,t)$, $\underline{s}^-(k,t)$, $\underline{g}(k,t)$ respectively. Further let $\underline{\sigma}^+(\omega) \triangleq \underline{\sigma}^+(1,\omega)$; $\underline{\sigma}^-(\omega) \triangleq \underline{\sigma}^-(1,\omega)$; $\underline{\gamma}(\omega) \triangleq \underline{\gamma}(0,\omega)$ and let $\underline{\sigma}^+ \triangleq \underline{\sigma}^+(\omega=0)$; $\underline{\sigma}^- \triangleq \underline{\sigma}^-(\omega=0)$, $\underline{\gamma} \triangleq \underline{\gamma}(\omega=0)$. Note for example, $\underline{\sigma}^+ = \int_0^\infty \underline{s}^+(t)dt$ so that the i^{th} row-sum of $\underline{\sigma}^+$ is the probability that the homogeneous process ever reaches the set $\{(\cdot,1)\}$ given that it starts in $(i,0)$. (The set $\{(\cdot,i)\}$ denotes the set of states in the column $K^H = i$.) An analogous interpretation applies to $\underline{\sigma}^-$. Because the drift is negative, $\underline{\sigma}^+$ is strictly substochastic as the process may never reach $K^H = 1$ from $K^H = 0$. Similarly, $\underline{\sigma}^-$ is stochastic since the process K^H does eventually decrease, with probability one. In the transform domain, the previous relations become

$$\underline{\sigma}^+(k,\omega) = [\underline{\sigma}^+(\omega)]^k$$

$$\underline{\sigma}^-(k,\omega) = [\underline{\sigma}^-(\omega)]^k$$

$$\underline{\gamma}(k,\omega) = \begin{cases} [\underline{\sigma}^+(\omega)]^k\, \underline{\gamma}(0,\omega) & k \geq 1 \\ [\underline{\sigma}^-(\omega)]^k\, \underline{\gamma}(0,\omega) & k \leq -1 \end{cases}$$

Now to the crux.

Keilson [15] has shown that if a boundary is inserted in the homogeneous process at $K^H = 0$ then

1) There exists a "compensation measure" $\underline{\Gamma}(k)$, $-\infty < k < \infty$ s.t. $\underline{e}(k)$ is a convolution in k of $\underline{\gamma}(k)$ and $\underline{\Gamma}(k)$. ($\underline{\gamma}(k) \triangleq \underline{\gamma}(k,\omega=0) = \int_0^\infty \underline{g}(k,t)dt$.)

Specifically

$$\underline{e}^T(k) = \sum_{\ell=-\infty}^{\infty} \underline{\Gamma}^T(\ell)\, \underline{\gamma}(k-\ell) \qquad k \geq 0.$$

2) $\underline{\Gamma}(k) = \underline{0}$ except at $k = -1$ and $k = 0$ so that

$$\underline{e}^T(k) = \underline{\Gamma}^T(0)\underline{\gamma}(k) + \underline{\Gamma}^T(-1)\underline{\gamma}(k+1)$$

3) $\underline{\Gamma}(-1) + \underline{\Gamma}(0) = \underline{0}$, $\underline{\Gamma}(-1) \leq \underline{0}$ so that

$$\underline{e}^T(k) = \underline{\Gamma}^T(0)[\underline{\gamma}(k) - \underline{\gamma}(k+1)]$$

Using the previous relations we then obtain

$$\begin{aligned}
\underline{e}^T(k) &= \underline{\Gamma}^T(0)[(\underline{\sigma}^+)^k - (\underline{\sigma}^+)^{k+1}]\underline{\gamma} \\
&= \underline{\Gamma}^T(0)[\underline{I} - \underline{\sigma}^+](\underline{\sigma}^+)^k\, \underline{\gamma} \\
&= \underline{\Gamma}^T(0)[\underline{I} - \underline{\sigma}^+]\underline{\gamma}\, \underline{\gamma}^{-1}(\underline{\sigma}^+)^k\, \underline{\gamma} \\
&= \underline{e}^T(0)\underline{\theta}^k
\end{aligned}$$

where $\underline{e}^T(0) = \underline{\Gamma}^T(0)[\underline{I} - \underline{\sigma}^+]\underline{\gamma}$ and $\underline{\theta} = \underline{\gamma}^{-1}\, \underline{\sigma}^+\, \underline{\gamma}$. Notice that $\underline{\theta}$ is similar to the strictly substochastic matrix $\underline{\sigma}^+$. Hence the eigenvalues of $\underline{\theta}$ are strictly inside the unit circle. This implies that the series $\sum_{\ell=0}^{\infty} \underline{\theta}^{\ell}$ is convergent and equals $[\underline{I}-\underline{\theta}]^{-1}$. From the relation $\sum_{k=0}^{\infty} \underline{e}(k) = \underline{p}$, we then obtain $\underline{e}^T(0) = \underline{p}^T(\underline{I} - \underline{\theta})$. Thus we do not need to explicitly find the compensation measure $\underline{\Gamma}(0)$ if $\underline{\theta}$ can be found. At this point we need to find $\underline{\gamma}$ and $\underline{\sigma}^+$. It turns out that $\underline{\gamma}$ is easily obtainable once $\underline{\sigma}^+$ and $\bar{\underline{\sigma}}$ are known, so we first proceed with $\underline{\sigma}^+$ and $\underline{\sigma}^-$.

Remark: The compensation technique also applies if K has a boundary at some positive integer, say M. In this case, the compensation measure also has mass at k = M and k = M+1. Thus one obtains a term of the form $\underline{\Gamma}^T(M)[\underline{\gamma}(k-M) - \underline{\gamma}(k-M-1)]$, and this can be expressed in terms of $\underline{\gamma}$ and powers of $\underline{\sigma}^-$.

D. Calculation of $\underline{\sigma}^+$ and $\underline{\sigma}^-$

For the homogeneous process define the matrix $\underline{\alpha}$ by

$$\alpha_{ij} = \Pr[K^H \text{ departs } k = 0 \text{ by going to } K^H = 1,\ A = j \text{ when this occurs } \begin{bmatrix} A(0) = i \\ K^H(0) = 0 \end{bmatrix}]$$

Similarly define $\underline{\beta}$ for going to $K^H = -1$. Notice the distinction between $\underline{\alpha}$ and $\underline{\sigma}^+$ and $\underline{\beta}$ and $\underline{\sigma}^-$. The $\underline{\sigma}$'s are first passage probabilities whereas $\underline{\alpha}$ and $\underline{\beta}$ are the $K^H = +1$, $K^H = -1$ probabilities when K^H leaves 0 (either to $K^H = +1$ or $K^H = -1$). If either $\underline{n} \neq 0$ or $\underline{r} \neq 0$, then $\underline{\alpha} + \underline{\beta}$ is stochastic since the process must eventually leave $K^H = 0$. The matrix $\underline{\alpha} + \underline{\beta}$ is in fact a transition probability matrix (on the speaker state space) for the embedded discrete time chain defined at the instants of changes in K^H.

As one suspects, there is a relationship among the $\underline{\alpha}$, $\underline{\beta}$, and $\underline{\sigma}$'s. Specifically

$$\underline{\sigma}^+ = \underline{\alpha} + \underline{\beta}(\underline{\sigma}^+)^2$$
$$\underline{\sigma}^- = \underline{\beta} + \underline{\alpha}(\underline{\sigma}^-)^2$$

The first equation says that the first passage to $K^H = 1$ may occur a) at the first departure from $K^H = 0$, the $\underline{\alpha}$ term, or b) if the first departure is to $K^H = -1$, reflected by the $\underline{\beta}$ term, then two first passages to the right are required to reach $K^H = +1$. Because of homogeneity, the transition from -1 to 0 has the same probabilistic structure as the one from 0 to 1. Thus the $(\underline{\sigma}^+)^2$ appears. A similar interpretation applies to $(\underline{\sigma}^-)^2$.

The iteration $\underline{\sigma}_{\ell+1} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}_{\ell}^{2}$. $\underline{\sigma}_{0} = 0$ (or reverse $\underline{\alpha}$ and $\underline{\beta}$ for $\underline{\sigma}^{-}$) does converge to the probabilistically correct solution. This algorithm is discussed later.

The matrices $\underline{\alpha}$ and $\underline{\beta}$ can easily be found. By a conditioning argument

$$\alpha_{ij} = \frac{\lambda_i}{\nu_i}\,\alpha_{i+1,j} + \frac{\mu_i}{\nu_i}\,\alpha_{i-1,j} + \frac{\eta}{\nu_i}\,\delta_{ij}$$

where $\nu_i = \eta+\lambda_i+\mu_i+ r_i$ = total exit rate out of a state $(i,\cdot)$ and δ_{ij} = Kronecker delta. In matrix form these equations are

$$\underline{\alpha} = (\underline{\eta}_D + \underline{r}_D - \underline{Q})^{-1}\,\underline{\eta}_D$$

where $\underline{Q}$ is the generator for the speaker process. Similarly $\underline{\beta} = (\underline{\eta}_D + \underline{r}_D - \underline{Q})^{-1}\,\underline{r}_D$. We can put these matrices into a form which more clearly indicates their probabilistic meaning. Let $\underline{\nu}_D = (\nu_i\,\delta_{ij})$ and let $\hat{\underline{Q}} = \underline{Q} + \mathrm{diag}(\lambda_i + \mu_i)$. Then we can write the equations as

$$\underline{\alpha} = (\underline{I} - \underline{\nu}_D^{-1}\,\hat{\underline{Q}})^{-1}\,\underline{\nu}_D^{-1}\,\underline{\eta}_D$$

$$\underline{\beta} = (\underline{I} - \underline{\nu}_D^{-1}\,\hat{\underline{Q}})^{-1}\,\underline{\nu}_D^{-1}\,\underline{r}_D$$

Essentially, we are considering the embedded discrete time chain. The matrix $\underline{\nu}_D^{-1}\,\hat{\underline{Q}}$ is a strictly substochastic matrix giving the probabilities that when a transition occurs, the K^H coordinate does not change. Analogously $\underline{\nu}_D^{-1}\,\underline{\eta}_D$ and $\underline{\nu}_D^{-1}\,\underline{r}_D$ give the probabilities that when a trasition occurs, K^H goes to +1, -1 respectively. Since $\underline{\nu}_D^{-1}\,\hat{\underline{Q}}$ is strictly substochastic,

$[\underline{I} - \underline{\nu}_D^{-1} \hat{\underline{Q}}]^{-1} = \sum_{\ell=0}^{\infty} (\underline{\nu}_D^{-1} \hat{\underline{Q}})^{\ell}$. The probabilistic meaning of these equations as accounting identities is as follows. A departure to $K^H = +1$, for example, occurs when there are, say, ℓ transitions at which K^H does not change -- this is reflected by the $(\underline{\nu}_D^{-1} \hat{\underline{Q}})^{\ell}$ term -- followed by a departure to $K^H = +1$ -- for which $\underline{\nu}_D^{-1} \underline{\eta}_D$ is the appropriate transition matrix. A simple calculation will indicate that $\underline{\alpha} + \underline{\beta} = (\underline{I} - \underline{\nu}_D^{-1} \hat{\underline{Q}})^{-1} \underline{\nu}_D^{-1}(\underline{\eta}_D + \underline{r}_D)$ is stochastic.

So far, we have not really relied on the particular properties of the voice/data model, except in using the relation $\sum_{k=0}^{\infty} \underline{e}(k) = \underline{p}$ to express $\underline{e}(0)$ in terms of $\underline{\theta}$ and $\underline{p}$. The procedure in fact, does work for general, homogeneous QBD processes. In the general case, $\underline{Q}$ is the generator for some Markov chain defined on $\{0,1,\ldots,N\}$, and $\underline{\eta}_D$ and $\underline{r}_D$ are replaced by general transition rate matrices $\underline{\eta}$ and $\underline{r}$, e.g. η_{i_1,i_2} is the rate at which $(i_1,k) \rightarrow (i_2,k+1)$ transitions occur. We later show that $\underline{e}(0)$ can be characterized as a left PRF eigenvector of a stochastic matrix in the general case, and $\underline{e}(0)$ is then uniquely specified by the requirement the $\sum_{i,k} e_{i_k} = 1$ or $\underline{e}^T(0)\ [\underline{I} - \underline{\theta}]^{-1}\ \underline{1} = 1$. But as indicated, the relation $\sum_{k=0}^{\infty} \underline{e}(k) = \underline{p}$, where $\underline{p}$ is the stationary distribution of the $\underline{Q}$ matrix in question, is generally not true since changes in the marginal "row process" are not totally accounted for by this $\underline{Q}$. Also, the simple drift criterion $\eta - \underline{r}^T \underline{p} < 0$ is replaced by a more

general condition.

For the voice/data model,we can prove the following assertions, assuming $\eta > 0$, $\underline{r} > \underline{0}$.

1) $\underline{\alpha} > \underline{0}$ $\beta > \underline{0}$

2) $\underline{g}^+ > \underline{0}$ $\underline{g}^- > \underline{0}$

3) The eigenvalues of $\underline{\alpha}, \underline{\beta}$ and $\underline{\alpha} + \underline{\beta}$ are real and positive

4) The eigenvalues of $\underline{g}^+$ and $\underline{g}^-$ are real and positive.

Proof:

1) In the speaker chain, any state A = j is reachable from any state A = i. If $\eta > 0$, then the transition $(j,0) \to (j,1)$ can occur $\forall$ j. So the sample paths that start in (i,0), wander in the set $\{(\cdot,0)\}$ until j is reached, and then go to (j,1) at the next transition, have positive probability, i.e. $\alpha_{ij} > 0$, $\forall$ i,j. Similarly, $\underline{r} > \underline{0}$ implies $\underline{\beta} > \underline{0}$.

2) From the matrix quadratic equations for the $\underline{g}$'s, it follows that $\underline{g}^+ \geq \underline{\alpha}$, $\underline{g}^- \geq \underline{\beta}$, so 2) follows from 1).

Remark: It follows from 1) and 2) and our discussion in Chapter II,that $\underline{\alpha}$, $\underline{\beta}$, $\underline{g}^+$, and $\underline{g}^-$ are primitive. Properties 1) and 2) are not too special. In the general homogeneous QBD process, they will follow from the irreducibility of the transition structure within a column and the condition $\underline{1}^T \underline{\eta} > \underline{0}^T$ and $\underline{1}^T \underline{r} > \underline{0}^T$, where $\underline{\eta}$ and $\underline{r}$ are the general rate matrices. The condition $\underline{1}^T \underline{\eta} > \underline{0}^T$ simply means that for each j,

there is some i s.t. an $(i,0) \to (j,1)$ transition is possible. (Similarly for $\underline{r}$.)

3) We use the following lemma.

Lemma: If $\underline{M}_1$ and $\underline{M}_2$ are two positive definite matrices (symmetry is included in our definition), then the eigenvalues of the product $\underline{M}_1 \underline{M}_2$ are real and positive. (Note $\underline{M}_1 \underline{M}_2$ need not be positive definite.)

Proof: $\underline{M}_1 \underline{M}_2 \underline{x} = \epsilon \underline{x}$ implies $\underline{M}_2 x = \epsilon \underline{M}_1^{-1} \underline{x}$, which implies $\underline{x}^{*T} \underline{M}_2 \underline{x} = \epsilon \underline{x}^{*T} \underline{M}_1^{-1} \underline{x}$, where * = complex conjugate. Now, $\underline{M}_1$ positive definite implies $\underline{M}_1^{-1}$ positive definite, so by the definition of positive definite $\underline{x}^{*T} \underline{M}_2 \underline{x}$ and $\underline{x}^{*T} \underline{M}_1^{-1} \underline{x}$ are real and positive. Thus ϵ must be real and positive.

(If either $\underline{M}_1$ or $\underline{M}_2$ is only semi-definite, a similar result holds with $>$ replaced by $\geq$, though the proof is more complicated.)

Now let $\underline{A} = (\underline{\pi}_D + \underline{r}_D - \underline{Q})$ and $\hat{\underline{A}} = \underline{p}_D^{\frac{1}{2}} \underline{A} \underline{p}_D^{-\frac{1}{2}}$, where $\underline{p}_D$ is the diagonal matrix obtained from the speaker ergodic distribution $\underline{p}$. By Girshgorin's theorem, the eigenvalues of $\underline{A}$ have positive real parts, and hence the eigenvalues of $\hat{\underline{A}}$ and $\hat{\underline{A}}^{-1}$ must also. By our remarks on detailed balance, $\underline{p}_D^{\frac{1}{2}} \underline{Q} \underline{p}_D^{-\frac{1}{2}}$ is symmetric, and hence $\hat{\underline{A}}$ and $\hat{\underline{A}}^{-1}$ are, since $\underline{\pi}_D$ and $\underline{r}_D$ are diagonal. Thus $\hat{\underline{A}}^{-1}$ is positive definite. Now $\underline{\alpha} = \underline{A}^{-1} \underline{\pi}_D$ which implies $\underline{p}_D^{-\frac{1}{2}} \underline{\alpha} \underline{p}_D^{\frac{1}{2}} = = \hat{\underline{A}}^{-1} \underline{p}_D^{-\frac{1}{2}} \underline{\pi}_D \underline{p}_D^{\frac{1}{2}} = (\hat{\underline{A}})^{-1} \underline{\pi}_D$ and similarly for $\underline{\beta} = \underline{A}^{-1} \underline{r}_D$ and $\underline{\alpha} + \underline{\beta} = \underline{A}^{-1}(\underline{\pi}_D + \underline{r}_D)$. By assumption, $\pi > 0$, $\underline{r} > \underline{0}$, so $\underline{\pi}_D$, $\underline{r}_D$, and $\underline{\pi}_D + \underline{r}_D$ are positive definite.

Invoking the lemma, we can conclude that $\underline{\alpha}$, $\underline{\beta}$, and $\underline{\alpha} + \underline{\beta}$ are similar to matrices having real positive eigenvalues.

4) We only prove the result for $\underline{\sigma}^+$ since the proof for $\underline{\sigma}^-$ is the same with rhe roles of $\underline{\alpha}$ and $\underline{\beta}$ reversed. For convenience, we drop the "+" from $\underline{\sigma}^+$.

By previous remarks, $\underline{\sigma} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^2$, so $\tilde{\underline{\sigma}} = \tilde{\underline{\alpha}} + \tilde{\underline{\beta}}\,\tilde{\underline{\sigma}}^2$ where $\tilde{\underline{M}} = \underline{p}_D^{-\frac{1}{2}}\,\underline{M}\,\underline{p}_D^{\frac{1}{2}}$, for $\underline{M} = \underline{\alpha}$, $\underline{\beta}$, or $\underline{\sigma}$. Now suppose $\underline{x}$ is an eigenvector of $\tilde{\underline{\sigma}}$ with eigenvalue ε. Then we obtain $\varepsilon\,\underline{x} = \tilde{\underline{\alpha}}\,\underline{x} + \varepsilon^2\,\tilde{\underline{\beta}}\,\underline{x}$. Recall from the proof of 3) that $\tilde{\underline{\alpha}} = (\hat{\underline{A}})^{-1}\,\underline{n}_D$ and $\tilde{\underline{\beta}} = (\hat{\underline{A}})^{-1}\,\underline{r}_D$, where $\hat{\underline{A}}$ is positive definite. Thus $\varepsilon\,\underline{x} = (\hat{\underline{A}})^{-1}\,(\underline{n}_D + \varepsilon^2\,\underline{r}_D)\,\underline{x}$, which implies $\varepsilon\,\hat{\underline{A}}\,\underline{x} = (\underline{n}_D + \varepsilon^2\,\underline{r}_D)\,\underline{x}$. Multiplying both sides on the left by $\underline{x}^{*T}$, we obtain $c_1\,\varepsilon = c_o + c_2\,\varepsilon^2$, where $c_1 = \underline{x}^{*T}\,\hat{\underline{A}}\,\underline{x}$, $c_2 = \underline{x}^{*T}\,\underline{r}_D\,\underline{x}$, $c_o = \eta$. (We assume $\underline{x}$ has unit norm.) From the definition of $\hat{\underline{A}}$ it follows that $c_1 = c_o + c_2 + c_3$, where $c_3 = -\underline{x}^{*T}\,\underline{p}_D^{\frac{1}{2}}\,\underline{Q}\,\underline{p}_D^{-\frac{1}{2}}\,\underline{x}$. From our remarks on detailed balance, $\underline{p}_D^{\frac{1}{2}}\,\underline{Q}\,\underline{p}_D^{-\frac{1}{2}}$ is symmetric. Since $\underline{Q}$ has nonpositive eigenvalues, it follows that $c_3 > 0$. Further, c_o and c_2 are positive since $\eta > 0$, $\underline{r} > \underline{0}$. A straighforward application of the quadratic formula then shows that both roots of the polynomial must be real and positive.

Remark: This indicates that $(\underline{\sigma}^+)^k$ will not exhibit "oscillatory behavior" as $k \to \infty$.

E. Calculation of γ

The transition probability matrix $\underline{g}(t)$, $g_{ij}(t) = \Pr[A(t) = j, K^H(t) = 0 | A(t) = i, K^H(0) = 0]$ satisfies integro-differential equations

$$\dot{\underline{g}}(t) = \underline{g}(t)[\underline{Q} - \underline{n}_D - \underline{r}_D] + (\underline{g}(t)\underline{n}_D)*\underline{s}^-(t) + (\underline{g}(t)\underline{r}_D)*\underline{s}^+(t)$$

$$\dot{\underline{g}}(t) = \underline{g}(t)[\underline{Q} - \underline{n}_D - \underline{r}_D] + \underline{s}^-(t)*(\underline{g}(t)\underline{n}_D) + \underline{s}^+(t)*(\underline{g}(t)\underline{r}_D).$$

(* = matrix-convolution product)

These equations can be derived using the conditioning arguments relating $g(t+dt)$ to $\underline{g}(t)$. In both equations, the first term comes from the fact that $(j,0)$ can be reached at $t + dt$ by being in $(j-1,0)$, $(j,0)$, $(j+1,0)$ at time t and then having a transition up, no transition, transition down, respectively. The matrix $[\underline{Q} - \underline{n}_D - \underline{r}_D]$ incorporates these possibilities. The other terms are in general different in the two equations since matrix multiplication (and hence *) is not commutative. The two equations derive from different conditioning arguments. We sketch the two for the "$\underline{g}\ \underline{n}_D\ \underline{s}^-$" term.

In the first equation, "$\underline{g}\ \underline{n}_D\ \underline{s}^-$" is obtained as follows. We wish to compute $g_{ij}(t+dt)$. One "set of paths" starting in $(i,0)$ at 0 and reaching $(j,0)$ at $(t+dt)$ is the following. At some intermediate time $t - \tau$, the process is in state $(\ell,0)$, with probability $g_{i\ell}(t-\tau)$. In the next dt it goes to $(\ell,1)$ with probability ηdt. Then the <u>first return</u> to the set

$\{(\cdot,0)\}$ occurs $\tau,\tau+d\tau$ later, and the entrance state is $(j,0)$, with probability $s^{-}_{\ell j}(\tau)d\tau$. Therefore $g_{ij}(t+dt)$ contains a term of the form $\sum_{\ell} \eta dt \int_0^t g_{i\ell}(t-\tau)s^{-}_{\ell j}(\tau)d\tau$ so that $\dot{g}_{ij}(t)$ contains a term $\eta\sum_{\ell} g_{i\ell} * s^{-}_{\ell j}$. Note that there is no double counting because $s^{-}_{\ell j}$ is a first passage time density.

In the second equation, the term is obtained as follows. The state $(j,0)$ can be reached at $t + dt$ if the process is in $(j,-1)$ at time t,this occurring with probability $g_{ij}(-1,t)$, and then a transition to $(j,0)$ occurs (with probability ηdt). But from previous results $\underline{g}(-1,t) = \underline{s}^{-}(t)*\underline{g}(t)$.

Transforming the first equation yields

$$\omega\underline{\gamma}(\omega) - \underline{I} = \underline{\gamma}(\underline{\omega})[\underline{Q} - \underline{\eta}_D - \underline{r}_D] + \underline{\gamma}(\omega)[\underline{\eta}_D\underline{\sigma}^{-}(\omega) + \underline{r}_D\underline{\sigma}^{+}(\omega)]$$

As $\underline{\gamma} = \underline{\gamma}(\omega=0)$, we conclude,

$$\underline{\gamma} = [\underline{\eta}_D + \underline{r}_D - \underline{Q} - \underline{\eta}_D\underline{\sigma}^{-} - \underline{r}_D\underline{\sigma}^{+}]^{-1}$$

The existence of $\underline{\gamma}$, i.e. the invertibility of the matrix in in brackets, can now be shown. Let $\nu_i = \eta + r_i + (N-i)\lambda + i\mu$ $\underline{\nu}_D = \text{diag}(\nu_i)$. Then the matrix in question can be rewritten as $\underline{\nu}_D[\underline{I} - \underline{B}]$ where $\underline{B} = \underline{\nu}_D^{-1}(\hat{\underline{Q}} + \underline{\eta}_D\underline{\sigma}^{-} + \underline{r}_D\underline{\sigma}^{+})$ and $\hat{\underline{Q}} = \underline{Q} +$ diagonal$((N-i)\lambda + i\mu)$. $\underline{B}$ is an irreducible nonnegative matrix, and because $\underline{\sigma}^{+}$ is strictly substochastic, $\underline{B}$ is also. Thus $s_p(B) < 1 \Rightarrow \underline{I} - \underline{B}$ is invertible.

Now $\gamma_{ij} = \int_0^{\infty} g_{ij}(t) = E\left\{\int_0^{\infty} dt I(K^H(t) = 0, A(t) = j \,|\, K^H(0) = 0, A(0) = i)\right\}$, where $I(\ )$ is the indicator function. That is, γ_{ij}

is the expected amount of time that $[A(t), K^H(t)]$ spends in $(j,0)$ given a start in$(i,0)$. With negative drift, $K^H(t) \to -\infty$ a.s., so that the column $K^H = 0$ is transient. The existence of $\underline{\gamma}$ shows that $g_{ij}(t)$ goes to 0 rapidly enough to be integrable. From this, it follows that $\underline{\gamma}(\omega) = \int_0^\infty e^{-\omega t}\underline{g}(t)dt$ is finite as well for Re $\omega \geq 0$.

F. Alternate Characterization of $\underline{e}(0)$

For the bounded process, let $\underline{f}(t)$ be the transition probability matrix for the K = 0 column

$$f_{ij}(t) = \Pr[A(t) = j,\ K(t) = 0 \mid A(0) = i,\ K(0) = 0].$$

If the drift is negative, the column k = 0 is recurrent in the bounded process, and $f_{ij}(t) \to e_{j0}$ as $t \to \infty$, independent of i. $\underline{f}(t)$ satisfies the followed differential equation which can be derived similarly to the one for $\underline{g}(t)$.

$$\dot{\underline{f}}(t) = \underline{f}(t)[\underline{Q} - \underline{\eta}_D] + \underline{f}(t)\underline{\eta}_D * \underline{s}^-(t)$$

Note the $\underline{r}_D$ terms are not present (except through $\underline{s}^-(t)$) because of the boundary. Also, for the bounded process, the one step left (i.e. $k \to k-1$) first passage time density for any $k \geq 1$ is the same as that in the homogeneous process because neither is bounded above. Thus the $\underline{s}^-(t)$ from the homogeneous process is the correct quantity to use in the above equation. It follows that $\underline{f}(\infty) = \underline{1}\,\underline{e}^T(0)$ and $\underline{e}^T(0)$ is a left eigenvector

$$\underline{e}^T(0)[\underline{Q} - \underline{\eta}_D + \underline{\eta}_D\underline{\sigma}^-] = \underline{0}^T.$$

From the discussion on uniformizing a chain, $\underline{e}(0)$ is also a left eigenvector with eigenvalue 1 of the matrix
$(\underline{I} + \frac{1}{\nu}[\underline{Q} - \underline{\eta}_D] + \frac{1}{\nu}\underline{\eta}_D\underline{\sigma}^-) \triangleq (\underline{a}_\nu^o + \underline{a}^+) = \underline{A}_\nu$ where
$\nu \geq \max(\eta + (N-i)\lambda + i\mu)$. Since $\underline{\sigma}^-$ is stochastic, the matrix $\underline{A}_\nu$ is stochastic and is a transition probability matrix for the discrete time chain embedded at clock "bong" instants when the process enters a state in the column K = 0. That is $(\underline{A}_\nu)_{ij}$

is the probability that at the next clock "bong" for which the chain enters a state $(\cdot,0)$, it enters $(j,0)$, given a start in $(i,0)$. Note that this may be the very next clock bong, with probabilities given by the $\underline{a}_{\nu}^{o}$ term. (Again $(a_{\nu}^{o})_{ii} > 0$ if $\nu > \eta + (N-i)\lambda + i\mu$ so $(i,0)$ itself can be the "entrance" state at the next bong.) Or, at the next clock time, the chain may go to $(i,1)$, with probability, $\frac{\eta}{\nu}$, in which case the state of return to$\{(\cdot,0)\}$is determined by the $\underline{g}_{ij}^{-}$'s.

G. Algorithm for Computing $\underline{\sigma}^+$ and $\underline{\sigma}^-$

(Much of this discussion is due to Zachmann [21] and Neuts [12].)

Let S_0 be the set of substochastic matrices ($\underline{A}\,\underline{1} \leq \underline{1}$), S_1 the set of stochastic matrices, ($\underline{A}\,\underline{1} = \underline{1}$), and S_2 the set of strictly substochastic matrices ($\underline{A}\,\underline{1} < \underline{1}$). Note $S_0 \supseteq S_1 \cup S_2$, and the containment is proper. We now study the equation $\underline{\sigma} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^2$ where $\underline{\alpha}, \underline{\beta} \in S_0, \underline{\alpha} > \underline{0}, \underline{\beta} > \underline{0}$, $\underline{\alpha} + \underline{\beta} \in S_1$. (Technically, the assumption $\underline{\alpha} > \underline{0}$, $\underline{\beta} > \underline{0}$ could be replaced by $\underline{\alpha} \geq \underline{0}$, $\underline{\beta} \geq \underline{0}$, and weaker assuptions about the irreducibility/primitivity of $\underline{\alpha}$ or $\underline{\beta}$ or even just $\underline{\alpha} + \underline{\beta}$, depending on the result. For our purposes, such technical details are of little added value, so we assume $\underline{\alpha} > \underline{0}$, $\underline{\beta} > \underline{0}$.)

Let f be the function $f(\underline{\sigma}) = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^2$. Note that if $\underline{\sigma} \in S_i$, then $f(\underline{\sigma}) \in S_i$, $i = 0,1,2$. By Brouwer's Fixed Point Theorem, f has at least one fixed point in S_1 (S_1 is compact, and f is continuous.) We investigate other solutions of probabilistic significance.

Consider the iteration $\underline{\sigma}_{k+1} = f(\underline{\sigma}_k)$. We denote k applications of the iteration by $A^k(\underline{\sigma})$ so $\underline{\sigma}_k = A^k(\underline{\sigma}_0)$. Let $\underline{\pi}$ denote the left eigenvector of the stochastic matrix $\underline{\alpha} + \underline{\beta}$, associated with the PRF root 1, s.t. $\underline{\pi}^T\,\underline{1} = 1$. Recall that $\underline{\alpha} + \underline{\beta}$ has an interpretation as a transition probability matrix of an embedded chain, so $\underline{\pi}$ is its stationary distribution. The convergence properties of the algorithm are as follows:

1) If $\underline{\sigma}_o = \underline{0}$ then the algorithm converges to a limiting matrix $A^\infty(\underline{0}) \in S_o$, and $A^\infty(\underline{0})$ is a fixed point of f.

2) If $\underline{\pi}^T(\underline{\alpha}-\underline{\beta})1 > 0$ then $A^\infty(\underline{0}) \in S_1$ and is the unique fixed point of f in S_o. (Thus it is the one guaranteed by Brouwer's theorem.) Further, for any $\underline{\sigma}_o \in S_o$, $A^\infty(\underline{\sigma}_o) = A^\infty(\underline{0})$.

3) If $\underline{\pi}^T(\underline{\alpha}-\underline{\beta})\underline{1} < 0$, then $A^\infty(\underline{0})$ is the unique fixed point in S_2. (If $\underline{\sigma}_o \neq \underline{0}$, the algorithm need not converge to the desired strictly substochastic solution. For example, if $\underline{\sigma}_o$ is stochastic then $A^k(\underline{\sigma}_o)$ is stochastic $\forall$ k.)

We now sketch the proof of these properties.

1. One can show that $A^{k+1}(\underline{0}) \geq A^k(\underline{0})$. Since $f(\underline{\sigma}) \in S_o$ if $\underline{\sigma}_o \in S_o$, the sequence $A^k(\underline{0})$ is bounded above, componentwise. Thus it must converge componentwise to a limit. This limit is clearly a fixed point.

2. One can also show that for any $\underline{\sigma}_o \in S_o$, $A^k(\underline{\sigma}_o) \geq A^k(\underline{0})$. Thus if the limit $A^\infty(\underline{0})$ is stochastic, $A^k(\underline{\sigma}_o)$ must also converge to $A^\infty(\underline{0})$. In particular, if $\underline{\sigma}_o$ is a fixed point, then $\underline{\sigma}_o = A^\infty(\underline{\sigma}_o) = A^\infty(\underline{0})$, so $A^\infty(\underline{0})$ is the unique fixed point in S_o if it is stochastic. The hypothesis $\underline{\pi}^T(\underline{\alpha}-\underline{\beta})\underline{1} > 0$ is needed to prove $A^\infty(\underline{0})$ stochastic.

The quantity $\delta = \underline{\pi}^T(\underline{\alpha}-\underline{\beta})\underline{1}$ also has an interpretation as a drift. The i^{th} component of $(\underline{\alpha}-\underline{\beta})\underline{1}$ is the difference between

the probabilities of going from $(i,k) \rightarrow \{(\cdot,k+1)$ and going from $(i,k) \rightarrow \{(\cdot,k-1)\}$. The $\underline{\pi}$ appropriately weights the phase state i at which the departure from the column occurs.

The condition $\delta < 0$ is the requirement for the ergodicity of the general, homogeneous QBD process with boundary at 0. For the voice/data model, it reduces to the previous stability condition $\eta < \underline{p}^T \underline{r}$. (This was pointed out to us by Humblet [22].) We now show this. $\underline{\Pi}$ is the ergodic distribution of the discrete time chain embedded at epochs when K^H changes. For the voice/data model, A(t) does not change when K^H changes, so by our remarks in Chapter II, $\underline{\Pi}$ is proportional to $(\underline{\eta}_D + \underline{r}_D)\underline{p}$, i.e. when $A(t) = i$, changes in K^H occur at rate $\eta + r_i$, so Π_i is p_i weighted by this factor. With normalization, $\Pi_i = \dfrac{p_i(\eta+r_i)}{\Sigma p_i(\eta+r_i)}$. The condition $\underline{\Pi}^T(\underline{\alpha}+\underline{\beta}) = \underline{\Pi}^T$ implies

$\underline{\Pi}^T(\underline{I} - (\underline{\eta}_D + \underline{r}_D)^{-1}\ \underline{Q})(\underline{\eta}_D + \underline{r}_D)^{-1}(\underline{\eta}_D + \underline{r}_D) = \underline{\Pi}^T$, which implies

$\underline{\Pi}^T(\underline{I} - (\underline{\eta}_D + \underline{r}_D)^{-1}\ \underline{Q}) = \underline{\Pi}^T$. (Note this equation implies $\underline{\Pi}^T(\underline{\eta}_D + \underline{r}_D)^{-1}\underline{Q} = \underline{0}^T$, which implies $\underline{\Pi}^T$ is proportional to $\underline{p}^T(\underline{\eta}_D + \underline{r}_D)$ as reasoned earlier.) Thus, $\underline{\Pi}^T(\underline{\alpha}-\underline{\beta})\underline{1} = \underline{\Pi}^T(\underline{\eta}_D+\underline{r}_D)^{-1}(\underline{\eta}_D-\underline{r}_D)\underline{1} = \dfrac{\Sigma p_i(\eta-r_i)}{\Sigma p_i(\eta+r_i)} = \dfrac{\eta - \underline{p}^T\underline{r}}{\eta + \underline{p}^T\underline{r}}$. Thus $\delta < 0$ iff $\eta < \underline{p}^T\underline{r}$.

Mathematically, δ enters the convergence proof as follows.

Let $g(x) \triangleq sp[\underline{\alpha}/x + x\underline{\beta}]$, $x > 0$. Since $\underline{\alpha} + \underline{\beta}$ is stochastic, $g(1) = 1$. $g(x)$ is strictly convex, and $g'(1) = -\delta$. The assertion $g'(1) = -\delta$ can be verified by direct computation assuming that $g(x)$ and the left PRF eigenvector $\underline{L}(x)$ (which we take as a row vector) are differentiable. Let $\underline{M}(x) = \underline{\alpha}/x + x\underline{\beta}$. Note $\underline{\alpha} > \underline{0}$ (or $\underline{\beta} > \underline{0}$) implies $\underline{M}(x)$ primitive $\forall\ x > 0$. Thus, the eigenvector $\underline{L}(x)$ can be chosen positive and normalized. We assume that, with this consistent choice, $\underline{L}(x)$ is differentiable. Then $\underline{L}(x)\ \underline{M}(x) = g(x)\ \underline{L}(x)$ implies $\underline{L}'(x)\ \underline{M}(x) + \underline{L}(x)[\underline{\beta} - \underline{\alpha}/x^2] = g'(x)\ \underline{L}(x) + g(x)\ \underline{L}'(x)$. Note $\underline{M}(1) = \underline{\alpha} + \underline{\beta}$ implies $\underline{L}(1) = \underline{\pi}^T$ and $\underline{M}(1)\underline{1} = \underline{1}$. Substituting $x = 1$ into the equation and taking the inner product of both sides with $\underline{1}$ shows $\underline{L}'(1)\underline{1} + \underline{\pi}^T(\underline{\beta} - \underline{\alpha})\underline{1} = g'(1) + \underline{L}'(1) \cdot \underline{1}$ which implies $g'(1) = -\delta$. See Kingman [23] for the convexity proof.

Now let $\underline{\sigma} \geq \underline{0}$ be any fixed point (not necessarily in S_o) and let $r = sp(\underline{\sigma})$. Note that $\underline{\sigma} \geq \underline{\alpha} > \underline{0}$ so $\underline{\sigma}$ is in fact irreducible. By the discussion in Chapter II, the matrix $\frac{\underline{D}^{-1}\underline{\sigma}\ \underline{D}}{r}$ is stochastic, where $\underline{D}$ is the diagonal matrix obtained from the PRF right eigenvector of $\underline{\sigma}$. Thus the identity

$$\frac{\underline{D}^{-1}\ \underline{\sigma}\ \underline{D}}{r} = \frac{\underline{D}^{-1}\underline{\alpha}\ \underline{D}}{r} + r\ \underline{D}^{-1}\underline{\beta}\ \underline{D}\left(\frac{\underline{D}^{-1}\underline{\sigma}\ \underline{D}}{r}\right)^2$$

shows that $\underline{\alpha}/r + r\underline{\beta}$ is similar to a stochastic matrix. This implies $sp(\underline{\alpha}/r + r\underline{\beta}) = 1$. Thus we have a characterization of the PRF root of a nonnegative fixed point. If $\delta > 0$ then $g'(1) < 0$. Together with convexity this shows that $r = 1$

is the only value of $r \leq 1$ s.t. $g(r) = 1$. Thus if $\underline{\sigma} \in S_0$ its spectral radius must be 1. (For any $\underline{M} \in S_0$, $sp(\underline{M}) \leq 1$.) From the discussion on PRF roots (specifically, the row sum bounds on r), the only <u>irreducible</u> substochastic matrices that have spectral radius one are in fact stochastic. Thus if $\underline{\sigma} \in S_0$ and $\underline{\sigma}$ is a fixed point, $\underline{\sigma} \in S_1$.

3) If $\delta < 0$ then $g'(1) > 0$. Since $\underline{\alpha} > \underline{0}$, $g(r) \to \infty$ as $r \to 0$. Together with convexity, this shows that there is exactly one value of $r < 1$ s.t. $g(r) = 1$. This allows for the possibility (though by itself doesn't prove the existence of) another fixed point in S_0, with spectral radius < 1. The matrix $A^{\infty}(\underline{0})$ can be shown to be such a fixed point, and $A^{\infty}(\underline{0})$ is in fact strictly substochastic and is the unique fixed point in S_2.

Let us apply this to the voice/data link. For the queue to be stable, δ must be negative. In this case, the iteration $\underline{\alpha} + \underline{\beta}\,\underline{\sigma}_k^2$, $\underline{\sigma}_0 = \underline{0}$ converges to $\underline{\sigma}^+$ and is strictly substochastic. The other fixed point (by Brouwer's Theorem) has no apparent probabilistic meaning. For the iteration $\underline{\beta} + \underline{\alpha}\,\underline{\sigma}_k^2$, the quantity $\underline{\pi}^T(\underline{\beta}-\underline{\alpha})\underline{1}$ is positive, so this converges to a stochastic solution which is $\underline{\sigma}^-$. As mentioned, the starting point in this case is not important, and $\underline{\sigma}^-$ is unique in S_0.

Extending the probabilistic reasoning used to derive the fixed point equation, we see that $\underline{\sigma}^+$ is the sum over all paths that are "loops followed by an alpha". Specifically, $\underline{\sigma}^+$ is

a sum over terms of the form

$$[\underline{\beta}^{i_1}\,\underline{\alpha}^{j_1} \ldots \underline{\beta}^{i_L}\,\underline{\alpha}^{j_L}]\underline{\alpha}$$

where $\sum_{\ell=1}^{L} i_\ell = \sum_{\ell=1}^{L} j_\ell$ and for any $M < L$, $\sum_{\ell=1}^{M} i_\ell < \sum_{\ell=1}^{M} j_\ell$. Such a term corresponds to a sample path in which the first i_1 changes in the K^H coordinate are negative, the next j_1 positive, etc. until the final j_L changes bring it back to 0, and then a transition to +1 occurs. With each step, the algorithm computes more of these terms. For example

$$\underline{\sigma}_0 = \underline{0}$$

$$\underline{\sigma}_1 = \underline{\alpha}$$

$$\underline{\sigma}_2 = \underline{\alpha} + \underline{\beta}\,\underline{\alpha}^2$$

$$\underline{\sigma}_3 = \underline{\alpha} + \underline{\beta}\,\underline{\alpha}^2 + \underline{\beta}\,\underline{\alpha}\,\underline{\beta}\,\underline{\alpha}^2 + \underline{\beta}^2\,\underline{\alpha}^3 + \underline{\beta}^2\,\underline{\alpha}^2\,\underline{\beta}\,\underline{\alpha}^2.$$

It seems difficult to find any characterization of the probability mass which is still uncounted after the kth iteration -- i.e. what is the convergence rate. This is a real problem. For $\underline{\sigma}^-$, we know that the final answer is stochastic. Thus one has an absolute test of convergence. For $\underline{\sigma}^+$ we have no such test yet. One can look at the successive componentwise differences, but it seems difficult to relate this "stepsize" to the true distance from the limit. This problem needs more investigation.

Finally, we discuss another algorithm. In the equation $\underline{\sigma} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^2$ one can "solve" for $\underline{\sigma}$, $\underline{\sigma} = [\underline{I} - \underline{\beta}\,\underline{\sigma}]^{-1}\,\underline{\alpha}$, and this leads to another iteration in an obvious manner. To see

the difference, consider the result of the second step, when $\underline{\sigma}_0 = \underline{0}$. Then $\underline{\sigma}_2 = [\underline{I} - \underline{\beta}\ \underline{\alpha}]^{-1}\ \underline{\alpha} = \sum_{\ell=0}^{\infty} (\underline{\beta}\ \underline{\alpha})^{\ell}\ \underline{\alpha}$. In this case the algorighm is counting all prefixes of the form $\underline{\beta}\ \underline{\alpha}\ \ldots\ \underline{\beta}\ \underline{\alpha}$ followed by an $\underline{\alpha}$. It is unclear as to when one algorithm "counts faster" than the other. One might even be able to devise a hybrid procedure.

H. An Example

To make some of this more "concrete", we work out the details of the case in which $r_i = \bar{r}$ V i. This problem is rather trivial in that A(t) and K(t) are now statistically independent, and K(t) is just the classical M/M/1 birth-death queuing process with parameters η, $\bar{r}$ and utilization $\rho = \eta/\bar{r} < 1$. For this queue (see Kleinrock [24]), the ergodic probability of k customers in the system is $(1-\rho)\rho^k$, k = 0,1,... Thus, for the vector process, $e_{ik} = p_i (1-\rho)\rho^k$, or $\underline{e}(k) = \underline{p}(1-\rho)\rho^k$. From the previous discussion, we know that $\underline{e}^T(k) = \underline{p}^T[\underline{I} - \underline{\theta}]\theta^k$, and one might suspect that $\underline{\theta}$ is simply $\rho\underline{I}$. This is not the case. In fact, we have shown that $\underline{\sigma}^+$ is primitive, and hence ρ (which will turn out to be the PRF root) is a simple eigenvalue. Thus $\underline{\theta}$, which is similar to $\underline{\sigma}^+$, cannot be $\rho\underline{I}$.

First we recall the following from Chapter III.

- The $\underline{Q}$ matrix for the speaker process has eigenvalues $S_i = -i(\lambda+\mu)$ i = 0,...N.
- $\underline{Q}$ is diagonalizable, i.e. $\underline{Q} = \underline{L}^{-1}\,\underline{S}\,\underline{L}$, where the i^{th} row of $\underline{L}$ is $\dfrac{[1\ \varepsilon]^{*N-i} * [1\ -1]^{*i}}{(1+\varepsilon)^{N/2}}$, $\varepsilon = \lambda/\mu$, and $\underline{L} = \underline{L}^{-1}$. (* = convolution)
- $\underline{p}$ is proportional to $\underline{L}(0)$, specifically, $\underline{p} = \dfrac{[1\ \varepsilon]^{*N}}{(1+\varepsilon)^N}$ or $\underline{p} = \underline{L}(0) \cdot (1+\varepsilon)^{-N/2}$

From the previous sections in this chapter,

$$\underline{\alpha} = ((\eta + \bar{r})\underline{I} - \underline{Q})^{-1}\,\eta \qquad \underline{\sigma}^{+} = \underline{\alpha} + \underline{\beta}(\underline{\sigma}^{+})^{2}$$

$$\underline{\beta} = ((\eta + \bar{r})\underline{I} - \underline{Q})^{-1}\,\bar{r} \qquad \underline{\sigma}^{-} = \underline{\beta} + \underline{\alpha}(\underline{\sigma}^{-})^{2}$$

$$\underline{\gamma} = ((\eta + \bar{r})\underline{I} - \underline{Q} - \bar{r}\,\underline{\sigma}^{+} - \eta\underline{\sigma}^{-})^{-1} \qquad \underline{\theta} = \underline{\gamma}^{-1}\,\underline{\sigma}^{+}\,\underline{\gamma}$$

where now $\underline{r}_D$ reduces to $\bar{r}\underline{I}$.

Now define $\hat{\underline{M}} = \underline{L}\,\underline{M}\,\underline{L}$ where $\underline{M}$ is any of the previous matrices. Set $\underline{D} = [(\eta + \bar{r})\underline{I} - \underline{S}]^{-1}$. $\underline{S}$ = diagonal (S_i) so $\underline{D}$ is diagonal. Then a simple computation shows $\hat{\underline{\alpha}} = \underline{D}\eta$, $\hat{\underline{\beta}} = \underline{D}\,\bar{r}$, $\hat{\underline{\gamma}} = (\underline{D}^{-1} - \bar{r}\,\hat{\underline{\sigma}}^{+} - \eta\,\hat{\underline{\sigma}}^{-})$, $\hat{\underline{\theta}} = \hat{\underline{\gamma}}^{-1}\,\hat{\underline{\sigma}}\,\hat{\underline{\gamma}}$. We will show that $\underline{\sigma}^{+}$ and $\underline{\sigma}^{-}$ are diagonal, which implies $\hat{\underline{\gamma}}$ diagonal, and thus $\hat{\underline{\theta}} = \hat{\underline{\sigma}}^{+}$, which implies $\underline{\theta} = \underline{\sigma}^{+}$. $\hat{\underline{\sigma}}^{+}$ wil have real positive diagonal terms, and ρ will be the largest with $\underline{\sigma}^{+}_{oo} = \rho$. Thus

$$\underline{e}^{T}(k) = \underline{p}^{T}[\underline{I} - \underline{\theta}]\underline{\theta}^{k} = \underline{p}^{T}\,\underline{L}\,[\underline{I} - \hat{\underline{\sigma}}^{+}]\underline{L}\,\underline{L}(\hat{\underline{\sigma}}^{+})^{k}\,\underline{L}$$

$$= \underline{p}^{T}\,\underline{L}(\underline{I} - \hat{\underline{\sigma}}^{+})(\hat{\underline{\sigma}}^{+})^{k}\,\underline{L} = (1+\varepsilon)^{-N/2}\,\underline{f}_{o}^{T}\,(\underline{I} - \hat{\underline{\sigma}}^{+})(\hat{\underline{\sigma}}^{+})^{k}\,\underline{L}$$

where $\underline{f}_o$ is the vector $(1,0,\ldots 0)$. (This follows because $\underline{p} = \underline{L}(0)(1+\varepsilon)^{-N/2}$ and $\underline{L} = \underline{L}^{-1}$, so $\underline{L}^{T}(0)\,\underline{L} = \underline{f}_{o}^{T}$.) Continuing, we obtain $\underline{e}^{T}(k) = (1+\varepsilon)^{-N/2}(1-\rho)\rho^{k}\,\underline{f}_{o}^{T}\,\underline{L}$ = $(1+\varepsilon)^{-N/2}\,(1-\rho)\rho^{k}\,\underline{L}^{T}(0) = \underline{p}^{T}(1-\rho)\rho^{k}$, as claimed.

Now consider the matrix quadratic equation $\underline{\sigma} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^{2}$. (We drop the "+" for convenience.)

We know that the equation has a unique strictly sub-stochastic solution, which we want, and at least one other solution in S_1. The meaningful one is the limit of the recursion $\underline{\sigma}_{\ell} = \underline{\alpha} + \underline{\beta}\,\underline{\sigma}^{2}_{\ell-1}$, $\underline{\sigma}_{0} = \underline{0}$. Conjugating the equation

by $\underline{L}$, we obtain $\hat{\underline{\sigma}} = \hat{\underline{\alpha}} + \hat{\underline{\beta}}\,\hat{\underline{\sigma}}^2$, where $\hat{\underline{\alpha}} = \underline{D}\eta$, $\hat{\underline{\beta}} = \underline{D}\,\bar{r}$.

Notice that if we start the transformed recursion with $\hat{\underline{\sigma}}_0 = \underline{0}$, then $\hat{\underline{\sigma}}_k$ is diagonal for each k. For the moment, let us assume that the fixed point $\hat{\underline{\sigma}}$ which corresponds to the desired $\hat{\underline{\sigma}}^+$, i.e. the $\hat{\underline{\sigma}}$ s.t. $\underline{L}\,\hat{\underline{\sigma}}\,\underline{L}$ is the correct $\underline{\sigma}^+$, is diagonal. Then, we are left with the individual equations

$$\hat{\sigma}_{ii} = \eta D_{ii} + \bar{r}\, D_{ii}(\hat{\sigma}_{ii})^2 , \quad i = 0,1,\ldots,N.$$

These equations are scalar quadratic equations of the form

$$z = a + b\, z^2$$

where a, b > 0 and a < b (since $\eta < \bar{r}$). Further a + b = 1 if i = 0, a + b < 1, i ≠ 0. From the quadratic formula, we obtain the roots

$$\frac{1 \pm \sqrt{1 - 4ab}}{2b}$$

Note $ab < \frac{1}{4}$ implies the roots are real. A simple geometric argument shows that both roots are positive, one root is always less than 1, and the other equals 1 if a + b = 1, and is greater than 1 if a + b < 1. Since we are after a strictly sub-stochastic matrix, we pick the root less than 1 for each i. This is

$$\frac{1 - (1 - 4\eta\,\bar{r}\,D_i)^{1/2}}{2\bar{r}} .$$

For $i = 0$, this is ρ , and for $i \neq 0$ it is less than ρ.

Technically, we still need to show that with this choice of $\hat{\underline{\sigma}}$, the matrix $\underline{L}\,\hat{\underline{\sigma}}\,\underline{L}$ is the desired $\underline{\sigma}^{+}$. Intuitively, this is pretty clear. A formal proof can be obtained by showing that if $\hat{\underline{\sigma}}$ is the limit of $\hat{\underline{\sigma}}_{\ell} = \hat{\underline{\alpha}} + \hat{\underline{\beta}}\,\hat{\underline{\sigma}}^{2}_{\ell-1}$, $\hat{\underline{\sigma}}_{0} = \underline{0}$, then $\underline{L}\,\hat{\underline{\sigma}}\,\underline{L}$ is the limit of the original recursion for $\underline{\sigma}^{+}$, starting with $\underline{\sigma}_{0} = \underline{0}$. A simple argument then shows that the $\hat{\underline{\sigma}}$ we have chosen, i.e. with eigenvalues <1, is the limit of $\hat{\underline{\sigma}}_{\ell} = \hat{\underline{\alpha}} + \hat{\underline{\beta}}\,\hat{\underline{\sigma}}^{2}_{\ell-1}$ starting with $\hat{\underline{\sigma}}_{0} = 0$.

Similar computations go through for $\underline{\sigma}^{-}$ except now we have the scalar equations $z = a + bz^{2}$ with $a > b$. For $i = 0$ (where $a + b = 1$) we obtain a root at $z = 1$ and a root $z > 1$. Thus we pick $z = 1$. For $i \neq 0$, one root is < 1, and the other is > 1. We pick the smaller.

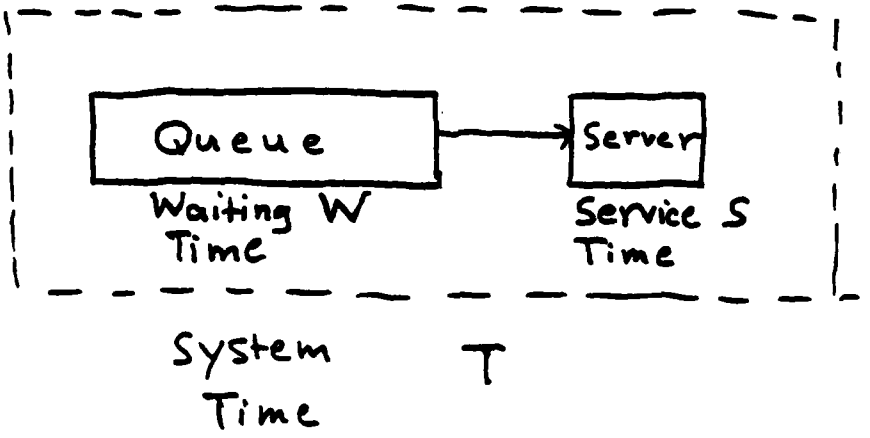
The eigenvalues of $\underline{\sigma}^{+}$ other than ρ have no apparent probabilistic meaning.

VI. Queue Statistics and Numerical Examples

A. Queue Statistics

The matrix series $\sum_{\ell} \underline{\theta}^{\ell}$ obeys the "ordinary calculus rules" for geometric series, e.g. $\sum_{\ell=0}^{\infty} \ell\underline{\theta}^{\ell} = \underline{\theta}[\underline{I} - \underline{\theta}]^{-2}$. Hence, computing moments of the distribution of the number in system poses no special problem. Also, one can easily obtain an expression for the actual "bit" backlog distribution. Specifically, let $f_k = \underline{e}^T(k)\,\underline{1}$ denote the marginal probability of k messages in the system, and let $[\exp(\xi)]^{*k}$ denote the k-fold covolution of the message length distribution exp (ξ) with itself. Then the backlog distribution is $\sum_{k=0}^{\infty} f_k[\exp(\xi)]^{*k}$

It is customary to separate a queuing system into the "queue" and the "service facility"



The total time spent in the system is denoted by T, the time a message spends in the queue is called the waiting time W, and S is the service time. It follows from this definition

that

$$\overline{KQ} = \overline{KS} - 1 + Pr[KS = 0]$$

where KS = number in system, KQ = number in queue, and over-bar denotes expectation.

For a constant rate server (and i.i.d. message lengths) this division is quite natural. A message's service time is linearly related to its length and is independent of its waiting time and the service and waiting times of other messages. That is, the queuing process only affects waiting times. For a voice/data queue model in which r_i varies with i, the distinction between service and waiting can be less meaningful as the following examples indicate.

Consider a service rate vector $(r_0,\ldots,r_N)$ s.t. $r_i = 0$ $\forall$ $i \neq 0$, and r_o is "infinite". Assume $\lambda = \mu = 1$. The mean first passage from A = 1 to A = 0 is $\frac{1-p_o}{p_1} \simeq \frac{2^N}{N}$. More generally, the passage time from any initial state to state 0 is dominated by the time to take the last step from 1 to 0, i.e. if a message arrives when $A \neq 0$, the mean time until the speaker process reaches 0 is still $\simeq \frac{2^N}{N}$. Now suppose that the mean message length $\xi^{-1} = 1$. There are two cases. If the mean interarrival time $\frac{1}{\eta} << \frac{2^N}{N}$, large queues accumulate while $A \neq 0$, and these are emptied "instantaneously" when A reaches 0. Thus most messages have a very large waiting time but essentially no serice time. If $\frac{1}{\eta} >> \frac{2^N}{N}$, then most messages arrive to an empty system. Their waiting times are zero, but their

service times are very large. In both cases, the mean system time is determined by the time it takes for a return to A = 0.

The point of this is as follows. A message is "inconvenienced" by having to share the link with both the speakers and other messages. To a certain extent, the "service time" reflects the sharing with the speakers, and the waiting time reflects the sharing with other messages. Although it is possible to compute the service time distribution given a start in some speaker state, the interaction between the queueing process and voice activity process can affect the distribution of service initiation states. That is, waiting and service times are not independent, and the distribution of service initiation states is generally not the same as the speaker ergodic distribution $\underline{p}$.

In any case, it is possible to derive expressions for the Laplace-Stieltjes transforms of the limiting distributions of T, W, and S. Let $\hat{X}_{ij}(z)$ denote the L - S transform of the joint [service time; voice completion state = j/A(0) = i] distribution, i.e.

$$\hat{X}_{ij}(z) = \int_0^\infty e^{-zt} \, d_t \Pr[\text{service time} \leq t, A = j \text{ at completion} | A(0) = i]$$

(See Chapter IV.)

Because message lengths are i.i.d., it follows that the analogous transform for the time to service k successive messages is $[\hat{\underline{X}}(z)]^k$. Further, $[\hat{\underline{X}}(0)]^k_{ij} = \Pr[\text{completion of k messages occurs when } A = j | A(0) = i]$. We also make use of

the following observations:

- Poisson arrivals take a "random" look at the system, so that the equilibrium distribution of [A, K] as seen by arriving messages is the same as the ergodic distribution $\{e_{ik}\}$. (This need not be the case for non-Poisson arrivals; see Kleinrock [24] .)
- If a message arrives to a nonempty system and A = j, the distribution of the time needed to complete service for the message in the service facility is the same as if a message with length ~ exp(ξ) starts service when A = j. (This follows from the memoryless property of the exponential distribution.)

From all this, it follows that

$$\hat{T}(z) = \sum_{k=0}^{\infty} \underline{e}^T(k)[\hat{\underline{X}}(z)]^{k+1}\,\underline{1} = \underline{p}^T[\underline{I} - \underline{\theta}][\underline{I} - \underline{\theta}\,\underline{X}(z)]^{-1}\,\hat{\underline{X}}(z)\,\underline{1}$$

$$\hat{W}(z) = \sum_{k=0}^{\infty} \underline{e}^T(k)\,[\hat{\underline{X}}(z)]^k\,\underline{1}$$

$$\hat{S}(z) = \sum_{k=0}^{\infty} \underline{e}^T(k)\,[\hat{\underline{X}}(0)]^k\,\hat{\underline{X}}(z)\,\underline{1}$$

The matrix $\hat{\underline{X}}(0)$ can be related to $\underline{\alpha}$ and $\underline{\beta}$, as follows. If $K(0) \geq 1$, the time to complete a service is the time until K(t) registers its first decrease. By this we do not mean the first passage time to K(0) - 1, but the time until the first

decrease occurs, with any possible number of intervening increases. Since K is not bounded above, it follows that

$$\underline{X} = \underline{\beta} + \underline{\alpha}\,\underline{X}$$

(The reasoning is similar to that used to derive the quadratic equation for the $\underline{\sigma}$'s.) Thus $\underline{X} = [\underline{I} - \underline{\alpha}]^{-1}\,\underline{\beta}$. A straightforward calculation will show that $[\underline{I} - \underline{\alpha}]^{-1}\,\underline{\beta} = (\xi\,\underline{r}_D - \underline{Q})^{-1}\,\underline{r}_D$, which is the expression derived in Chapter IV.

B. Numerical Examples

We have applied the matrix-geometric algorithms to a voice/data queue with 10 speakers. The other speaker parameters are $\lambda = .75$, $\mu = .81$ (per second) so mean silence $= \lambda^{-1} = 1.34$sec. and mean talkspurt $= \mu^{-1} = 1.23$ sec. [6]. The link has capacity 320 Kbps, and each speaker demands 32 Kbps when in talkspurt. Thus $r_i = 320 - 32i$ Kbps, so $\bar{r} \approx 166$ Kbps.

One important question is how the mix of data traffic affects performance, i.e. for fixed total average data rate, η/ξ, how does performance depend on η and ξ individually. We consider three mixes for each value of the utilization $\rho = .1,.2,\ldots,.9$, where $\rho = \eta/(\xi\bar{r})$. The three mixes are referred to as cases A, B, and C, and the respective mean message lengths are 500, 1000, and 2000 bits. Thus, in each case, η is varied to obtain the appropriate value of ρ.

Notice that for states 5 - 10, $r_i \leq \bar{r}$. The total occupation fraction of these states is about .57. We have

also computed the mean first passage times from each speaker state to state $\lceil A \rceil = \lceil 4.8 \rceil = 5$. These are

State	Time(sec)	State	Time (sec)
0	1.23	6	.36
1	1.1	7	.61
2	.93	8	.80
3	.72	9	.95
4	.43	10	1.08

The results are organized in tabular form, and the notation is as follows:

$\overline{KS}$ = mean number in system

VARKS = variance of KS

STDVKS = $\sqrt{VARKS}$

COEFVAR = STDVKS/$\overline{KS}$

$\overline{KQ}$ = mean number in queue

$\overline{T}$ = mean system time

$\overline{W}$ = mean waiting time .

$\overline{SERV}$ = mean service time

Pr[0] = Pr[KS = 0]

- DRIFT = $-\underline{\pi}^T(\underline{\alpha} - \underline{\beta})\underline{1} = \frac{\xi\overline{r} - \eta}{\xi\overline{r} + \eta}$ (see Chapt.V)

MAXDEC = maximum asymptotic decay rate, i.e. PRF root of $\underline{\sigma}^+$.

Our "benchmark" for comparison is an M/M/1 queue with service rate $\overline{r} = \Sigma\rho_i\, r_i$, i.e. $r_i = \overline{r}$ V i. This is appropriate since we wish to determine how the service rate variability affects performance. Each of the first nine tables represents the above statistics for all cases (one table for each value of ρ). Each of the next two tables presents selected

statistics for all values of ρ and all mixes. In the second of these two tables, the quantity $\overline{SERV}$ (random) refers to $\Sigma p_i \overline{SERV_i}$ where $\overline{SERV_i}$ is the mean effective service time given a start in state i, and p_i is the ergodic probability that A = i. This is the average service time that a message arriving "randomly" to an empty system would incur. As indicated, this is not necessarily the actual mean service time incurred by messages since the queuing process does interact with the voice activity process. For purposes of comparison, we mention that $\overline{SERV}$(random) is 3.5 ms, 6.3 ms and 13.6 ms in cases A, B, and C respectively. Finally, the last three tables present some points from the tails of the respective distributions, e.g. TAIL 5 = Pr[KS $\geq$ 5].

In the case of $\underline{\sigma}^-$, the iteration for $\underline{\sigma}_\ell = \underline{\beta} + \underline{\alpha}\,\underline{\sigma}^2_{\ell-1}$ was run until $\underline{\sigma}_\ell$ was stochastic to within 10^{-4} or so. In the case of $\underline{\sigma}^+$ (for which we have no absolute test), the procedure was run until $\max_{i,j} |(\underline{\sigma}_\ell)_{ij} - (\underline{\sigma}_{\ell-1})_{ij}| \leq 10^{-7}, 10^{-8}$. We did not perform an error propagation analysis, but this stopping criterion seemed reasonably adequate. Occasionally, we run some cases until the maximum difference was 10^{-12} or so. This did not result in any drastic changes.

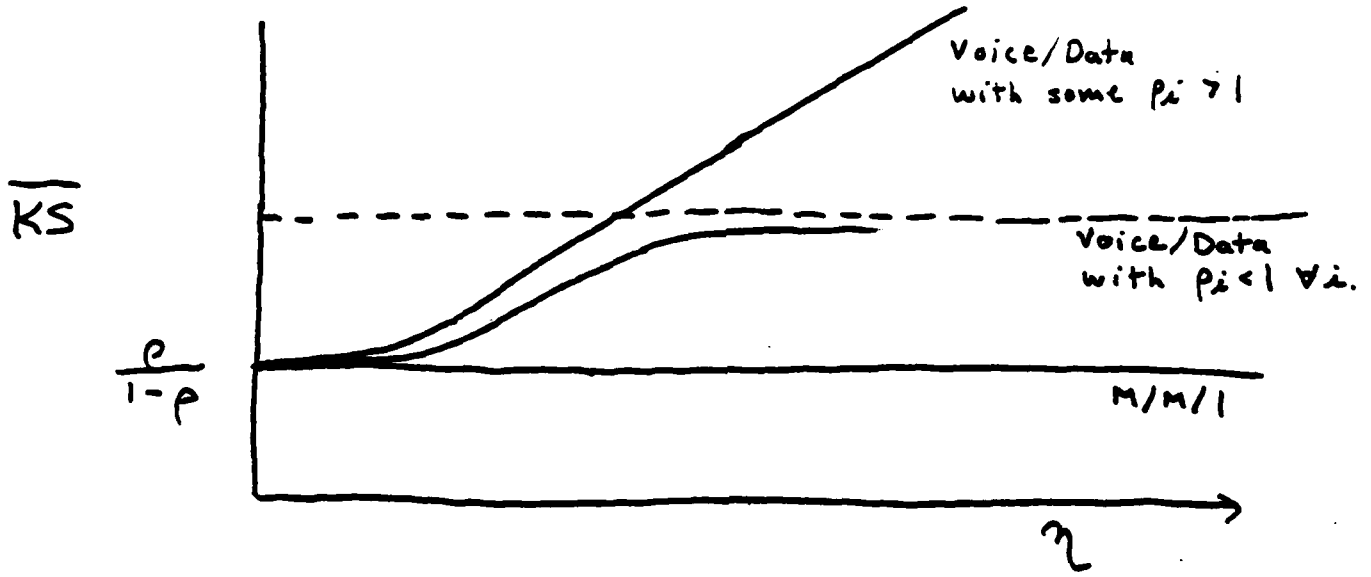
For an M/M/1 queue, the PRF root is ρ and $f_k = (1-\rho)\rho^k$, i.e. the distribution of KS is independent of the load mix. Thus for fixed ρ, the average system time approaches zero as $0(\frac{1}{\eta})$ as $\eta, \xi \to \infty$. For the voice/data queue, the distribution

of KS is not independent of the load mix, and we can qualitatively explain the observed behavior as follows. (In this discussion we consider ρ fixed, i.e. we let η, ξ approach either 0 or ∞ with η/ξ remaining fixed.) As indicated in Chapter IV, the behavior of the voice/data queue approaches that of the analogous M/M/1 queue as $\eta, \xi \to 0$. This is because the speaker process moves "infinitely fast" relative to the data arrival and service processes so that each message is effectively served at a constant rate $\bar{r}$. As $\eta, \xi \to \infty$, the individual queuing processes in each speaker state become "decoupled", and there are two possible types of behavior. If $\rho_i = \eta/(\xi r_i) < 1 \; \forall \; i$, then the behavior approaches that of $N+1$ decoupled, stable, M/M/1 queues with utilizations ρ_i. For example, $\overline{KS}$ will approach $\Sigma p_i \frac{\rho_i}{1 - \rho_i}$. Note, by convexity,

$$\Sigma p_i \frac{\rho_i}{1 - \rho_i} \geq \frac{\Sigma p_i \rho_i}{1 - \Sigma p_i \rho_i} \geq \frac{\eta(\Sigma \xi p_i r_i)^{-1}}{1 - \eta(\Sigma \xi p_i r_i)^{-1}} = \frac{\rho}{1 - \rho}$$

That is, the variable service rate does degrade performance, though $\overline{KS}$ remains bounded as $\eta \to \infty$. If there is at least one overloaded or unstable state, i.e. some i s.t. $\rho_i > 1$, the behavior is qualitatively different. While the voice process is sitting in state i, the mean number in system is growing as $(\eta - \xi r_i)t$. As long as overall stability, i.e. $\eta < \xi \bar{r}$, is present, these backlogs are eventually emptied. However, as $\eta, \xi \to \infty$, the contribution to $\overline{KS}$ made by the unstable states becomes the dominant contribution, and, asymptotically,

$\overline{KS}$ will grow linearly with η. The value of the slope of this growth depends on many factors, and there does not appear to be a simple estimate. Thus, we have the following qualitative picture of the behavior of $\overline{KS}$.



(No assertions about monotonicity or convexity are intended in these graphs, though one would certainly expect at least monotonicity.)

For the M/M/1 and voice/data with $\rho_i < 1$, $\forall$ i, $\overline{KS}$ remains bounded, and it follows that as $\eta, \xi \to \infty$, the average system time goes to 0. For the third case, the average number in system grows linearly with η , but since the average message length is scaled accordingly, it follows that the average backlog in bits and hence the system time, remain bounded. In fact, if $\overline{KS}$ grows as $b\eta$ it follows from Little's Theorem that $\overline{T} \to b$. Thus, the existence of an unstable state does not

preclude overall stability, but it does lead to the somewhat "unconventional" result that even if individual messages become small, their delay is bounded away from zero, for fixed total data rate.

As $\eta, \xi \to \infty$, the arriving data stream effectively becomes a steady flow of rate η/ξ bps. This is because during an interval of length t, the mean number of arriving bits is $\eta t/\xi$ but the variance is $2\eta t/\xi^2$, and this goes to zero as $\eta, \xi \to \infty$, η/ξ fixed. This suggests using a "flow model" for the backlog when η and ξ are large. That is, one assumes that the backlog $X(t)$ grows (or shrinks) deterministically at rate $\eta/\xi - r_i$ when $A = i$. This model should give the correct asymptotic dependence of the backlog on η, i.e. if $\overline{KS} \to b\eta$, the flow model will show that the average backlog $\overline{X}$ is $(\eta/\xi)b$. For the flow model, the vector process $[A(t), X(t)]$ is a Markov process, and its ergodic probability distribution vector $\underline{G}(x)$, where $G_i(x) = \Pr[A = i, X \leq x]$, satisfies the system of differential equations

$$\frac{d}{dx}\,\underline{G}^T(x)\left((\eta/\xi)\underline{I} - \underline{r}_D\right) = \underline{G}^T(x)\,\underline{Q} \tag{1}$$

subject to the boundary conditions

- $G_i(0) = 0$ if $\eta/\xi > r_i$, i.e. the backlog cannot be empty in unstable states,
- $G_i(\infty) = p_i$, i.e. the marginal distribution of A is the original ergodic distribution $\underline{p}$.

(This model was used by Berger in [6], and the reader

can find a detailed exposition there.)

It is sometimes easier to work with these equations directly rather than with limiting arguments on the previous expression $\sum_{k=0}^{\infty} f_k[\exp(\xi)]^{*k}$. For example, it follows from the structure of the differential equation and the fact that the boundary conditions depend only on λ/μ, that the backlog random variable is linear in $1/\lambda$, $1/\mu$ for λ/μ fixed, i.e. $\underline{G}(x,s\lambda, s\mu) = \underline{G}(sx,\lambda,\mu)$. This implies that the average system time is linear in $1/\lambda$, $1/\mu$, in the limit of small, frequent messages.

	A		B		C		
	Voice/Data	M/M/1	Voice/Data	M/M/1	Voice/Data	M/M/1	
$\overline{KS}$	.138	.11	.135	.11	.132	.11	
VARKS	.226	.123	.181	.123	.164	.123	
STDVKS	.475	.351	.426	.351	.406	.351	
COEFVAR	3.4	3.2	3.2	3.2	3.1	3.2	
$\overline{KQ}$	.024	.011	.021	.011	.019	.011	$\rho = .1$
$\overline{T}$ (ms)	4.1	3.3	8.1	6.6	15.9	13.2	
$\overline{W}$ (ms)	.73	.3	1.3	.6	2.3	1.2	
$\overline{SERV}$ (ms)	3.4	3.0	6.8	6.0	13.6	12.0	
Pr[0]	.886	.9	.886	.9	.887	.9	
-DRIFT	.82		.82		.82		
MAX DEC	.81		.69		.53		

	A		B		C	
$\overline{KS}$	.348	.25	.33	.25	.32	.25
VARKS	1.16	.31	.71	.31	.54	.31
STDVKS	1.07	.56	.84	.56	.74	.56
COEFVAR	3.1	2.24	2.25	2.24	2.3	2.24
$\overline{KQ}$	.122	.05	.104	.05	.092	.05
$\overline{T}$ (ms)	5.2	3.7	9.9	7.4	19	14.8
$\overline{W}$ (ms)	1.8	.7	3.2	1.4	5.6	2.8
$\overline{SERV}$ (ms)	3.4	3.0	6.8	6.0	13.5	12.0
Pr[0]	.774	.8	.774	.8	.775	.8
-DRIFT	.67		.67		.67	
MAXDEC	.90		.83		.71	

$\rho = .2$

	A		B		C	
$\overline{KS}$	.7	.43	.64	.43	.59	.43
VARKS	5.5	.61	2.48	.61	1.6	.61
STDVKS	2.3	.78	1.6	.78	1.2	.78
COEFVAR	3.3	1.8	2.5	1.8	2.03	1.8
$\overline{KQ}$	.39	.13	.30	.13	.26	.13
$\overline{T}$ (ms)	7.25	4.3	13	8.6	24	17.2
$\overline{W}$ (ms)	3.9	1.3	6.1	2.6	10	5.2
$\overline{SERV}$ (ms)	3.4	3.0	6.8	6.0	13	12
Pr[0]	.66	.7	.66	.7	.66	.7
-DRIFT	.54		.54		.54	
MAXDEC	.94		.89		.8	

$\rho = .3$

	A		B		C	
$\overline{KS}$	1.47	.67	1.21	.67	1.05	.67
VARKS	23.7	1.11	9.3	1.11	4.6	1.11
STDVKS	4.87	1.05	3.04	1.05	2.14	1.05
COEFVAR	3.3	1.57	2.5	1.57	2.0	1.57
$\overline{KQ}$	1.03	.26	.77	.26	.61	.26
$\overline{T}$ (ms)	11.1	5.0	18.2	10.0	32	20
$\overline{W}$ (ms)	7.7	2.0	11.6	4.0	18.4	8
$\overline{SERV}$ (ms)	3.4	3.0	6.6	6.0	13.6	12
Pr[0]	.56	.6	.56	.6	.56	.6
-DRIFT	.43		.43		.43	
MAXDEC	.96		.92		.85	

$\rho = .4$

	A		B		C	
$\overline{KS}$	3.14	1	2.31	1	1.86	1
VARKS	96	2	32.6	2	13.8	2
STDVKS	9.8	1.4	5.7	1.4	3.72	1.4
COEFVAR	3.12	1.4	2.5	1.4	2	1.4
$\overline{KQ}$	2.6	.5	1.76	.5	1.32	.5
$\overline{T}$ (ms)	18.9	6.02	27.7	12.0	45	24
$\overline{W}$ (ms)	15.6	3.01	21	6	31	12
$\overline{SERV}$ (ms)	3.3	3.01	6.7	6	14	12
Pr[0]	.44	.5	.45	.5	.45	.5
-DRIFT	.33		.33		.33	
MAXDEC	.97		.94		.89	

$\rho = .5$

	A		B		C		
$\overline{KS}$	6.97	1.5	4.68	1.5	3.42	1.5	
VARKS	367	3.75	116	3.75	42.7	3.75	
STDVKS	19.2	1.94	10.8	1.94	6.54	1.94	
COEFVAR	2.75	1.29	2.31	1.29	1.91	1.29	
$\overline{KQ}$	6.3	.9	4.03	.9	2.78	.9	
$\overline{T}$ (ms)	35	7.5	47	15	68	30	ρ = .6
$\overline{W}$ (ms)	31.7	4.5	40	9	55	18	
$\overline{SERV}$ (ms)	3.3	3.0	7	6	13	12	
Pr[0]	.35	.4	.35	.4	.35	.4	
-DRIFT	.25		.25		.25		
MAXDEC	.98		.96		.92		

	A		B		C		
$\overline{KS}$	16.3	2.33	9.8	2.33	6.7	2.33	
VARKS	1.4×10^3	7.78	399	7.78	139	7.78	
STDVKS	37.3	2.79	20	2.79	11.8	2.79	
COEFVAR	2.29	1.20	2.04	1.20	1.76	1.20	
$\overline{KQ}$	15.5	1.63	9.1	1.63	6	1.63	$\rho = .7$
$\overline{T}$ (ms)	70	10	84	20	115	40	
$\overline{W}$ (ms)	66.8	7	78	14	102	28	
$\overline{SERV}$ (ms)	3.2	3	6	6	13	12	
Pr[0]	.25	.3	.26	.3	.26	.3	
-DRIFT	.18		.18		.18		
MAXDEC	.98		.97		.95		

	A		B		C	
$\overline{KS}$	41.4	4	23.7	4	14.7	4
VARKS	5.8×10^3	20	1.6×10^3	20	513	20
STDVKS	75.9	4.47	40.4	4.47	22.7	4.47
COEFVAR	1.83	1.11	1.7	1.11	1.55	1.11
$\overline{KQ}$	40.5	3.2	22.9	3.2	13.8	3.2
$\overline{T}$ (ms)	156	15	179	30	221	60
$\overline{W}$ (ms)	152	12	172	24	208	48
$\overline{SERV}$ (ms)	3.2	3	6.3	6	13	12
Pr[0]	.16	.2	.16	.2	.17	.2
-DRIFT	.11		.11		.11	
MAXDEC	.99		.98		.97	

$\rho = .8$

	A		B		C		
$\overline{KS}$	134	9	73	9	42.2	9	
VARKS	3.57×10^4	90	9850	90	2690	90	
STDVKS	189	9.5	99	9.5	54.4	9.5	
COEFVAR	1.41	1.06	1.36	1.06	1.28	1.06	
$\overline{KQ}$	133	8.1	72.2	8.1	41.2	8.1	$\rho = .9$
$\overline{T}$(ms)	449	30	489	60	564	120	
$\overline{W}$(ms)	446	27	483	54	552	108	
$\overline{SERV}$(ms)	3	3	6	6	12	12	
Pr[0]	.07	.1	.08	.1	.08	.1	
-DRIFT	.05		.05		.05		
MAXDEC	.995		.991		.983		

		A		B		C	
		Voice/Data	M/M/1	Voice/Data	M/M/1	Voice/Data	M/M/1
$\overline{KS}$	ρ=.1	.138	.11	.135	.11	.132	.11
STDVKS		.475	.351	.426	.351	.406	.351
COEFVAR		3.4	3.2	3.2	3.2	3.1	3.2
	.2	.348	.25	.33	.25	.33	.25
		1.07	.56	.84	.56	.74	.56
		3.1	2.24	2.5	2.24	2.3	2.24
	.3	.7	.43	.64	.43	.59	.43
		2.3	.78	1.6	.78	1.2	.78
		3.3	1.8	2.5	1.8	2.03	1.8
	.4	1.47	.67	1.21	.67	1.05	.67
		4.87	1.05	3.04	1.05	2.14	1.05
		3.3	1.57	2.5	1.57	2.0	1.57
	.5	3.14	1	2.31	1	1.86	1
		9.8	1.4	5.7	1.4	3.72	1.4
		3.12	1.4	2.5	1.4	2	1.4
	.6	6.97	1.5	4.68	1.5	3.42	1.5
		19.2	1.94	10.8	1.94	6.54	1.94
		2.75	1.29	2.31	1.29	1.91	1.29
	.7	16.3	2.33	9.8	2.33	6.7	2.33
		37.3	2.79	20	2.79	11.8	2.79
		2.29	1.2	2.04	1.2	1.76	1.2
	.8	41.4	4	23.7	4	14.7	4
		75.9	4.47	40.4	4.47	22.7	4.47
		1.83	1.11	1.7	1.11	1.55	1.11
	.9	134	9	73	9	42.2	9
		189	9.5	99	9.5	54.4	9.5
		1.41	1.06	1.36	1.06	1.28	1.06

Selected Queue Length Statistics for All Values of ρ and All Cases

	ρ	A Voice/Data	A M/M/1	B Voice/Data	B M/M/1	C Voice/Data	C M/M/1
$\overline{T}$	.1	4.1	3.3	8.1	6.6	15.9	13.2
$\overline{W}$		.7	.3	1.3	.6	2.3	1.2
Actual $\overline{SERV}$		3.4	3.0	6.8	6.0	13.6	12.0
$\overline{T}/\overline{SERV}$ (random)		1.17	1.1	1.29	1.1	1.17	1.1
	.2	5.2	3.7	9.9	7.4	19	14.8
		1.8	.7	3.2	1.4	5.6	2.8
		3.4	3.0	6.8	6.0	13.5	12.0
		1.49	1.23	1.57	1.23	1.4	1.23
	.3	7.25	4.3	13	8.6	24	17.2
		3.9	1.3	6.1	2.6	10	5.2
		3.4	3.0	6.8	6.0	13	12.0
		2.05	1.43	2.06	1.43	1.76	1.43
	.4	11.1	5.0	18.2	10	32	20
		7.7	2.0	11.6	4	18.4	8
		3.4	3.0	6.6	6	13.6	12
		3.17	1.66	2.8	1.66	2.35	1.66
	.5	18.9	6.0	27.7	12	45	24
		15.6	3.00	21	6	31	12
		3.3	3.00	6.7	6	14	12
		5.4	2	4.4	2	3.31	2
	.6	35	7.5	4.7	15	68	30
		31.7	4.5	40	9	55	18
		3.3	3.0	7	6	13	12
		10	2.5	7.46	2.5	5	2.5
	.7	70	10	84	20	115	40
		66.8	7	78	14	102	28
		3.2	3	6	6	13	12
		20	3.3	13	3.3	8.46	3.2
	.8	156	15	179	30	221	60
		152.8	12	172.7	24	208	48
		3.2	3	6.3	6	13	12
		45	5	28.4	5	16	45
	.9	449	30	489	60	564	120
		446	27	483	54	552	108
		3	3	6	6	12	12
		128.3	10	77.6	10	41.5	10

Selected Time Statistics for All Values of ρ and All Cases (in milliseconds)

TAIL 5

		A	B	C	M/M/1
ρ =	.1	.00075	.00046	.00024	.00001
	.2	.0057	.0045	.0033	.00032
	.3	.023	.019	.015	.0024
	.4	.061	.055	.047	.001
	.5	.131	.121	.11	.031
	.6	.241	.221	.21	.078
	.7	.381	.361	.34	.171
	.8	.561	.551	.53	.33
	.9	.771	.761	.75	.59

TAIL 10

	A	B	C	M/M/1
ρ = .1	.0002	.000051	.0000065	1×10^{-10}
.2	.0016	.0008	.00028	1×10^{-7}
.3	.0078	.0047	.0024	5.8×10^{-6}
.4	.025	.018	.011	.0001
.5	.067	.052	.037	.0098
.6	.14	.12	.094	.006
.7	.27	.24	.20	.028
.8	.46	.42	.38	.11
.9	.7	.67	.64	.35

TAIL 15

	A	B	C	M/M/1
ρ = .1	7.1×10^{-5}	7.8×10^{-6}	2.8×10^{-7}	1×10^{-15}
.2	.00011	.00026	4.4×10^{-5}	1.3×10^{-10}
.3	.0043	.0019	.00061	1.4×10^{-8}
.4	.015	.009	.0041	1.1×10^{-6}
.5	.046	.031	.017	3.1×10^{-5}
.6	.11	.082	.054	.00047
.7	.22	.18	.14	.0047
.8	.40	.36	.30	.035
.9	.66	.62	.57	.20

C. A Queuing Inequality

From the previous discussion, it appears that the M/M/1 queue, i.e. $\underline{r} = \bar{r}\,\underline{1}$, is the "best" queue in the sense of minimizing $\overline{KS}$ or $\overline{T}$ among service rate vectors with mean $\bar{r}$. This is consistent with the queuing theory "metaprinciple" that, for fixed average values, "performance" degrades as "randomness" increases. One can attempt to make this notion precise in several ways, which we now explore. First some notation and "context".

For a random variable X, let F_X denote its distribution function and $F_X^c = 1 - F_X$. Note that $E(X) = \int_0^\infty F_X^c(s)ds - \int_{-\infty}^0 F_X(s)ds$. (Use integration by parts.) We assume that each term in the difference is finite, i.e. $E(|X|) < \infty$. For two random variables, X, Y, we write $X \leq_1 Y$, if $F_X^c(s) \leq F_Y^c(s)$ for all s. This notion of "inequality" is sometimes called stochastic dominance and is quite strong. For example, if X and Y are nonnegative, then $X \leq_1 Y$ implies $E(X^n) \leq E(Y^n)$, $n = 1,2,\ldots$ since $E(X^n) = n\int_0^\infty s^{n-1} F_X^c(s)ds$ for a nonnegative random variable X.

We write $X \leq_2 Y$ if $\int_{-t}^\infty (F_X^c(s) - F_Y^c(s))ds \leq 0$, for all t. This is equivalent to $E((X-t)^+) \leq E((Y-t)^+)$ for all t, where $(x)^+ = x$, $x \geq 0$; $(x)^+ = 0$, $x < 0$. Although not as strong as $\leq_1$, the notion $\leq_2$ does retain some features of a measure of the"smallness" and "determinism" of a random variable, as the following properties indicate.

1) If $X \leq_2 Y$ then $E(X) \leq E(Y)$.

Proof: By previous remarks, it suffices to show $\int_{-\infty}^{\infty}(F_X^c(s) - F_Y^c(s))ds \leq 0$. Set $G(t) = \int_t^{\infty}(F_{X(s)}^c - F_Y^c(s))ds$ $X \leq_2 Y$ implies $G(t) \leq 0$ for all t. Further, the assumptions $E(|X|)<\infty$, $E(|Y|) < \infty$ rule out the possibility that G(t) "oscillates" as $t \to -\infty$, i.e. these assumptions imply $\int_{-\infty}^{\infty}|F_X^c(s) - F_Y^c(s)|ds < \infty$ so that G(t) must converge as $t \to -\infty$. ||

2) If $E(X) = E(Y)$ and $X \leq_2 Y$, then $\text{var}(X) \leq \text{var}(Y)$.

Proof: It suffices to show $E(X^2) \leq E(Y^2)$. This follows from the hypothesis $X \leq_2 Y$ and the identity $E(X^2) - E(Y^2) = 2\int_{-\infty}^{\infty}dt[\int_t^{\infty}(F_X^c(s) - F_Y^c(s))ds]$. ||

As a "partial converse" to 2) we have

3) If X is deterministic and $E(X) \leq E(Y)$ then $X \leq_2 Y$.

Proof: If $t \geq X$, then $E((X-t)^+) = 0 \leq E((Y-t)^+)$. For $t < X$ we reason as follows. $E(X) \leq E(Y)$ implies $\int_t^{\infty}(F_X^c(s) - F_Y^c(s))ds \leq -\int_{-\infty}^{t}(F_X^c(s) - F_Y^c(s))ds$. Now for any $s \leq t$, $F_X^c(s) = 1 \geq F_Y^c(s)$ since $t \leq X$. Thus the right-side of the last inequality is ≤ 0. ||

Stoyon [25] has used $\leq_2$ to make the "metaprinciple" precise in the following way. Let A_i and B_i, i = 1,2,...

denote the interarrival and service times respectively for two G/G/1 queuing systems. Then: if $E(A_1) = E(A_2)$, $A_1 \leq_2 A_2$, $B_1 \leq_2 B_2$, then $W_1 \leq_2 W_2$, W_i = limiting waiting time. In particular, for fixed arrival time distribution A and mean service time $\overline{B}$, the A/D/1 is "best" among all A/$\overline{B}$/1. (Apply property (3).) Similarly D/B/1 is "best" among $\overline{A}$/B/1 for B fixed in distribution but A fixed only in mean. (D = deterministic.)

For the voice/data queue, we have been able to establish the following inequality.

Proposition: Let $B(t, \underline{r})$ denote the backlog (in bits) at time t for a voice/data queue with service rate vector $\underline{r}$ and with an initial speaker state drawn from the stationary distribution $\underline{p}$, and an initially empty backlog. Then

$$B(t, \overline{r}\,\underline{1}) \leq_2 B(t, \underline{r})$$

where $\overline{r} = \underline{p}^T \underline{r}$, and all other parameters, i.e. λ, μ, η, ξ, are the same in the two cases.

Our proof uses the following characterization of the backlog.

Lemma: Let U(t) denote the total number of bits arriving up to time t. Let $R(t\ \underline{r})$ denote the total potential amount of service up to time t, i.e. $R(t, \underline{r}) = \int_0^t r(A(s))ds$, $r(A=i) = r_i$. Then

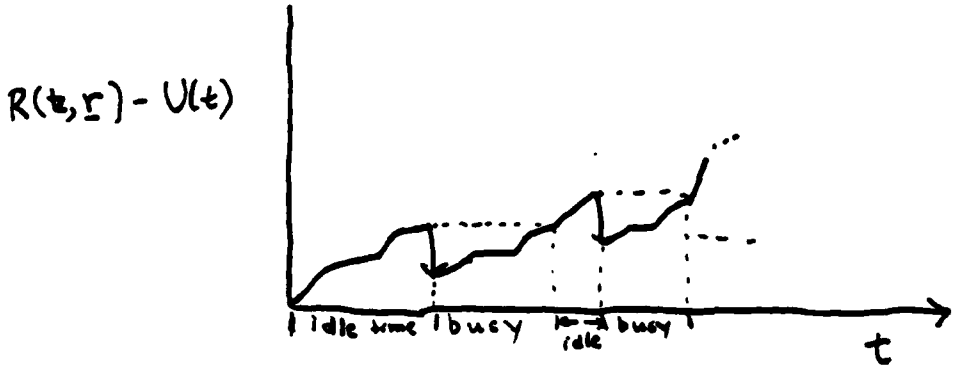
$$B(t,\underline{r}) = \sup_{0 \leq y \leq t} (U(t) - U(y) - R(t,\underline{r}) + R(y,\underline{R})).$$

Proof:

$$B(t,\underline{r}) = U(t) - R(t,\underline{r}) + \int_0^t I[B(s,\underline{r}) = 0]r(A(s))ds,$$

where $I(\cdot)$ is the indicator function. This integral term gives the amount of service that is counted in $R(t,\underline{r})$ but which should not be counted because work is not done during idles. That is, the current backlog = total arrival - total potential work + work counted during idles. The following picture "shows" that the integral is given by

$$\sup_{0\le y\le t} (R(y,\underline{r}) - U(y)).$$



It is clear that a new idle starts whenever the graph crosses its previous maximum. Thus the sup counts that part of $R(t,\underline{r})$ which accumulates in idles.||

Proof of Proposition:

Set $G(t,y,\underline{r}) = U(t) - U(y) - R(t,\underline{r}) + R(y,\underline{r})$. Since $(a-b)^+ = \sup(a,b) - b$ for any numbers a,b, the proposition is equivalent to

$$E \sup(x, \sup_{0 \le y \le t} G(t,y,\bar{r}\,\underline{1})) \le E \sup(x, \sup_{0 \le y \le t} G(t,y,\underline{r})) \text{ for all } x.$$

Since $E \sup(a_y) \ge \sup(E(a_y))$ for any collection of random variables $\{a_y\}$, since $\sup(x, \sup_y(a_y)) = \sup_y(\sup(x,a_y))$, and since $U(t)$ and $A(t)$ are independent we obtain

$$E_{A,U} \sup(x, \sup_{0 \le y \le t} G(t,y,\underline{r})) \ge E_U \sup_{0 \le y \le t} (E_A \sup(x, G(t,y,\underline{r})))$$

$$\ge E_U \sup_{0 \le y \le t} (\sup(x, E_A\, G(t,y,\underline{r}))) = E_U \sup_{0 \le y \le t} (\sup(x, G(t,y,\bar{r}\,\underline{1}))$$

The last equality uses, $E(R(s,\underline{r})) = \bar{r}\, s$, which holds because the initial speaker state $A(0)$ is drawn from the stationary distribution.

Remark: Our proof did not rely on $U(t)$ consisting of Poisson arrivals with exponential length messages, and it should work for any "nonpathological" bit arrival process, as long as arrivals and speaker activity are independent. Similarly, one should also be able to extend it to more general stationary service processes.

VII. A CONTROL PROBLEM

So far, we have discussed methods for analyzing the data queue performance as a function of given service rates that depend only on the speaker activity A. From the discussion of the TDM architecture, it is evident that there is no particular "physical" reason to change the allocation only when speaker activity changes. That is, for purposes of transmission, voice and data blocks are indistinguishable, and one can easily vary the allocation from frame to frame. In an idealized model then, one might assume that the allocation can be varied "instantaneously" (i.e. neglect the frame structure) and seek an optimal control (allocation policy) with respect to some overall cost for voice and data performance. For the data component of this cost, one can take some function of the delay or backlog. For voice, one might consider two types of costs. First, one might assume that speech is kept in some finite buffer with overflow speech being discarded, and then take the voice cost as some function of the delay/loss. Alternatively, one can assume that speech is not buffered but that there is simply a "fidelity cost per unit time" $h(i,r)$ of encoding the output of i active speakers with $(C - r)$ bps. (C = link capacity.) The average cost $\lim_{T\to\infty} E(\frac{1}{T}\int_0^T h(A(t), r(t))dt)$ can then be taken as the voice component of the total cost. We adopt this structure.

If r depends only on A, then the voice cost equals its ensemble average, $\bar{h}(r) = \sum_{j=0}^{N} p_j\, h(j,r_j)$, since the speaker chain is ergodic. If the data cost can also be expressed as some "closed form" function $\bar{f}(r_o,\ldots,r_N)$, then the optimization problem "reduces" to a static problem, e.g. $\min_{r_o,..,r_N} \bar{f}(r_o,\ldots,r_N) + \bar{h}(r_o,\ldots r_N)$ or $\min_{r_o,\ldots r_N)} \bar{f}(r_o,\ldots,r_N)$ subject to $\bar{h}(r_o,\ldots,r_N) \leq h_o$ etc. Such a "simple form" for the data cost does not appear to be forthcoming. However, by using the method of chapter V, we can, in principle, "optimize by numerical trial and error", i.e. $\bar{f}$ can be evaluated numerically for any $r_o,..,r_N$ (in the M/M case anyway) and, presumably, $\bar{h}$ can be evaluated since the p_j are known.

If the data service rate, r, is allowed to depend on speaker activity, the data backlog size, and perhaps time explicitly, the problem is more complicated, and we have no procedure for finding the equilibrium behavior of the data queue for a fixed policy. (To begin with, the notion of "stability" is more complicated. For example, one can conceive of policies that keep the backlog bounded, but that do lead to a well-defined "steady-state" behavior.) However, it is sometimes possible to characterize an optimal control, without "reducing" the problem to the static case by finding a formula relating cost and control. As a first step in exploring the complete stochastic control problem, we have

managed to characterize the optimal control for the much simpler problem of emptying an initial backlog, assuming no data arrivals or speaker activity changes. The formulation and analysis follow.

Backlog Emptying Problem

- $X(t)$ denotes the backlog at time t, $X(0) > 0$.
- $r(t)$ is the data service rate so that

$$X(t) = (X(0) - \int_0^t r(s)ds)^+$$

We allow r to be piecewise continuous with at most a finite number of jump discontinuities. At such a jump, we choose r so that it is right continuous. $r(t)$ is constrained to be in $[0,C]$ for each t.

- The voice cost per unit time is $h(r)$. We assume $h(0) = 0$, $h(r)$ nondecreasing and piecewise differentiable on either $[0,C)$ or $[0,C]$ i.e. we allow $h(r) \to \infty$ as $r \to C$. At $r = 0$, we take the right-sided derivative $h'(0^+)$ and at $r = C$, we take the left-sided derivative $h'(C^-)$ (which might be infinite). At all other points $0 < r < C$, both $h'(r^+)$ and $h'(r^-)$ exist, and $h'(r^+) = h'(r^-)$ at all but a finite number of points.
- The total cost to be minimized is

$$J(r) = \int_0^\infty X(t)dt + \int_0^\infty h(r(t))dt$$

Remark: The time, T, at which X reaches 0 is a free parameter

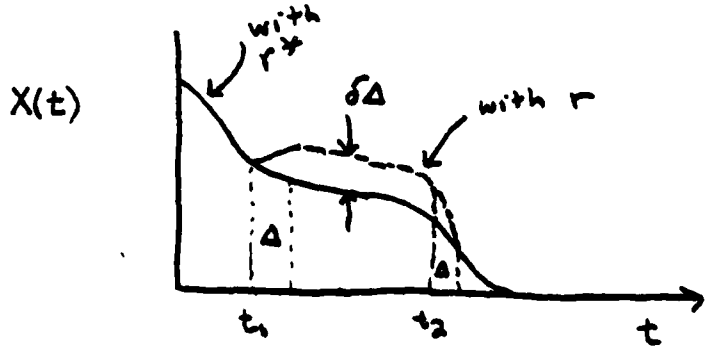
in the problem, and clearly, an optimal control $r^*(t)$ must be zero if $t > T$. A priori, T might be infinite for some h, i.e. X(t) approaches zero without ever reaching it. We will show that if there is an optimal conrrol, then this cannot be the case.

Necessary Condition for Optimality

Suppose t_1 and t_2 are two times s.t. $r^*(t_1) > 0$, $r^*(t_2) < C$. By right continuity, there exist $\delta, \Delta > 0$ such that the control

$$r(t) = \begin{cases} r^*(t) - \delta, & t \in [t_1, t_1+\Delta] \\ r^*(t) + \delta, & t \in [t_2, t_2+\Delta] \\ r^*(t), & \text{otherwise} \end{cases}$$

is admissible.



To first order in Δ, δ, the difference in costs $J(r) - J(r^*)$ is

$$(t_2 - t_1)\delta\Delta + \delta\Delta\left[\frac{h(r^*(t_1) - \delta) - h(r^*(t_1))}{\delta}\right]$$

$$+ \delta\Delta\left[\frac{h(r^*(t_2+\delta) - h(r^*(t_2))}{\delta}\right]$$

If r^* is optimal, this must be nonnegative, so we obtain

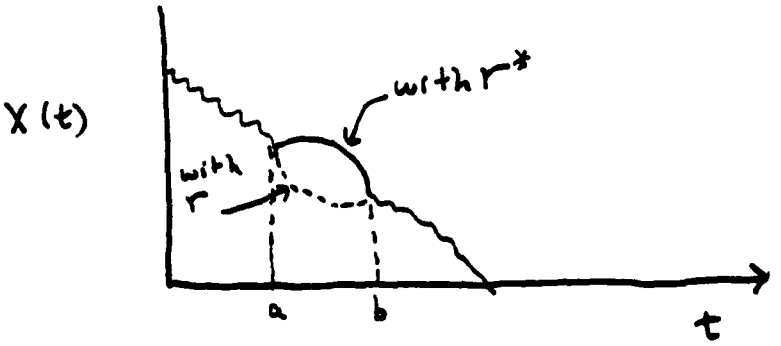
$$t_2 + h'(r^+)\Big|_{r=r^*(t_2)} \geq t_1 + h'(r^-)\Big|_{r=r^*(t_1)} \qquad (1)$$

(We have shown $t_1 < t_2$ in the picture, but this was not used in deriving (1), i.e. (1) is a property of an optimal control at any times t_1, t_2 meeting the specified conditions. In particular, if $C > r^*(t_1), r^*(t_2) > 0$, (1) also holds with the roles of t_1 and t_2 reversed.)

We can now show that there is a time T at which X does reach 0 (assuming an optimal r^* exists). First, we note that, trivially, there is a control with finite cost, e.g. $r^* = r_o\, I(0 \leq t < X(0)/r_o)$ where $I(\ )$ is the indicator function and $0 < r_o < C$ (since $h(r) < \infty$ if $r < C$). Second, we note that an optimal control must be nonincreasing. To see this, observe that if r^* is increasing on [a,b], then the control

$$r = \begin{cases} r^*(t) & t \notin [a,b] \\ r^*(b-(t-a)) & t \in [a,b], \end{cases}$$

(which just reverses r^* in time over [a,b]) has the same voice cost but lower data cost.



Now, evidently, there is some time t_o at which r^* assumes a value r_o less than C. Since $h'(r^+)$, $h'(r^-)$ are nonnegative and $h'(r^+) < \infty$ if $r < C$, it follows from (1) that $r^*(t)$ can be positive only if $t \leq t_o + h'(r_o^+)) < \infty$. Thus there is some time t_1 at which $r^*(t_1) = 0$. By monotonicity, $r^*(t) = 0$, for $t > t_1$, so if $X(t_1) \neq 0$, $J(r^*)$ is infinite and r^* cannot be optimal. Letting $T = \inf \{t:X(t) = 0\}$ it follows from the continuity of $X(t)$ (r^* has no "impulse") and the right continuity of r^* that $X(T) = r^*(T) = 0$; $x(t) > 0$, $r^*(t) > 0$, for $t < T$.

Since $r^*(t) > 0$, $t < T$, we can apply (1) to obtain, for $0 \leq t \leq T$

$$h'(r^-)\Big|_{r=r^*(T-t)} \leq t + h'(0^+) \tag{2}$$

$$t + h'(r^-)\Big|_{r=r^*(T^-)} \leq h'(r^+)\Big|_{r^*(T-t)} \tag{3}$$

$$\text{if } r^*(T-t) < C.$$

(In (3), if $r^*(T^-) = 0$, we mean $\lim_{r\to 0^+} h'(r^-)$ which must be $h'(0^+)$.) It follows from either (2) or (3) that $h'(r^-)\Big|_{r=r^*(T^-)} \leq h'(0^+)$. Now we can use these conditions to construct r^*.

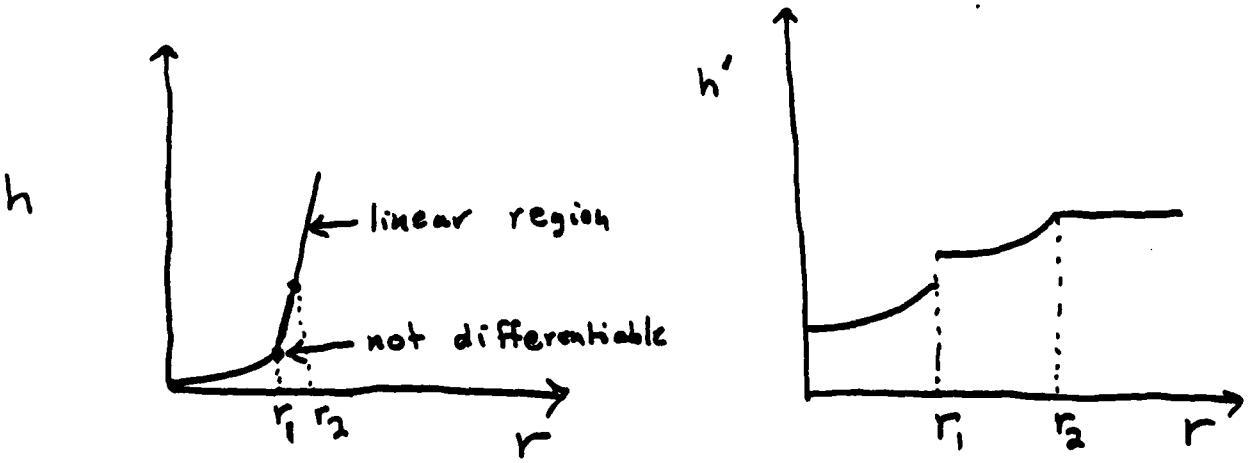
Case 1: h is convex $\cup$. In this case, $h'(r^-) \geq h'(0^+)$ for all $r > 0$. Thus, by the previous remark, we can conclude $h'(r^-)\Big|_{r=r^*(T^-)} = h'(0^+)$, so from (2) and (3) we obtain

$$h'(r^-)\Big|_{r=r^*(T-t)} \leq t + h'(0^+) \tag{4}$$

$$h'(r^+)\Big|_{r=r^*(T-t)} \geq t + h'(0^+) \tag{5}$$

$$\text{if } r^*(T-t) < C.$$

The basic idea is to work "backwards" in time from T and use (4) and (5) to pick off a value for $r^*(T-t)$. The construction stops when, for some t_o, $\int_{T-t_o}^{T} r^*(s)ds = X(0)$, so that one then "redefines the origin" and $T = t_o$, i.e. the free parameter T is determined by this condition. We illustrate the possibilities by the following example.



In the region $0 \leq r < r_1$, h is strictly convex and differentiable so (4) and (5) hold with equality, and there is a unique solution, $r^*(T-t) = [h']^{-1}(t + h'(0^+))$. For t in the region $h'(r_1^-) - h'(0^+) \leq t \leq h'(r_1^+) - h'(0^+)$, any $r \leq r_1$ satisfies (4) but only $r = r_1$ satisfies (5). Thus $r^*(T-t)$ sits at r_1 until t reaches $h'(r_1^+) - h'(0^+)$, and it then proceeds up the next portion of the curve. Once t exceeds $h'(r_2)$, any $r_2 \leq r \leq C$ satisfies (4), but if $r < C$, (5) does not hold. Thus $r^* = C$ must be chosen and then maintained until the condition specifying T is met. More generally, we see that a linear portion in h causes a jump discontinuity, and a nondifferentiable point causes r^* to remain constant for some time. Otherwise $r^*(T-t)$ increases monotonically and continuously with t.

Case 2: h is not convex. This reduces to the previous case as follows. Let $\hat{h}$ be the convex hull of h. Then, since $\hat{h} \leq h$, an optimal policy for $\hat{h}$ which only uses values for r^* at which h and $\hat{h}$ agree, must also be optimal for h. Now in an interval (a,b) in which $\hat{h}$ and h disagree, $\hat{h}$ is linear and "joins" h at the endpoints. By the previous construction, an optimal policy for $\hat{h}$ will only use a or b if it uses any r in [a,b].

VIII. FURTHER RESEARCH

• Control Problem - Our aim has been to determine the extent to which data and voice can share a link's transmission capacity. A basic theme has been the "service demand vs. service supply" mismatch problem, and we have seen that changes which allow faster movement between "high" and "low" data service states improve performance. The analysis of data queue performance for fixed rates $r_0,\ldots,r_N$ was roughly based on the assumption that, because of its delay requirements, voice must have "nearly" complete priority, i.e. r cannot change until A changes. (One could perhaps optimize data queue performance over choices of $r_0,\ldots,r_N$ that yield equivalent overall speech quality levels, but there is not too much flexibility.) An important conceptual question is whether a "dynamic control", i.e. r depends on voice activity and backlog size, offers substantial improvement. The following "heuristic" argument indicates a possible reason for expecting improvement. It seems that a large part of the delay in queuing systems is incurred by a small percentage of customers, in particular, those arriving during or right after "surges" which overwhelm the server. A numerical example supporting this "interpretation" can be found by considering a truncated M/M/1 queue, with room for say L customers. (Overflow is "lost".) For $\rho = .9$ and $L = 20$ the loss is 1.3%, and the average number in system is $.6\ \frac{\rho}{1-\rho}$. At $L = 10$, the

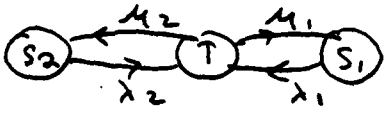
loss is 5%, and average number in system is .26 $\frac{\rho}{1-\rho}$. The same reduction in delay would not occur if we rejected 1.3% or 5% of customers at random, i.e. the finite buffer "targets" the rejection during surges. In the context of the voice/data queue, this might mean that we can significantly improve data performance by providing brief but large "bursts" of capacity at the right moments, i.e. by "backing off" on voice priority at certain times we might be able to improve data performance while maintaining acceptable speech quality. It is not clear how brief these bursts can be and still help the data.

- More Queuing Inequalities - We conjecture that other queuing inequalities of the type in VII.C are true. One might consider inequalities involving system time, number in system, backlog etc. and various notions of "stochastically smaller", but the basic conjectures fall into these categories.

  - All other things being equal, performance improves monotonically as λ, μ increase, λ/μ fixed.

  - All other things being equal, performance improves as one "equalizes" service rates, i.e. raise some r_i that is below $\bar{r}$ and reduce some r_i above $\bar{r}$ in such a way that the average is maintained. Since the vector $\bar{r}\,\underline{1}$ is the "ultimate equalization", a result of this type involving the backlog and the stochastic inequality measure $\leq_2$, would imply our previous result in VI C.

- Matrix-Geometric Analysis With Other Voice Models -
  The matrix geometric approach works with any underlying Markov chain for the phase process. It would be interesting to carry out the numerical calculations for the three state single speaker model



  (In this case, the phase process for N speakers would be the Markov process $(T(t), S_1(t))$, see I.C.) This model is more realistic for small numbers of speakers since the effects of two different types of silences do not "wash out" until, apparently, N is 25-30. Note that in this model, a substantial portion of the silence time comes from the "long, infrequent" silences in S_2, i.e. the alternation between talkspurt and a part of the silences is on a slower scale. Thus, one would expect degraded performance.

- General Length Distributions (i.e. the M/G case) -
  If message lengths are not exponentially distributed, then the process [A(t), K(t)] is not Markov. However, the process [A(t), X(t)], X(t) = bit backlog, is, and one can derive coupled integro-differential equations for the joint equilibrium probabilities $G_i(x) = \Pr[A = i, X \leq x]$. (The equations are similar to those in the flow model in Chapter VI, except that the discrete nature of message arrivals leads to terms which are

convolutions of the $G_i(x)$ and the message length distribution.) These equations have been studied by Halfin and Segal [26], [27]. Since they are more difficult to handle (than the matrix-geometric approach), one would be interested in knowing the "sensitivity" of the data queue performance to the message length distribution, i.e. for what purposes can one assume exponentially distributed message lengths.

- The equation $\underline{\sigma} = \underline{\alpha} + \underline{\beta}\ \underline{\sigma}^2$. The expressions for $\underline{\alpha}$ and $\underline{\beta}$ are relatively "simple", and, more important, the basic parameters of the problem appear in a rather "direct" way. It would be nice to have some relation (or bound) on some quantities pertaining to $\underline{\sigma}$ (e.g. eigenvalues) in terms of the corresponding quantities for $\underline{\alpha}$ and $\underline{\beta}$. Also, the convergence rate of the iteration when $\Pi^T(\underline{\alpha} - \underline{\beta})1 < 0$, i.e. when we are computing the strictly substochatic matrix $\underline{\sigma}^+$, needs to be determined.

APPENDIX A.

The exact expression for T_n^- is

$$\frac{1}{n\mu}\ \frac{1}{p_n}\ \sum_{j=n}^{N} p_j\ ,$$

so our task is to estimate this, when the $\{p_j\}$ are the terms of a binomial distribution. For convenience, we will only consider symmetric distributions i.e. $\lambda = \mu$. This does not result in loss of generality since the asymptotic expressions can be appropriately modified for asymmetric distributions. The value of some constant might change, but we are not concerned with this. Also, we will assume $\mu = 1$, since for fixed λ/μ, T_n^- is inversely proportional to μ, i.e. the $\{p_j\}$ depend only on $\frac{\lambda}{\mu}$. Any result we quote is taken from Feller [28], Chapter 6.

Notations:

- 2N is the number of speakers, so $\bar{A} = N$, and the standard deviation $\sigma_N = \sqrt{N/2}$.
- Let $\phi(y) = (\sqrt{2\pi})^{-1} \exp(-y^2/2)$, $\Phi(y) = \int_{-\infty}^{y} \phi(t)dt$
- $p(n) = \binom{2N}{n} 2^{-2N}$
- $H(n) = \sum_{j=n}^{N} p(j)$

We consider the asymptotic behavior of

$$T^{-}_{G(N)} = \frac{N(G(N))}{\mu_{G(N)}\, p_{G(N)}} \quad \text{for } G(N) = N + f(N)$$

Region I: $\lim f(N)/\sqrt{N} < \infty$

From Feller:

$$p(G(N)) \sim \frac{1}{\sqrt{\pi N}} \exp(-f^2(N)/N)$$

$$H(G(N)) \sim 1 - \Phi(f(N)/\sigma_N)$$

$$\therefore \quad T^{-}_{G(N)} \sim \frac{1}{N+f(N)} \; \frac{\sqrt{N\pi}\,(1 - \Phi(f(N)/\sigma_N))}{\exp(-f^2(N)/N)}$$

e.g.

- $f(N)/\sqrt{N} \to 0, \quad T^{-}_{G(N)} \sim \frac{1}{2}\sqrt{\frac{\pi}{N}}$

- If $f(N)/\sigma_N \to \gamma, \quad T^{-}_{G(N)} \sim \sqrt{\frac{\pi}{N}} \; \frac{1 - \Phi(\gamma)}{\exp(-\gamma^2/2)}$

For large values of γ, we can apply the assymptotic result $1 - \Phi(y) \sim \frac{\phi(y)}{y}$ to obtain $T^{-}_{G(N)} \sim \sqrt{\frac{\pi}{N}} \; \frac{1}{\gamma\sqrt{2\pi}} = \frac{1}{2\gamma\,\sigma_N}$

Region II: $f(N) = X_N\,\sigma_N$ where $X_N \to \infty$ but $X_N/\sigma_N \to 0$.

Let $s(X_N) = \sigma_N^2 \sum_{\ell=3}^{\infty} \frac{(\frac{1}{2})^{\ell-1} + (-\frac{1}{2})^{\ell-1}}{\ell(\ell-1)} (\frac{X_N}{\sigma_N})^{\ell}$

then from Feller, we have

$$p(G(N)) \sim \frac{\phi(X_N)}{\sigma_N} \exp(- s(X_N))$$

$$H(G(N)) \sim (1 - \Phi(X_N)) \exp (-s(X_N))$$

Combining this with $1 - \Phi(y) \sim \frac{\phi(y)}{y}$, we obtain

$$T^-_{G(N)} \sim \frac{1}{N+f(N)} \frac{\sigma_N}{X_N}$$

$$\sim \frac{1}{N+X_N \sigma_N} \frac{\sigma_N}{X_N}$$

$$\sim \frac{1}{2X_N \sigma_N}$$

Region III: $f(N)/N \to \gamma$, $0 < \gamma \leq 1$

From the simple one step bounds $\frac{1}{\mu_n} \leq T^-_n \leq \frac{1}{\mu_n - \lambda_n}$,

we obtain $\frac{1}{N+f(N)} \leq T^-_{G(N)} \leq \frac{1}{2f(N)}$. For $f(N)$ in this region , these become $\frac{1}{(1+\gamma)N} \leq T^-_{G(N)} \leq \frac{1}{2\gamma N}$, in the limit.

We will show that the upper bound is exact, in the asymptotic sence, i.e. for any $\delta > 0$, $\exists\, N_0$ s.t. $\forall\, N \geq N_0$

$$T^-_{G(N)} \geq \frac{1}{2\gamma N} - \frac{\delta}{N} \quad .$$

Hereafter, all $\geq$ statements used in the asymptotic sence will be in the above context, and we write $\geq \sim$ without going through the "for any δ, $\exists$ N" argument in a formal way.

Let j be a given positive integer. Then $\frac{H(j)}{p(j)} = \left(\sum_{\ell=j}^{2N} \binom{2N}{\ell}\right) \Big/ \binom{2N}{j}$

Now $\binom{2N}{j+m} = \binom{2N}{j} \prod_{\ell=1}^{m} \frac{2N-j-\ell+1}{j+\ell}$. Therefore $\frac{H(j)}{p(j)} = 1 + \frac{2N-j}{j+1} +$

$+ \frac{(2N-j)(2N-j-1)}{(j+1)(j+2)} + \cdots\cdots$

Now fix an integer k. Because each term in the sum is positive and because, for fixed x, $\frac{x-y}{x+y}$ is decreasing in y, we obtain the bound

$$T^{-}_{G(N)} = \frac{1}{G(N)} \frac{H(G(N))}{p(G(N))} \geq \sum_{\ell=0}^{k} \left(\frac{N-f(N)+k}{N+f(N)+k+1}\right)^{\ell}$$

$$= \frac{1 - \left(\frac{N-f(N)+k}{N+f(N)+k+1}\right)^{k+1}}{1 - \frac{N-f(N)+k}{N+f(N)+k+1}} \frac{N+f(N)+k+1}{N+f(N)}$$

$$= \frac{1 - \left(\frac{N-f(N)+k}{N+f(N)+k+1}\right)^{k+1}}{2f(N)+1} + \left[1 + \frac{k+1}{N+f(N)}\right]$$

Now observe that if $f(N)/N \to 0$ as $N \to \infty$, then $\left(\frac{N-f(N)+k}{N+f(N)+k+1}\right)^{k+1} \to 1$ as $N \to \infty$, and the bound is useless. However, if $f(N)/N \to \gamma$, then this expression approaches

$$\frac{1 - \left(\frac{1-\gamma}{1+\gamma}\right)^{k+1}}{2\gamma N}$$

Since k is arbitrary, we can conclude $T^{-}_{G(N)} \geq \sim \frac{1}{2\gamma N}$.

REFERENCES

[1] Weinstein, C.J., "Fractional Speech Loss and Talker Activity Model for TASI and for Packet-Switched Speech", IEEE Trans. on Comm., Vol. 26, No.8, Aug. 1978, pp. 1253-1257.

[2] Gitman I. and Frank, H., "Economic Analysis of Integrated Voice and Data Networks: A Case Study", Proceedings of the IEEE, Vol. 66, No. 11, Nov. 1978, pp. 1549-1570.

[3] Bially, T., Gold, B., and Seneff, S., "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks", Lincoln Lab Report, Jan 1978.

[4] Brady P., "A Model for Generating on-Off Speech Patterns in Two-Way Convesation", Bell Syst. Tech. Journal, Sept. 1969, pp. 2445-2471.

[5] Feller, William, An Introduction to Probability Theory and Its Application, Volume II, Wiley Press, 1971.

[6] Berger, R., "The Queuing Behavior of a Buffered TASI Multiplexer", Lincoln Lab Report, Aug. 1979.

[7] Keilson, J., Markov Chain Models -- Rarity and Exponentiality, Springer-Verlag, Applied Mathematical Sciences Series, Number 28, 1979.

[8] Gantmacher, F.R., The Theory of Matrices Vol. II., Chelsea Publ., 1960.

[9] Strang, G., Linear Algebra and Its Applications, Second Ed., Academic Press, 1980.

[10] Karlin, S. and MacGregor, J.K., "On a Genetics Model of Moran", Proc. Comb. Philos.Soc., pp. 299-311, 1962.

[11] Bellman, R. and Harris, T., "Recurrence Times for the Ehrenfest Model", Pacific Journal of Mathematics 1 (1951), pp. 179-193.

[12] Neuts, M., Matrix-Geometric Solutions in Stochastic Models, 1980 Draft, to appear.

[13] Keilson, J.K., Private Communication.

[14] Keilson, J., Zachmann, M., and Sumita, U., "Row Continuous Finite Markov Chains: Structure and Algorithms", M.I.T. Laboratory for Information and Decision Systems, LIDS-P-1078, March 1981.

[15] Keilson, J., Green's Function Methods in Probability Theory, Charles Griffin, 1965.

[16] Yechiali, U. and Naor, P., "Queueing Problems with Heterogeneous Arrivals and Service", Opns. Res., 19, 1971, pp. 722-734.

[17] Yechiali, U., "A Birth and Death Process", Opns. Res., 21, 1971, pp. 604-609.

[18] Loynes, R.M., "A Continuous-Time Treatment of Certain Queues and Infinite Dams", Journal of Australian Math. Soc., pp. 484-498, 1961-1962.

[19] Loynes, R.M., "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times", Proc. Camb. Philos.Soc., 58, pp. 497-520, 1962.

[20] Loynes, R.M., "Stationary Waiting-Time Distributions For Single-Server Queues", Annals of Math. Stat., pp.1323-1339, 1962.

[21] Zachmann,M.,Private Communication.

[22] Humblet, P.A., Private Communication.

[23] Kingman, J.F.C., "A Convexity Property of Positive Matrices", Quarterly Journal of Mathematics Oxford (2) 12 (1961), pp. 283-284.

[24] Kleinrock, L., Queueing Systems Vol. I, Wiley and Sons, 1975.

[25] Stoyon, From P.A. Humblet, Private Communication.

[26] Halfin, S. and Segal M., "A Priority Queueing Model for a Mixture of Two Types of Customers", SIAM J. Appl. Math., Vol. 23, No. 3, Nov. 1972, pp. 369-379.

[27] Halfin, S., "Steady-State Distribution for the Buffer Content of an M/G/1 Queue with Varying Service Rate", SIAM J. Appl. Math., Vol. 23, No. 3, pp. 356-363.

[28] Feller, William, An Introduction to Probability Theory and Its Applications", Vol. I, Wiley Press, 1968.