PAPER P-1155

ON CAPACITY ALLOCATION STRATEGIES FOR DoD COMMUNICATION SATELLITES

January 1976

Joseph M. Aein
Ostap S. Kosovych

INSTITUTE FOR DEFENSE ANALYSES
SCIENCE AND TECHNOLOGY DIVISION

IDA Log No. HQ 75-17866
Copy 139 of 215 copies

The work reported in this document was conducted under contract DAHC15 73 C 0200 for the Department of Defense. The publication of this IDA Paper does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that agency.

PAPER P-1155

ON CAPACITY ALLOCATION STRATEGIES FOR DoD COMMUNICATION SATELLITES

January 1976

Joseph M. Aein
Ostap S. Kosovych

INSTITUTE FOR DEFENSE ANALYSES
SCIENCE AND TECHNOLOGY DIVISION
400 Army-Navy Drive, Arlington, Virginia 22202

Contract DAHC15 73 C 0200
Task D-10

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Paper P-1155	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): On Capacity Allocation Strategies For DoD Communication Satellites.	5. TYPE OF REPORT & PERIOD COVERED: FINAL January 1975-January 1976
6. PERFORMING ORG. REPORT NUMBER: P-1155	
7. AUTHOR(s): Joseph M. Aein; Ostap S. Kosovych	8. CONTRACT OR GRANT NUMBER(s): DAHC 15-73-C-0200
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Institute for Defense Analyses, 400 Army-Navy Drive, Arlington, Virginia 22202	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Task D-10
11. CONTROLLING OFFICE NAME AND ADDRESS: Assistant Director (General-Purpose Systems) Office of Director, Telecommunication and Command and Control Systems (DTACCS)	12. REPORT DATE: January 1976
	13. NUMBER OF PAGES: 330 (12) 315 pp.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): DTACCS	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(18) IDA/HQ

(19) 75-17966

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

None

18. SUPPLEMENTARY NOTES

N/A

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

communication satellite capacity allocation, circuit-switched service, store-and-forward service, message-switched service

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

Several allocation strategies are investigated for sharing the satellite capacity among user communities requiring circuit-switched or store-and-forward (message-switched) communication service. Teletraffic and queueing models are used to obtain analytical results by which the allocation strategies are compared. It is shown that the capacity required to provide a specified grade of service is strongly dependent upon the particular allocation strategy chosen. Therefore, an allocation

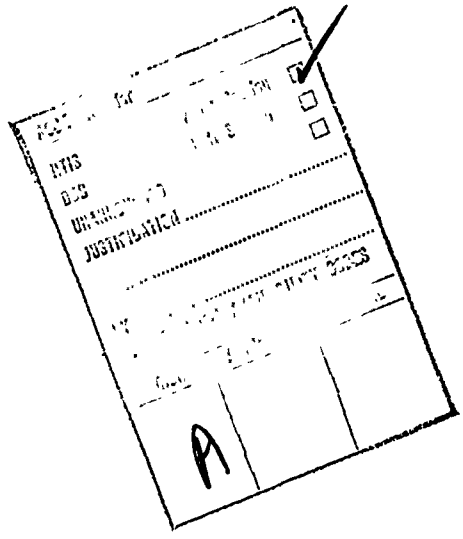
OVER

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

UNCLASSIFIED 403 108

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. strategy can be selected which requires less capacity and hence provides more efficient utilization of the satellite capacity for a specified grade of service.



CONTENTS

GLOSSARY

ABNCP	Airborne Command Post
ADP	Automatic Data Processing
AF	Air Force
AFSAT	Air Force Satellite
AFSATCOM	Air Force Satellite Communications
AM	Amplitude Modulation
AUTOSEVOCOM	Automatic Secure Voice Communications
BPO	Blocking Probability Objective
BPSK	Binary Phase-Shift Keyed
CAU	Central Acquisition Unit
CDMA	Code-Division Multiple Access
CINC	Commander in Chief
CINCNET	CINC Network
Comsat	Communication Satellite Corporation
CPU	Central Processing Unit
CUDIXS	Common User Digital Information Exchange Subsystem
DA	Demand Access
DCA	Defense Communications Agency
DD	Destroyer
DE	Destroyer Escort
DL	Destroyer Leader
DoD	Department of Defense
DSCS	Defense Satellite Communications System
DTACCS	Director, Telecommunications and Command and Control Systems

EAM	Emergency Action Message
ECM	Electronic Countermeasures
EIRP	Effective Isotropic Irradiated Power
FDMA	Frequency-Division Multiple Access
FLEETSAT	Fleet Satellite
FLEETSATCOM	Fleet Satellite Communications
FLTCOM	Fleet Command Voice Network
FLTSATCOM	Fleet Satellite Communications
FM	Frequency Modulation
FSB	Fleet Satellite Broadcast
FSK	Frequency-Shift Keyed
GOS	Grade of Service
HICOM	High Command Network
IDA	Institute for Defense Analyses
IF	Intermediate Frequency
I/O	Input/Output
INTELSAT	International Telecommunications Satellite
IXS	Information Exchange System
K-P	Khintchine-Polloczek
NAVCOMSTA	Naval Communication Station
NCA	National Command Authorities
OSD	Office of the Secretary of Defense
PSK	Phase-Shift Keyed
QPSK	Quaternary Phase-Shift Keyed
R&D	Research and Development
RF	Radio Frequency
RLC	Resistance, Inductance, Capacitance
Satcom	Satellite Communications
SHF	Super High Frequency
SIOP	Single Integrated Operational Plan

SSIXS	Submarine Satellite Information Exchange System
SSMA	Spread-Spectrum Multiple Access
SSTIXS	Small Ship Teletype Information Exchange System
STDM	Synchronous Time-Division Multiplex
TACINTEL	Tactical Intelligence Digital Circuit
TADIXS	Tactical Data Information Exchange System
TDMA	Time-Division Multiple Access
TRITAC	Tri-Service Tactical Communications System
TSCIXS	Tactical Support Center Information Exchange Subsystem
TTY	Teletype
UHF	Ultra High Frequency
U.S.	United States

EXECUTIVE SUMMARY

A. OBJECTIVE AND SCOPE

This paper presents the results of part of a task undertaken by the Institute for Defense Analyses (IDA) at the request of the Director, Telecommunications and Command and Control Systems (DTACCS), Office of the Secretary of Defense. The objective is to support the development and evaluation of military command and control concepts and techniques that use satellite relays, placing emphasis on network control concepts and key technical problem areas. In this study, conceptual strategies are investigated for the efficient, flexible, and timely allocation of DoD satellite communication (satcom) resources to terrestrial terminals.

As another part of the same task, an assessment has been made of R&D issues relating to nuclear-event-induced propagation degradations to satcom systems. That assessment is reported upon separately in IDA Paper P-1154.

In the operational use of DoD satcom systems, it is expected that large fluctuations in communication traffic volume, variations in available satellite capacity, and network reorganization may occur. Consequently, the objective of network control must be to balance available capacity against traffic demand. Within the current DoD satcom programs* there will be a wide diversity of terminal capabilities, data rates, and interconnections, with limited satcom resources. The existing methods of telecommunication system traffic engineering are deemed inadequate for the emerging DoD satcom programs.

---

*I.e., the Defense Satellite Communication System (DSCS), Fleet Satellite Communications (FLEETSATCOM), and Air Force Satellite Communications (AFSATCOM).

Therefore, the primary effort in this study is devoted to initiating the development of a quantitative methodology for

- Identifying key system traffic factors and system capacity-sharing techniques
- Evaluating expected performance
- Comparing capacity-sharing techniques.

This study addresses the concept of control as an allocation of capacity to networks of users. Not directly addressed are issues of hardware for implementing control. Those issues are under study by the respective DoD program offices and have been very specific and particularized to individual system design detail. Rather, effort here is mainly devoted to developing an analytical methodology for relating user terminal traffic to satellite capacity and comparing alternative arrangements or strategies for a sharing of available capacity by user networks. Analytical tools for describing networks, traffic, and capacity relationships also provide support in the following areas:

Policy Formulation

- Numerical assessment of network architectural options for the capacity needed
- Relation of network to spacecraft architecture
- Guidance of multiple-access and base-band technology development
- Structuring of control issues.

Transmission System Operations

- Assessment of system operations
- Adjustment of transmission and traffic flow parameters within a net

changes in traffic demand or available capacity.

Planning

- Theoretical guidelines for system design
- Guidance in selection of values for simulation parameters
- Checkpoints for simulation results.

The analysis builds upon the theoretical methods of conventional telephonic traffic engineering and of the newly developing field of computer data communications. The new dimensions inherent in DoD satcom systems are:

- Reduction in the connectivity constraints of geography and time, due to wide-area satellite coverage
- Variability in satellite power and bandwidth needed to support a link, due to variations in terminal antenna size, power, noise environment, and data rate
- Variability in available satellite capacity, due to preemptions of transponder assets, enemy jamming, and the eventual partial failure of transponders.

Also, for store-and-forward traffic, the effects of long-path delay that are inherent in any satellite communications must be taken into account.

B. APPROACH

The principal contribution of this study is to demonstrate by example that, for wide variations in DoD satcom use, mathematical models can be constructed that provide quantitative measures for judging the equity of allocations of satcom capacity to meet traffic demands.

Communication services of the circuit and store-and-forward types are studied separately. Hybrid systems having circuits

with store-and-forward traffic intermingled in transmission are not investigated.

As modeled, traffic is generated by stationary random sources of exponential type, characterized by two parameters: (1) an average rate of message generation or call request and (2) an average message length or call duration. The traffic parameters can themselves vary but are assumed to change slowly with respect to the time needed to achieve a steady-state condition in system traffic flow. The theory of queues is then employed as the principal mathematical tool. The communication transmission system is viewed as a structured and adjustable set of servers in support of the traffic customers. As the traffic parameters change, the servers can be readjusted.

With suitable multiple-access modulation, it is possible to functionally amalgamate much, if not all, of the individual capacity provided by each of the operating transponders on a communications satellite. The resulting capacity is then shared among the accessing terminals on *a demand basis, constrained* by rules of access [i.e., sharing strategies, according to terminal class and (sub)network]. This concept is depicted in Fig. 1, wherein the overall capacity is partitioned into three categories, (1) denied, (2) circuits, and (3) store and forward. Denied capacity is capacity lost or not available because of special preemption, jamming, or partial system failure. The remaining available capacity is divided between circuit service and store-and-forward nets.

Strategies for sharing circuit capacity among classes of users are studied first. Two common candidate strategies are:

1. Dedicated--Subdivision of the total available functional capacity into separate pools, each dedicated to its subnetwork of user terminals but having reallocable capacity balanced against traffic.

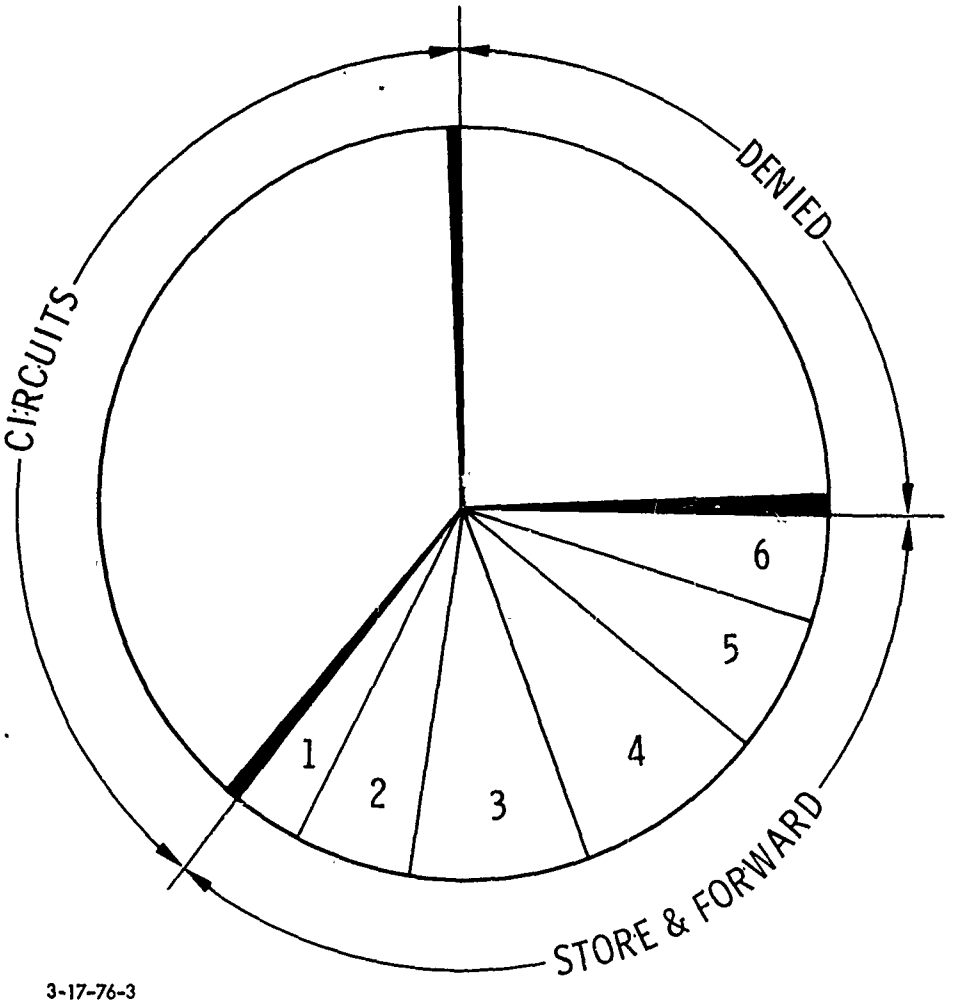


FIGURE 1. Conceptual Capacity Allocation Architecture

2. Fully Shared--Combination of all the terminals into one (super) network that fully shares all the available capacity.

Various intermediate strategies exist, of which one described in this paper is of special interest for internetwork traffic and for traffic overflow.

For store-and-forward data service (Fig. 1), each store-and-forward net is provided its own capacity. Consideration is then given to time-sharing strategies for data exchange within a

net, as well as to the capacity the net requires in order to operate. Thus, store-and-forward nets can be thought of as operating on dedicated circuits.

A critical concept is that any slicing of the capacity pie of Fig. 1 should not be permanent. Capacity can be made reallocable. At the very least, as the levels of average traffic demand vary within user nets, capacity allocations should be adjusted to maintain overall system balance. In a jamming environment, the traffic models developed are applicable provided enough residual capacity is left beyond what is needed to operate a minimal system-control orderwire. In this event, the issue of balancing traffic and capacity becomes especially significant, as it is necessary to off-load traffic as well as to reallocate capacity.

The general approach used here is not tied to any particular multiple-access modulation system. To a first approximation, the model applies to time-division, frequency-division, or spread-spectrum multiple access. Whichever technique is employed, the model assumes that capacity is subdivided linearly among the active users, each consuming an amount proportionate to his terminal capability (receive sensitivity and transmit power) and information rate.

C. LIMITATIONS

This study may be viewed as a useful beginning in relating traffic engineering to communication satellite technology for DoD applications. By no means should it be taken as final. It is the belief here that the subject should evolve with satellite technology as military needs and traffic patterns develop.

Time and a lack of basic theory prohibited pursuit of several areas of special importance to military communication systems. Chief among these areas is the problem of analyzing the effects of priority and preemption features. To a very

limited extent, such analysis can be accommodated within the dedicated strategies, as these have the most equivalence to theory applicable to the existing terrestrially oriented systems. For any of the other sharing strategies, the analytical problems tend to become intractable within the currently available mathematical formulations. Thus, it is reasonable to look to a simulation approach to these problems.

Previous experience indicates that priorities weakly affect average traffic flow or throughput, although they very much affect speed of service to individual call or message requests. Consequently, analytical results without the priority/preempt features may aid in choosing the key ranges of parameter values for a simulation. They can also aid in interpreting the simulation results.

In a like vein, only quasi-stationary exponentially distributed traffic models are employed. Beyond the use of satellite circuits as cable equivalents, there is as yet no data base from which real traffic statistics may be derived. The quasi-stationary exponential models have been useful in the nonmilitary sector. Whether they are reasonable in the military sector remains to be seen. However, models assuming quasi-stationary behavior are useful for determing critical operating points for system stability, while models of exponential-type traffic are useful in guiding the collection of real traffic data and its reduction to traffic models.

Finally, not all multiple-access modulation techniques subdivide capacity in a linearly additive manner, as assumed here. For those which divide capacity in an algebraically nonlinear way (e.g., spread-spectrum systems), the theory would appear to be easily extendable. For those which show frequency-dependent nonlinearity (e.g., frequency-division systems), the analysis appears to be intractable, and the system in all likelihood would have to be operated in a restricted manner so as to emulate no worse than algebraically nonlinear capacity division.

D. RESULTS

1. Circuit Traffic

The users are grouped into classes, each of which is characterized by its traffic level and the amount of circuit capacity used per call. For K classes of users, the state of the system is given by the number of calls in progress (i.e., active circuits) from each of the K user classes. It is shown that the strategy, or rules, by which available satellite capacity is shared amongst the user classes is represented by the allowable states of the system. For example, certainly no state of the system is allowed that would exceed the available capacity. Arriving calls are blocked from service if they would move the system to an unallowed state. The effect of choosing a sharing strategy is manifested in the resulting blocking performance to each user class. In every case, the sharing strategy can be represented geometrically by an object in K dimensions (e.g., the dedicated strategy is a rectangular solid, while the fully shared strategy is a tetrahedron). The upper edge surfaces of the sharing strategy contain blocking states* for the different user classes.

The probability of call blocking is the conventional measure of performance or grade of service and is used here. It is found by determining the probability of occurrence of each of the allowable states and summing over those blocking edge states. The form of this state probability is found to have an important fundamental form for the case studied (i.e., blocked calls cleared). Independent of the sharing strategy, the probability of each allowed state is always a product of two factors. One factor depends only on the traffic characteristics of the user classes and each of the allowed states of the system. It can be generated from traffic requirements independently from, and in advance of, specification of a

*Those states in which new call requests cannot be accepted.

transmission system. The other factor depends on the system, the sharing strategy, and traffic levels but is independent of system state. Varying the sharing strategy does not affect the first (requirements) factor, but only the second factor. Furthermore, the second factor is physically interpreted as the probability of there being no calls in progress. Consequently, in monitoring or simulating traffic, a fundamental parameter to observe is the percentage of time there are no calls in progress.

Given the basic probabilistic form, the computational challenge of evaluating each state probability and thence the blocking state probabilities can become stressing for systems with a large number of states. A computer program has been written which evaluates and compares the dedicated and fully shared strategies (for up to three classes of users) on their need for capacity and average utilization factor to meet a specified grade of service.

2. Store-and-Forward Traffic

For each store-and-forward net, an appropriate level of capacity is dedicated to operate each net. Three classes of substrategies for further time-sharing the capacity of a net are then analyzed and compared as follows:

1. Fixed Assignments. Each member of the net is assigned a specified interval of transmission time that occurs periodically and that is his and only his to use. The assignments are made by a net controller on the basis of estimates of the average traffic needs of the net members and are then distributed to the net in an assignment message.

2. Access with Polling. The net members are sequentially polled according to a predetermined order by the net controller, with the polled user transmitting until

the message buffer is empty. At the end of transmission, the next member is polled and the process repeats.

3. Access with Reservations. When a message arrives to be transmitted, the net user requests transmission time for the message from the net controller. The controller allocates transmission time to the requesting user, and at the appropriate time the user transmits the message.

The two grades of service measures used for evaluating a sharing strategy are (1) the average queue length of messages awaiting transmission by a terminal and (2) the average message queueing time measured from generation by a user to the time to completely transmit the message. Expressions for the minimum capacity needed by a net are determined in order to prevent net saturation or instability. Then, an analytical basis is provided for trading off additional capacity allocated to a subnet among a grade of service desired by the users, the total available satellite capacity, and demand by other subnets. The time-sharing strategies studied are then compared on the basis of the capacity required by a strategy to meet a specified grade of service (e.g., average message queueing time).

E. CONCLUSIONS

In general, no single sharing strategy or arrangement is uniformly best.

- Changes in traffic levels and operational constraints can alter the technique preferred and the grade of service provided and certainly will alter the system operating parameters of a chosen technique, thus exhibiting the need to develop quantitative relationships between traffic, capacity allocation, and control.

- Analytical results can provide alternative measures for the assessment of network status, which is needed for system control. Monitoring of delay or blockage at key points in a system can be used to estimate operating traffic levels and to confirm or possibly eliminate traffic monitoring by each user in the system.

For circuit traffic,

- Sample calculations show that, above certain minimal levels of net traffic, no significant savings in the overall capacity required accrue due to amalgamating the users and fully sharing all of the available capacity rather than allocating capacity separately to each user net. This conclusion is justified only if capacity can be reallocated to each net in order to match the level of traffic existent in the net.
- On the other hand, at low levels of traffic within a net, the dedicated allocation of capacity can require as much as 50 percent more capacity than if the net were merged with other low-traffic nets.

For store-and-forward data transmission systems,

- The more time-dynamic strategies (polling or reservations, as opposed to fixed assignments) usually provide savings in the capacity required to achieve a specified message queueing time for a given level of message traffic. However, the fixed assignment technique is more attractive for those systems where the time a net controller takes to make transmission assignments to user terminals is large compared to the average message transmission time.
- For the fixed assignment technique, the user time allocation must be carefully determined, as system

performance can become sensitive to traffic load and allocation. Optimal time allocations can result in delays as much as 50 percent less than those caused by intuitive allocations. For polling or reservation systems, time allocations are intrinsic in the technique and therefore require no optimization. All techniques require a determination of traffic needs and adequate capacity allocation from the overall satellite capacity to support the net traffic.

I. INTRODUCTION

A. BACKGROUND

In August 1974 IDA was requested* by OSD/DTACCS to undertake studies in the area of military satellite communication (satcom) systems important to the command and control of U.S. forces. Two principal areas of concern were identified as follows:

- The impact of nuclear-event-induced propagation degradations to satcom systems

and

- "A study of the technical means for achieving, allocating, monitoring and controlling efficient, flexible, and timely netting of mobile satcom radio terminals in the general purpose forces."

A study of the first of these areas is reported upon separately in IDA P-1154 (Ref. 1). This paper addresses itself to the second area, netting satcom terminals. In particular, effort is devoted to developing theoretical methodology for data exchange applicable to satellite systems as a means for quantitatively relating user-community traffic to needed satellite capacity and comparing methods for sharing available capacity between user communities.

*The task statement is reproduced in Appendix A.

The current DoD satcom programs* which establish a context for this study are the Defense Satellite Communication System (DSCS), Air Force Satellite Communications (AFSATCOM), and Fleet Satellite Communications (FLEETSATCOM). The systems are summarized in Appendix B. Of these, FLEETSATCOM provided the initial motivation, but the results should also be applicable to the DSCS and AFSATCOM as they evolve. In Appendix B, FLEETSATCOM serves as an example for exhibiting specific system parameters and their relation to the various elements of the traffic theory studied.

B. OBJECTIVE AND SCOPE

The DoD satcom programs, especially those utilizing mobile or transportable earth terminals, are characterized by a diversity of (1) terminal capability, (2) interconnections and data flow in terminal networks or subnetworks according to military mission and/or command, and (3) electronic countermeasure environment. The interplay of these factors is manifested by asking whether a quantitative relationship can be developed between

- Communication traffic requirements, i.e., networks or subnetworks of terminals operating together through a given satellite

and

- The management or allocation of satellite capacity, bandwidth, and power to support the various terminal networks.

The principal objective of the study is to develop quantitative means for judging an equitable balance between communication

*Additional summary information regarding these programs is contained in Ref. 2.

traffic and the capacity allocated to service it. The principal contributions of this report are a mathematical formulation of the problem and a set of analytical techniques.

The general statement of the objective is in the tradition of classic telephonic traffic engineering, dating back to the theoretical methods first initiated by Erlang in 1917. The features of DoD satcom systems which add new dimensions to classic traffic engineering are:

- The physical variability of satellite capacity required to support links between differing types of terminals according to terminal receiver sensitivity, transmit power, and link information bandwidth.
- The temporal variability of available satellite capacity through spacecraft transponder asset reassignment, capacity reduction by electronic warfare, and partial failure of transponder assets.
- The removal of connectivity and transmission constraints. In satcom systems terminal interconnections and line speed can be altered dynamically within a matter of tens of minutes, if not instantly, whereas months may be required in terrestrial systems.

The store-and-forward transmission of digital data messages, blocks, or packets, as opposed to assignment of circuits, presents a different body of theory, one which is currently being developed for computers-to-remote-terminal communications (i.e., teleprocessing) in terrestrial facilities. Modification of these techniques is implicit in satcom applications due to the long-path delay inherent in satellite communications and and the attendant timing considerations.

As an example, for a given information bandwidth, satellite capacity required to support links to the various mobile Naval

terminal configurations for use with FLEETSATCOM could vary by more than a factor of eight.* Depending on communication modes, information bandwidths can vary by a factor of 210 (from teletype bit rates up to secure voice). In an extreme case, a terminal with low receiver sensitivity operating with a large information bandwidth would require a dedicated transponder. This, in turn, would reduce available satellite capacity to the remaining mix of terminals capable of sharing transponders.

With suitable multiple-access modulation** it is possible in principle to amalgamate the available capacity from each of the operating transponders and then share this among the accessing terminals by rules or constraints of access, i.e., sharing strategies, according to terminal class or subnetwork. For example, two possible strategies could be (1) to subdivide the total capacity into separate pools, one for each subnetwork of terminals, or (2) to combine all the terminal subnetworks into one class that fully shares all the available capacity. Many intermediate strategies exist. In addition, for operating store-and-forward data nets, consideration must be given to substrategies for data exchange within a net.

A critical concept is that partitioning capacity need not be done on a permanent basis. It is here postulated that capacity can be assigned according to a balance between the time-dependent traffic needs of a terminal net and the capacity available. Capacity reallocation should occur as the gross traffic levels vary within the user communities. Thus, the principal objective becomes defining terminal net traffic,

---

*That is to say, receiver gain varies more than 6 dB and system noise by more than 3 dB, so that the ratio of terminal antenna gain to receiver noise temperature (G/T) can vary by more than 9 dB.

**For AFSATCOM and FLEETSATCOM, the time-division multiple-access (TDMA) technique is suitable, while for DSCS, the frequency-division multiple-access (FDMA) and spread-spectrum multiple-access (SSMA) techniques may also be used. A general reference on multiple access is provided in Ref. 3.

capacity-sharing methods, and a quantitative method for negotiating the amount of capacity to be allocated for achieving a communication objective or grade of service.

In a jamming environment, the models are assumed to apply, the total available capacity being reduced commensurately with the effect of the jammer in power-robbing the satellite transponders. This would include any transponder suppression effects. A central issue is traffic off-loading and capacity reallocation. Consequently, the issue of communication resource utilization in an ECM environment further motivates the desire for quantitative methods of balancing capacity use against traffic needs.

Issues relating to the specifics of system control are not directly addressed in this report. Through discussion with DTACCS, a decision was made to concentrate effort on a study of capacity allocation methods for the following reasons:

- There does not yet appear to be an adequate theoretical base upon which to build quantitative methods of judging issues of system control, whereas such a base does exist for capacity allocation.
- The need for quantitative understanding of capacity allocation issues was judged to take precedence over the need to investigate control techniques. A general approach to capacity allocation suitable to DoD satellites, although feasible, was not being pursued in the DoD community.
- In view of the above, a general approach to the control problem as applied to the varicus DoD satcom systems did not seem feasible. Control methods currently evolve intrinsically with

specific transmission system design and development.* That is to say, control and call signalling are implemented for a given type of spacecraft transponder (e.g., DSCS, FLEETSAT), earth terminal parameters (e.g., EIRP, G/T), multiple access (e.g., TDMA, FDMA, SSMA), and terrestrial user switch interface (e.g., TCC-39).

Thus, the lack of a general approach to the control problem, the wish to avoid needless duplication of effort on specific DoD satcom systems, and the potential for a general approach to capacity allocation led to the concentration of effort herein reported.

The physical transmission system serves to define the values of capacity available for use and may constrain the latitude with which it may be allocated. Within the latitude provided by the "hardware," there is still the need to make capacity assignments reasonably "efficient." The control system then affects the capacity assigned to the users as nets and also affects their individual operation within a net. Advanced terminal hardware components, especially multiple-access modems and RF bandwidth selection, have been designed without the ability to compare the overall system cost/benefits of implementing flexible capacity reallocation. In the absence of such an assessment, simpler, less costly terminal components are procured which inhibit capacity reallocation. It is hoped

*Several DoD contractors are pursuing system-specific control studies. The interested reader is referred to Refs. 4-6 for application to FLEETSAT TDMA. An advanced development model of a TDMA modem is constructed for the U.S. Army in Ref. 7. Reference 8 further addresses control for field army demand-access satcom features. Reference 9 reviews control mechanization chosen for INTELSAT application.

that the approach taken here will serve to provide a bridge between satellite transmission technology and multiple-access modulation methods on the one hand and utility to variable user traffic on the other.

C. APPROACH

The study seeks mathematically representative models of satellite traffic demand and means of organizing the demand (i.e., access to satellite capacity) with requisite levels of satellite capacity to satisfy some measure of adequate performance (e.g., channel or circuit availability, delay, throughput). The model objectives are completely analogous to those of conventional telephonic traffic engineering and more recent computer communication engineering.

The general approach used here is not tied to any particular method of multiple access. Thus, to a first approximation, it is equally valid for the frequency-division multiple-access (FDMA), time-division multiple-access (TDMA), and spread-spectrum multiple-access (SSMA)* techniques. The multiple-access technique, whichever used, is modeled as providing a functional pool of capacity which can be subdivided and assigned to terminals individually or to nets of terminals, which, in turn, could further share their allotted capacity. The terminals and/or nets take, on a per-user basis, different amounts of capacity, depending on terminal sensitivity, power, and bandwidth. The users in each network aggregate generally request transmission service in some form of demand assignment mode. The capacity in use at any one time is modeled to be the linear sum of all the capacity consumed by the actively transmitting users. Strategies for sharing out the capacity

*The literature also calls this class of technique "code-division multiple access (CDMA)."

and methods for evaluation are then examined in the context of this abstract, simplified (e.g., linear) model of a sharing mechanization.

For example, the per-user capacity taken from a satellite by a terminal with a receiver sensitivity figure of merit G/T,* for a downlink with information bandwidth B, would ideally be proportional to $B(G/T)^{-1}$. The maximum capacity available would depend on the satellite power, the largest terminal (G/T) value in the system, and the available RF bandwidth. In principle, TDMA comes closest to realizing this ideal. Capacity allotment on the downlink is controlled by the amount of time allotted to each assessing terminal in proportion to $B(G/T)^{-1}$.

It is important to digress a moment in order to amplify upon the relationship between linear use of capacity and the two other classes of multiple-access techniques, namely SSMA and FDMA. SSMA introduces uplink effects on transponder power-sharing among active carriers. As long as the sum of the power of all active uplink carriers remains constant, capacity will divide linearly amongst the users according to their proportion of uplink power. If the total carrier power arriving at the satellite should vary with active traffic, capacity would subdivide algebraically according to the ratio of desired to total uplink transmitter power. In this case, however, the system could be linearized to the "worst-case" maximum total uplink power. Interestingly enough, should one carrier (or, more likely, a jammer of constant power) dominate the uplink, the residual capacity left to the remaining uplink carriers would again divide linearly, independently of the summed power of the desired carriers.

*The ratio of a terminal's antenna gain to receiver noise temperature.

For FDMA, capacity division depends not only on the nature of the uplink power levels but also on the intermodulation product spectrum generated. Should an uneven intermodulation spectrum result, capacity division would become frequency dependent.

Thus, extension of the linear capacity division studied here to nonlinear capacity division as determined by multiple-access modulation techniques is an important avenue for future study. Certainly, for TDMA and linearized versions of SSMA and FDMA, the models used here apply. Moreover, it is believed that the models can be generalized to include algebraically nonlinear capacity division.

Returning to the development of the analytical methods used in this study, it is further assumed that the interface between the terminal-associated base-band digital system (e.g., teletype, printers, secure voice codes) and the terminal modem(s) can be such that the satellite transmission system (from the transmitting terminal through the satellite transponder to the receiving terminal) *can be made transparent to the operation of the users' base-band digital system.* This assumption represents a desirable practical goal. Space-related technology and user digital-device (i.e., computer) technology are independently evolving, each at its own rapid rate. It is desirable in satcom system development and application to preserve flexibility between satellite transmission technology and user data-system technology.

The methods studied are separated into two distinct areas of application: (1) data circuit connections and (2) store-and-forward data message transmission. For both classes of service, user traffic is modeled as a stochastic process, and the transmission facility is modeled as a server system with constraints imposed by the transmission system. Queueing-theory analysis is then employed to relate traffic load, system capacity

allocation, and a measure of performance (e.g., message delay or calls blocked). In this manner, the available satellite capacity is assumed to be split between circuit modes and store-and-forward data modes of communication service. This partition can be made variable in accordance with the traffic needs of the two classes of service. Each class of service is then studied to make the most efficient use of its available capacity.

The first area addressed is terminal access configurations suitable for supporting data circuits for such communication service as secure voice and bulk data transfer. This effort builds upon conventional traffic engineering theory with the introduction of new features associated with multirate circuits and terminal variations in power and sensitivity. Next addressed are various store-and-forward transmission strategies for sending data blocks. This form of communication is relevant to the exchange of teletype narrative messages, and, perhaps as important in the near future, it will support inquiry/response data flow between computer(s) and remote terminals. This portion of the effort builds upon current data communication research studying computer polled terminals on a multidrop telephone line (or data loop).

The dichotomy imposed between data-circuit as opposed to store-and-forward service is representative of current data transmission practice. It also reflects current planning for FLEETSATCOM and AFSATCOM use. Time did not permit a study of intermingling the circuit and store-and-forward service with hybrid transmission or of converting one type of service into another, as, for example, packetized voice. Also not studied in depth were random-access transmission systems, typified by the ALOHA approach, which seek to minimize if not eliminate access control. These approaches could become attractive for future systems.

D. ORGANIZATION OF THIS REPORT

Chapter II summarizes the study in greater depth and presents its findings in more detail than was possible in the brief Executive Summary. Chapters III and IV are central to this report and are relatively self-contained, treating the development of theoretical results for the circuit and store-and-forward types of traffic, respectively. Appendix A presents a copy of the task statement, Appendix B surveys the current DoD satellite communication programs, and Appendixes C through E give mathematical detail in support of the main text.

II. DISCUSSION AND FINDINGS

A. GENERAL

For both circuit and store-and-forward* applications, traffic is modeled as being generated by purely random traffic sources of exponential type characterized by two parameters: (1) an average rate of message or call generation and (2) an average length of message or call duration. The satellite communication system is modeled as an appropriate server system which provides a transmission service to each piece of generated traffic or "customer." Various queueing models appropriate to the transmission strategies are then analyzed to mathematically relate traffic load, capacity usage or utilization, and a performance measure such as average message delay or number of calls blocked. Mixes of traffic sources having differing traffic parameters are important features studied. However, store-and-forward sources are separated from circuit sources. Traffic is assumed to be quasi-stationary in time, i.e., slowly varying in a time frame relative to achieving steady state.

In all cases, traffic arrives with a demand for transmission service. This is provided on a first-come, first-served basis, with access constraints as determined by the transmission system and sharing strategy. Dedicated service is not addressed other than to observe that if such capacity is needed, it reduces

*Store-and-forward service, also called message traffic, includes not only narrative (telegram) messages but computer-to-remote-terminal data exchange messages.

the capacity available for demand service. Perhaps of more military significance, priority traffic and preempt features are not addressed. These introduce theoretical complications beyond the resources available for this study. This is clearly an important area for future investigation.

The sources of circuit-type traffic are divided into a finite number of user classes. The members of each class are modeled to have the same traffic level and per-circuit capacity requirement, but these parameters vary between the classes. The total capacity in use at any moment is assumed to be the linear sum from all of the user classes of all the per-circuit capacity taken by each of the calls in progress at that moment.

A newly arrived call request for a circuit will be blocked unless the unused capacity allocated to or shared with other user classes is adequate to support the circuit capacity required by the call request at its moment of arrival. The average percentage of calls blocked defines a performance measure or grade of service. Given a total available level of satellite capacity for circuit service, the problem is then how it shall be further apportioned and/or shared amongst the user classes. The study objective is *to select and compare capacity sharing and partitioning strategies against user class traffic level and per-circuit capacity draw for a given probability of call blocking or "grade of service."*

A store-and-forward system is assumed to support a group or net of terminals having message sources with unequal traffic parameters. A common satellite channel* would be time-shared

*This may be a physical piece of satellite hardware, but ideally a virtual channel would be provided. That is to say, an assignment of a portion of satellite total available capacity would be allocated to the net according to net data traffic load. In this regard, it is tacitly assumed that all members in a store-and-forward net operate with a common amount of capacity, the maximum needed by the weakest net terminal. There is no reason for this restriction other than the time and resources available for this study. Removal of this limitation remains as an interesting theoretical extension to the work reported here.

sequentially by the net terminal members operating at a common transmission speed determined by channel capacity. Messages would arrive at net member terminals and be buffered, awaiting each terminal's turn to transmit. The rules by which net members are ordered to transmit their messages and the length of transmission time allotted to each net member determine a sharing strategy. The average delay experienced by a typical message is a measure of performance and depends on the net traffic load and the channel capacity allocated.

Circuit traffic does not allow subdivision of a call for piecemeal transmission service, as do store-and-forward messages. Consequently, there are considerably more possible transmission strategies for store-and-forward traffic than for circuit traffic. For circuit systems without priorities and preempts, the only latitude is the scope permitted user groups to access the total capacity. For store-and-forward systems, considerably more degrees of freedom in design are available as to how terminals sequentially transmit their subdivided messages. This results in more diversity in the analytical technique needed to study store-and-forward systems. On the other hand, in the study of circuit systems, at least in the more elementary initial phase pursued here, a more unified theoretical base may be used.

Beyond the lack of service priority/preempts and the separation of service into message traffic and circuit traffic, several other important theoretical restrictions are placed upon the models and analysis used here. These are addressed further later in the report. They are aggregated into important areas for further research in Section II-E at the end of this chapter.

B. CIRCUIT SWITCHED SYSTEMS

1. System Model

Circuit switched systems are modeled as comprising K user classes. Each class is characterized by the number of traffic

sources N in the class, the call intensity of a source is given by the average number of calls placed during the average call holding time, and the per-circuit capacity is drawn from the satellite by an active call from the class.

Each user from each class generates calls independently of all other users. When a user makes a call, he demands service. If no capacity is available, the source is blocked and is modeled to return to a status "statistically" identical to not having attempted the call (i.e., "Blocked Calls Cleared"). That is to say, the effect (frustration) of being blocked is not modeled, and there is an immediate call reattempt. No priorities for service within a class are assumed. More satisfactory procedures for handling blocked calls ("Blocked Calls Held" or "Delayed") can be implemented along with priority/preempt features. These features add considerable complexity to the theory. Although not analyzed, they are discussed in Chapter III.

The active number of calls in progress from each of the K user classes is a "state." A state is described by a K-dimensional vector $j \equiv (j_1, j_2, \ldots j_K)$ whose coordinates j_i are nonnegative integers representing the calls in progress for each of the K user classes. Constraints are placed upon the states j which are allowed by the availability of capacity and the manner in which it is shared between user classes. For example, if there is a total capacity C_o available and each user from class i, i = 1, 2, ... K, uses c_i units of capacity for a circuit, then surely the capacity in use, $c_1j_1 + c_2j_2 + \ldots c_Kj_K$, must not exceed C_o. If a new call request were to arrive such that the capacity in use would exceed C_o, it could not be accepted, and that call would be blocked. Consequently, the allowed states would form a set Ω for which each vector j in Ω would satisfy the inequality

$$c_1j_1 + c_2j_2 + \ldots c_Kj_K \leq C_o.$$

Additional constraints can be imposed on the manner in which states are allowed to occur. Suppose the available capacity C_o were partitioned into K separate capacity pools, C_1, C_2, ... C_K, so that $C_1 + C_2 + \ldots + C_K = C_o$. Then, assign each pool of capacity to its correspondingly numbered user class for dedicated use by it and no other class. This can be represented as an allowable class of states A by j vectors whose j_i coordinates satisfy the following inequalities:

$$c_1 j_1 \leq C_1$$

$$c_2 j_2 \leq C_2$$

$$c_K j_K \leq C_K \ .$$

The access strategy is mathematically represented by specifying a set of allowable states A. Clearly, the set A must be contained in the set Ω. The two examples given represent two opposite strategies. The first, denoted as fully shared, is the Ω set which aggregates all user classes with equal access to any part of the total available capacity C_o. The second strategy, opposite to the first, allocates the total capacity into K separate pools or dedicated channels for exclusive use by each user class. (Note that although dedicated to class i, the capacity C_i assigned should be adjusted to balance the i user class traffic.) Other example strategies (A-sets) are described in Chapter III. In one strategy of potential interest, most of the available capacity is dedicated to the K user classes, but some is held in reserve as an "overflow channel." Calls that are blocked on the dedicated channel overflow to the reserve channel, where they fully share the reserve capacity with all other overflowed calls.

Interest in examining different A-sets is driven by the blocking interactions that result between user classes. The dedicated strategy completely eliminates interclass blocking

effects. Generally, the fully shared strategy uses least capacity but may not adequately satisfy blocking objectives of different user classes.

In every case the A-set can be represented as a geometrical object in K dimensions. This is shown in Fig. II-1 for K = 2 and 3. The dedicated A-set is a K-dimensional rectangular box resting on the positive coordinate axes with dimensions $C_1 \times C_2 \times \ldots C_K$. The fully shared A-set (i.e., Ω) is the tetrahedron enclosed between the positive coordinate axes and a tilted plane* offset from the origin by a distance proportional to the total capacity C_o. Thus, more available capacity permits more allowable states. Note that the rectangular box of the dedicated A-set fits inside the fully shared volume Ω with its furthermost corner point touching the upper bounding plane of Ω.

The subset of states which cause newly arrived calls to be blocked lies on the outer upper edges of the A-set. Blocking probability or grade of service will be determined by summing the probability of being in an edge state of the A-set. Thus, the transmission capacity, the strategy for sharing it, and the blocking conditions are mathematically characterized by the A-set. The connection between the user traffic and the transmission system is made by determining the steady-state probability $P(j)$ that a particular state j in A is occupied. $P(j)$ is physically the probability that $j_1, j_2, \ldots j_K$ circuits are in progress from the K user classes.

*The equation for this plane is $\Sigma c_i j_i = C_o$. This is a direct property of the linear model for capacity use. It is believed that the theory can be extended to algebraic nonlinearities which can be expressed geometrically as smooth bounding curved linear surfaces in the K-dimensional state space.

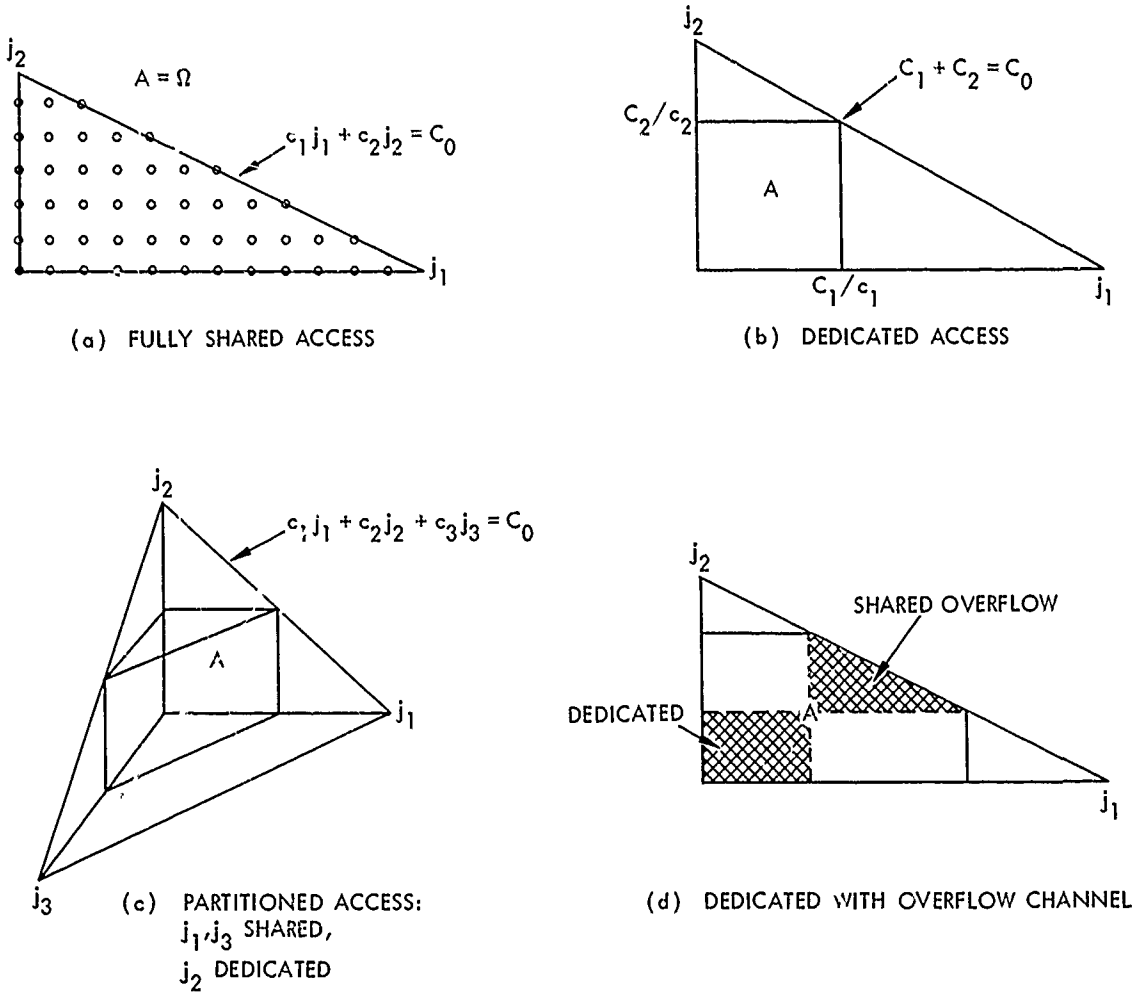


FIGURE II-1. Access Strategies and A-Sets

The steady-state probability values $P(j)$ are found by solving a second-order K-dimensional difference equation (i.e., "equation of state") whose coefficients are determined by the traffic parameters and whose boundary conditions are determined by the edge surfaces of the A-set. Because the traffic is statistically modeled as purely random with exponential distributions, the equation of state is of the

Birth-Death type Markov process. When the dedicated A-set is chosen, it is physically clear that one is dealing with K non-interacting channels in parallel.* This, then, reduces to the classical (one-dimensional) Erlang model (which must be solved K times, once for each channel). For the fully shared (and all other) A-sets, one seeks to "separate" the equations of state in a manner analogous to the dedicated A-set. The separating property is of great practical as well as theoretical value, and its study was one of the major efforts as shown in Chapter III and Appendix C.

Once the basic probabilities $P(j)$ of each allowed state or call occupancy state are found, key system performance measures can be calculated. The grade of service or blocking probabilities are computed by summing $P(j)$ along the appropriate boundaries of the A-set. The average of the number of calls in service at any moment can be found. This, in turn, determines system utilization, defined as the average capacity in use divided by the total available capacity.

2. Theoretical Results

The principal theoretical result is a characterization of the general validity of the solution form** for $P(j)$ independent of the A-set, provided the A-sets satisfy a very desirable system objective. Specifically, $P(j)$ has the following form, independent of the A-set sharing strategy, *provided the A-set permits the circuit capacity used by an active call to be immediately returned for reuse when the call is completed:*

$$P(j) = P_A(0) \cdot \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i} , \qquad \text{(II-1)}$$

*Note, however, that one is still at liberty to choose the separate capacity allotments to each channel.

**Remember that the boundary conditions for the state equation depend on the A-set so that the solution form for $P(j)$ could depend on the A-set.

where, for each user class i = 1, 2, ... K,

$j = (j_1, j_2, \ldots j_K)$ = state vector

j_i = number of calls in progress from user class i

N_i = number of sources in class i

$\binom{N_i}{j_i}$ = number of combinations of N_i objects taken j_i at a time

$a_i = \lambda_i/\mu_i$ = the source call intensity given by the average number of call generations per source during average call holding time for class i users

$P_A(0)$ = a normalization constant, independent of j.

$$1/P_A(0) = \sum_{j \in A} \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i} \qquad \text{(II-2)}$$

Equation II-1 has an extremely important form. It is a product of two quantities. Notice that the K-fold product of combinatorial factors (the traffic factor) *depends only on the traffic and is totally independent of the transmission system and sharing strategy (A-set).* These numbers can be computed directly from projected traffic levels ("requirements") and the independent variables j_i, the number of calls in progress from class i, with no reference to the transmission system structure.

On the other hand, the normalization constant $P_A(0)$ or system factor in Eq. II-1 is completely independent of any particular state j in the A-set. For a given specification of traffic level and per-circuit capacity draw of each user class, the system factor $P_A(0)$ can be calculated for different A-sets (this also includes changes in available capacity). Thus, *the form of* $P(j)$ *does not change* as the A-set or the traffic level is varied.

Although analytically elegant and simply stated, the solution of Eq. II-1 is deceptive in that evaluation of $P_A(0)$ in Eq. II-2 computationally grows quickly on the number of user classes K and the number of allowable states j or, equivalently, large numbers of users with adequate capacity to support them. Thus, there should be future interest in developing numerical and asymptotic techniques for reducing the computational load.

Another highly significant result derives from the form of Eq. II-1. Examination of Eq. II-1 shows that $P_A(0)$ equals the probability that the system is empty (i.e., no calls in progress). Consequently, if one were to simulate a system and its variants rather than directly computing $P_A(0)$ via Eq. II-2, the basic outcome of the simulation that would need to be monitored is the number of times the system is empty. In the limit of a large number of observations, the percentage of them in which there are no calls in progress is $P_A(0)$.

It is also shown that the following theoretical properties hold:

- Although $P(j)$ in Eq. II-1 is a product of terms in j_i for any valid A-set, the calls in progress from each user class are *not* statistically independent. Only when capacity is allocated on a dedicated basis to each user class do the j_i become statistically independent.
- The conventional relationship between time and call congestion* for one user class holds for K user classes.

*The probability $P(j)$ is referred to as "time congestion." Call congestion is the probability of circuit occupancy as seen by a newly arriving call. If the conventional relation holds, the numerical difference between call and time congestion is significant. The theoretical relation is provided in Appendix C.

- For the fully shared A-set, if the user classes are ordered by ascending need of per-circuit capacity, then, independent of traffic levels, the probability of blocking for each user class follows the same order.

3. Numerical Results

A computer program was developed (Appendix C) to evaluate Eq. II-1 and numerically compare dedicated and fully shared strategies for the amount of capacity C_o needed for a specified probability of blocking. It included both finite and infinite user source models. This program was written in Fortran IV and was executed on a remote time-sharing computer system. It uses a computationally straighforward approach and, as a consequence, is limited to systems of relatively small scale. No more than three user classes are permitted (e.g., $K \leq 3$, small, medium, and large capacity terminals). In addition, the determination of the blocking probability for the fully shared strategy is approximated with an upper bound calculation (see Appendix C). Since the true probability of blocking is less than or equal to the bound, the capacity required for a given blocking objective will be somewhat overestimated for the fully shared case.

The program is given the traffic parameters (N_i, $a_i = \lambda_i/\mu_i$) and the per-circuit capacity need c_i, as well as the blocking objective or grade of service GOS_i (i.e., probability of blocking = GOS_i) for each user class $i = 1, 2, 3$. It then calculates, using the conventional one-dimensional Erlang B (or Engset) equations, the dedicated capacity C_i needed for each user class to achieve its GOS_i objective. The total capacity C_o is then the sum of the C_i.

Using the C_o calculated for the dedicated strategy per the above, the program then calculates an upper bound approximate to the blocking probabilities that result with the fully shared

strategy. In all cases, this will result in a blocking probability which is better than the grade-of-service (GOS) objective. The value of C_o is then reduced by a small fixed amount, and the process is reiterated until the blocking objective is just met. The total capacity calculated on the "iterate" just prior to that exceeding the GOS objective is then that needed for the fully shared strategy. Note that the calculations assume that traffic parameters are known (i.e., measurable) and that capacity is allotted in a manner to match the traffic need.

The dedicated and shared strategies for capacity usage are compared first as a function of offered traffic for two different levels of blocking probabilities (0.05 and 0.01) and two mixes of user class characteristics. The two mixes of user sources considered are

1. Balanced Traffic: the product $N_i c_i$ = constant
2. Equal Traffic: N_i = constant, thus $N_3 = N_2 = N_1$.

For this comparison, the per-circuit capacity c_i is taken to be $1 = c_1 = c_2/2 = c_3/4$. For example, user bit rates of 1/2 kb/s, 2.4 kb/s, and 4.8 kb/s would generate the indicated c_i. The finite source model (Engset) is used, and the source traffic parameters $a = \lambda/\mu$ are all taken to be equal to 0.1 (a 10 percent user duty factor). The offered traffic is varied by changing the number of sources N_i in the user class. A sample plot of capacity required and the average fraction of capacity used (utilization) versus offered traffic levels* is shown in Fig. II-2 for a GOS objective of 0.05. Additional calculations for smaller traffic levels and a comparison of infinite source models with finite ones are provided in Chapter III.

*For the Engset model, offered traffic level is determined for each user class by the quantity $[a/(1+a)]N$ and is measured in dimensionless units denoted as "Erlangs." The total traffic offered is the linear sum of traffic from each user class.

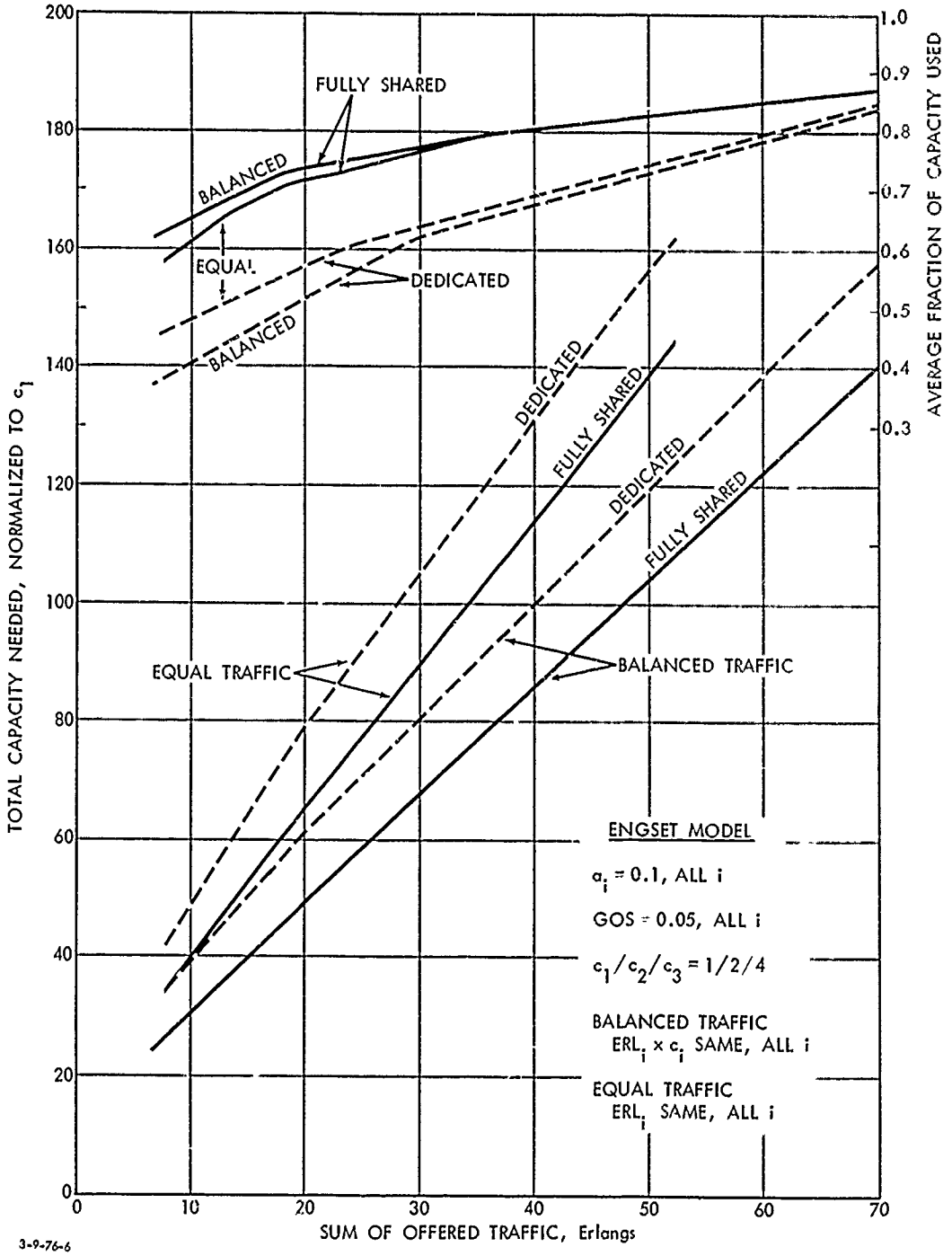


FIGURE II-2. Capacity Comparison, Dedicated versus Fully Shared for Two Traffic Mixes: Balanced and Equal

As a last example, a hypothetical mix of Naval task force users is considered. The three classes of circuits modeled* are as follows:

Class 1. Leader Class

Typical of circuits off medium flagships [e.g., destroyer leader (DL)] with moderate traffic source parameter $a_1 = 0.1$, moderate bit rate/circuit, and moderate terminal EIRP and G/T. Thus, the satellite capacity** needed, c_1, is taken as unity for this class.

Class 2. Force Element Class

Typical of circuits off force element ships [e.g., destroyer (DD,DE)] with low traffic source parameter $a_2 = 0.01$ and low bit rate/circuit, but even lower G/T. Hence, capacity/circuit $c_2 = 4$.

Class 3. Major Flagship Class

Typical of circuits off major flagships (carrier or cruiser) with high traffic source parameter $a_3 = 0.5$, but, although high G/T, even higher bit rate, so $c_3 = 4$.

In a task force, one might expect overall user traffic would tend to correlate with the number of sources N_3 on a major flagship.*** Consequently, required satellite capacity for dedicated versus shared allocation was computed as a function of the number

---

*This model is exceptionally simple as it does not address source-sink pairs of circuits.

**Capacity is taken proportional to $(\text{bit rate})^{-1} \times \text{G/T}$.

***There may be more than one user circuit per flagship and/or more than one flagship per task force.

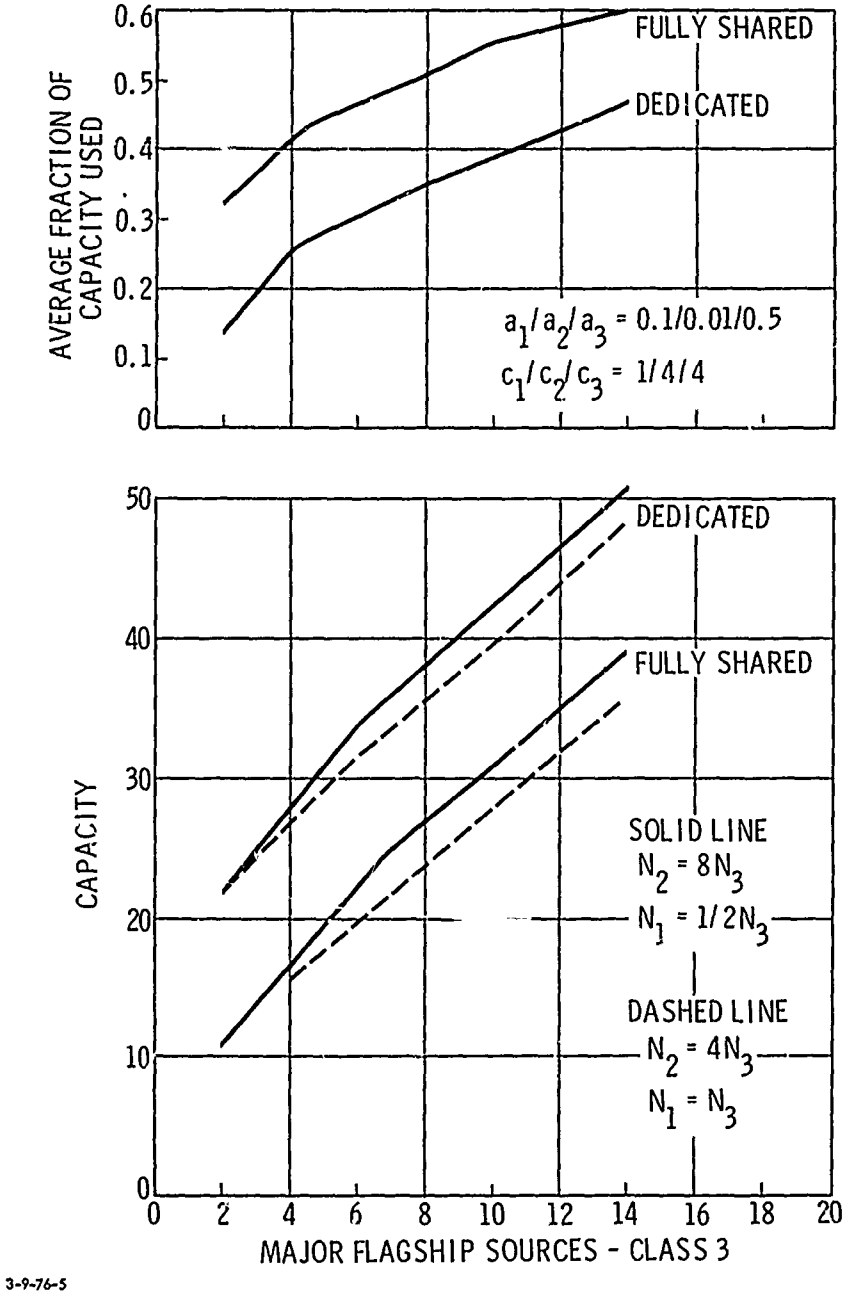


FIGURE II-3. Hypothetical Naval Task Force Example

of major flagship sources N_3 (rather than offered traffic, as in the previous figure). Two different mixes of N_1 and N_2 sources were used. Case 1 has eight times as many force element sources as major flagship sources and half as many leader sources as major flagship sources, $N_1 = N_3/2$, $N_2 = 8N_3$. In case 2, there are four times as many force element sources as major flagship sources, and there are equal numbers of leader sources and major flagship sources, $N_1 = N_3$, $N_2 = 4N_3$. For the example described, Fig. II-3 plots the required capacity as a function of the number of major flagship traffic sources for both cases of source mix. Note the small difference between the two mixes. The average fraction of capacity used (utilization) is shown only for Case 1; Case 2 is quite similar.

The results, such as Figs. II-2 and II-3, can be used to compare the levels of traffic that can be accommodated in given allocations of capacity. This is the converse problem to allocating a capacity to a given level of traffic, which can arise, for example, through loss in system capacity. For example, in Fig. II-2 with balanced traffic and 80 units of capacity available, 37 as opposed to 30 Erlangs can be accommodated with the fully shared strategy. On the other hand, with only 30 units of capacity, 11 versus 7 Erlangs of traffic can be accommodated, a significant improvement.

4. Conclusions

The results of the numerical examples indicate the following:

1. *Above moderate levels of offered traffic capacity,* needs increase linearly with offered traffic and at about the same rate, whether dedicated or fully shared allocation strategy is used. The capacity difference between the two is constant as traffic varies. Thus, the percentage increase in needed capacity at the higher level of offered traffic levels for the dedicated strategy tends to zero.

2. The numerical value of the constant difference in required capacity between dedicated and shared allocations increases with increased grade of service, (i.e., reduced probability of blocking).

3. *At low levels of traffic,* the percentage increase in capacity needed by the dedicated allocation can be as high as 50 percent.

4. Connectivity permitting, the effect on capacity utilization of balancing the user class parameters (traffic activity and per-circuit capacity draw) can be as large as or larger than the spread between dedicated and shared allocation strategies. The spread in needed capacity between balanced and equal source mixes in the example of Fig. II-2 is greater than between dedicated and shared allocation strategies. In this example, the rate of increase in needed capacity with offered traffic is 25 percent faster for the equal traffic user classes than for the balanced traffic case. Balanced network sources consume capacity at the rate of approximately 1.8 capacity units per Erlang of traffic, while equal network sources have a rate of approximately 2.4 capacity units per Erlang of traffic.

5. For total offered traffic in excess of 5 Erlangs, the Erlang and Engset models produce essentially the same results. Using the Erlang model as an approximation to the Engset model provides a conservative estimate for required capacity to meet a blocking objective.

With reference to the hypothetical Naval task force example, Fig. II-3 indicates that:

6. The results are qualitatively similar to those of the fixed and shared allocation schemes in the previous

example. At a moderate number of major flagship sources (8), the percentage capacity gain for fully shared is about 50 percent. This diminishes percentagewise with higher traffic levels.

7. The difference in required satellite capacity between the two cases of source population mix chosen is small.

The above points suggest the following conjecture:

8. When dealing with many user classes ($K > 3$), it may be possible, for the purpose of reducing the number of classes to be analyzed, to computationally treat as a single compound class those individual classes with a close balance in the product of traffic offered times capacity needed per circuit.

C. STORE-AND-FORWARD SYSTEMS

1. Introduction

For store-and-forward ("message" switched) systems, several techniques are investigated of time-sharing a satellite communication channel among many users with data messages. The users are organized into a net. The net is allocated a channel with sufficient capacity, represented by an equivalent line speed or bit rate, to handle the generated message traffic. The objective is to develop system performance measures by which a quantitative comparison of the techniques could be made.

The time-sharing techniques considered are:

1. Fixed Assignments--Each member of the net is assigned a specified transmission time that occurs periodically and that is his and only his to use. The users know when and for how long they can transmit.

2. Access with Polling--The net members are sequentially polled according to a predetermined order by the net controller, with the polled user transmitting until the message buffer is empty. At the end of transmission, the next member is polled and the process repeats.

3. Access with Reservations--At the time of a message arrival, the user requests transmission time for the message from the net controller. The controller allocates transmission time to the requesting users, and at the appropriate time the user transmits the message.

4. Contention or Random Access--At the message arrival, the user transmits all or part of the message on the available channel without coordination with the other users in the net. If two or more transmissions overlap, each message or parts of each message may be blocked, and the affected users try again.

Fixed assignments and polling are cyclic time-sharing techniques. The time behavior of the net operating under either fixed assignments or polling is exemplified by Fig. II-4. The transmission time allocations are subdivided into the time required to transfer control (walking time) from one user to another (t_w, considered overhead) and the time allocated for data transmission (t_i for the i^{th} user). The cycle length M is the time difference between consecutive transmissions by one user. In fixed assignments, the user data transmission times are predetermined and constant from cycle to cycle, and therefore the cycle length is also constant. The control transfer time consists of the preamble transmission time required to establish bit synchronization at the receiving terminal and the guard time between consecutive transmissions to ensure that the user transmissions do not overlap and interfere. Under access with polling, the control transfer time consists of the preamble transmission time and the single-hop propagation time to the satellite. The transmission time allocated to the users is not predetermined, but the polled user is allowed to transmit until the message buffer is empty. The cycle length is random because the user "allocations" are random.

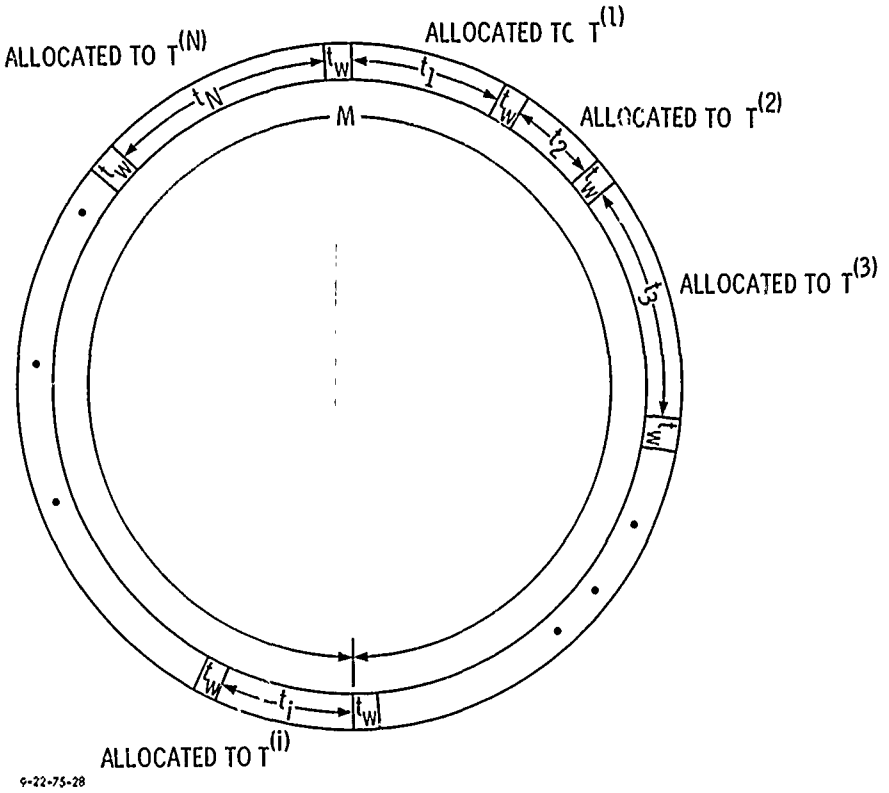


FIGURE II-4. Time Behavior of Net Operating Under Fixed Assignments or Polling

Under reservations and random access, the time sequence of user transmissions is not cyclic but is determined by the arrival (generation) time of messages. In reservations, coordination for user transmissions is provided by the net controller, who allocates transmission time to the requesting users. There is no coordination in random access, and therefore transmissions can overlap and interfere. The techniques are summarized in Table II-1, and some of the effects of traffic variations on the net operation are presented in Table II-2.

The methodology and the system performance measures used in evaluating these allocation techniques are based on queueing theory. The communication system is viewed as a service system, the messages being the customers, the channel being the server, and the transmission of data from one user to another being the

TABLE II-1. PROPERTIES OF THE SHARING TECHNIQUES

	Fixed Assignments	Polling	Reservations	Contention
User Transmission Sequence	Predetermined and cyclic.	Predetermined and cyclic.	Random, based upon message arrivals and queueing discipline.	Random, based upon message arrivals and blocking.
User Transmission Time Allocations	Assigned, constant, and dependent upon average message traffic.	Coordinated, variable, and dependent upon time to empty transmitting user's buffer.	Dynamically assigned, variable and dependent upon message length.	No coordination and no assignments.
Queue Discipline	Effected only at each user terminal and not over the net.		Effected over the net.	At each user terminal only.
Control Instructions Required	No dynamic instructions, allocations and sequence predetermined.	Transmission turn-on time.	Transmission turn-on time and duration.	None, but may require control for stability.

TABLE II-2. EFFECTS[a] OF TRAFFIC CHANGES ON NET OPERATION

	Fixed Assignments	Polling	Reservations	Contention
Changes in Average Traffic Load	Requires redetermination of the allocations because of the possible degradation of operating at a non-optimum point and possible saturation at a user.	None (System self-compensates)		None
User Entry or Departure	Requires alteration of the transmission sequence and possibly the user allocations.	Requires alteration of the transmission sequence.	None	None

[a]Exclusive of common effects (for example, either the capacity required to satisfy a performance objective or the performance measures for a given capacity will change).

service performed. There is a queue of data messages (buffer queue) that are waiting to be transmitted at each user's terminal. Two variables are used to quantify the effects of the allocation techniques on the user: (1) the length of the buffer queue and (2) the message queueing time (the waiting time in the buffer queue plus the message transmission time). The system performance measures used in evaluating the techniques are the numerical averages, over the users in the net, of (1) the expected value of the user buffer queue lengths and (2) the message queueing times at the user under steady-state conditions. The system saturates when the allocated capacity to a user in the net is insufficient to handle the average message traffic. Saturation causes both the user buffer contents and the message queueing times to increase without bound to infinity.

The system performance measures (the grades of service provided) determine the effects of the time-sharing techniques on the user, for a specified net size and capacity allocation. It still must be decided what capacity (average bit rate) will be allocated to the net. The allocated capacity affects both the channel utilization,* which impacts on the transmission facility cost, and the system performance measures. The system designer usually must make a tradeoff between the transmission facility costs and the grade of service provided to the users. The techniques are also compared in terms of the capacity required to provide a specified grade of service.

2. Results

Analytical expressions for the user buffer queue length and message delays are obtained for the time-sharing techniques under general conditions (compound Poisson distributed data

*Channel utilization is average arrival rate multiplied by the average transmission time of a message.

unit arrivals and identical user transmission capabilities), and therefore can be used within the stated assumptions for comparisons with other techniques. For polling, the message arrival rates and average message lengths are further assumed to be identical for all users in the net, while for reservations, the average message lengths are equal, with no further restriction on the arrival rates. At present, the analytical results either exist or could be derived only on the basis of these assumptions. For the implementation of reservations, the allocated net capacity is subdivided to provide a data channel plus an orderwire channel over which the time requests are transmitted using random access. The probability of blocking on the orderwire is assumed to be zero in Chapter IV, but this assumption is removed in Appendix E, where the orderwire traffic and output message process are assumed to be Poisson. The results presented here incorporate the results of Appendix E. The transmitted messages are assumed to be received without error to ensure that the effects on the buffer content and queueing times are attributable to the allocation schemes and are not caused by the channel. Errors in the data messages appear to have the same effect in all of the schemes, but errors which occur in the control messages associated with polling and reservations will have different effects and should be examined in future work.

Random-access or ALOHA-type systems were not investigated further (a limited discussion on the minimum net capacity required is presented in Section IV-F) because:

1. For Poisson data-unit arrivals, saturation in slotted ALOHA (random access starting at specified times only) occurs at a channel utilization of $1/e = 0.37$, while for fixed assignments and polling, saturation occurs at a utilization of 1. Therefore, a contention system could require up to three times as much capacity and does not seem attractive in situations where the transmission channel costs are high.

2. An uncontrolled random-access system may become unstable even with utilizations smaller than the saturation value.

3. The message throughput of an unstable, uncontrolled random-access system decreases with time and eventually becomes zero, whereas there still is message throughput in a saturated fixed-assignment or polling system.

To compare fixed assignments with the other two techniques, it is first necessary to determine the sizes of the transmission times assigned to the users. On the other hand, the user allocations in polling and reservations are dynamic and intrinsic in the technique. In a polling system, the polled user transmits for as long as required to empty the message buffer. In a reservation system, the users are given sufficient time to completely transmit each generated message.

The user allocations in fixed assignments are chosen to be those that minimize either the average buffer contents or the message delay* (queueing time minus message transmission time) over the users in the net. Unfortunately, the optimum user time allocations, which depend upon the net overhead, the message parameter values, and the transmission rate, could not be determined analytically and hence must be obtained numerically for specific examples. The net overhead (the total overhead in a cycle) consists of walking time (the preamble transmission time plus the guard time between consecutive transmissions) and the time reserved for the requests of new users to join

*Both the delay and queueing time are minimized at the same user allocations because the message transmission time is independent of the user allocation. The queueing time is used as a performance measure because, under reservations, the message transmission time is larger than that for polling or fixed assignments due to the capacity removed for the orderwire.

the net. As examples, the three values of overhead parameters presented in Table II-3 are considered with the user transmission and message parameters of Table II-4. The net in this example (Table II-4) is assumed to consist of ten small-volume users (class 1) and one large-volume user (class 2).

The behavior of the average buffer contents and message delay in the net for a fixed cycle length with overhead case I is demonstrated in Fig. II-5 for the user parameter values of Table II-4. It is seen that both performance measures are sensitive to the transmission allocations chosen. For the three overhead cases, the user allocations, which minimize the net message delay, and the resultant message delay values are presented in Table II-5, from which the effects of reducing overhead are observed. Reducing the overhead causes both the average message delay and the optimum user allocations to decrease, although the reductions in the delay are not as dramatic as the reductions in the user allocations and cycle lengths. The best performance (the smallest message delay) is obtained with zero overhead (case III). The delay reduction obtained from reducing the overhead from case II to III (removing the new entry request time) is larger than the reduction obtained from reducing the overhead from I to II (reducing the transmission overhead). The system can compensate for the transmission overhead by increasing the user allocations because the value of this overhead is independent of the allocation size. On the other hand, the size of the new entry time was chosen to be proportional to the cycle length, and therefore its effect on system performance is independent of user allocation size.

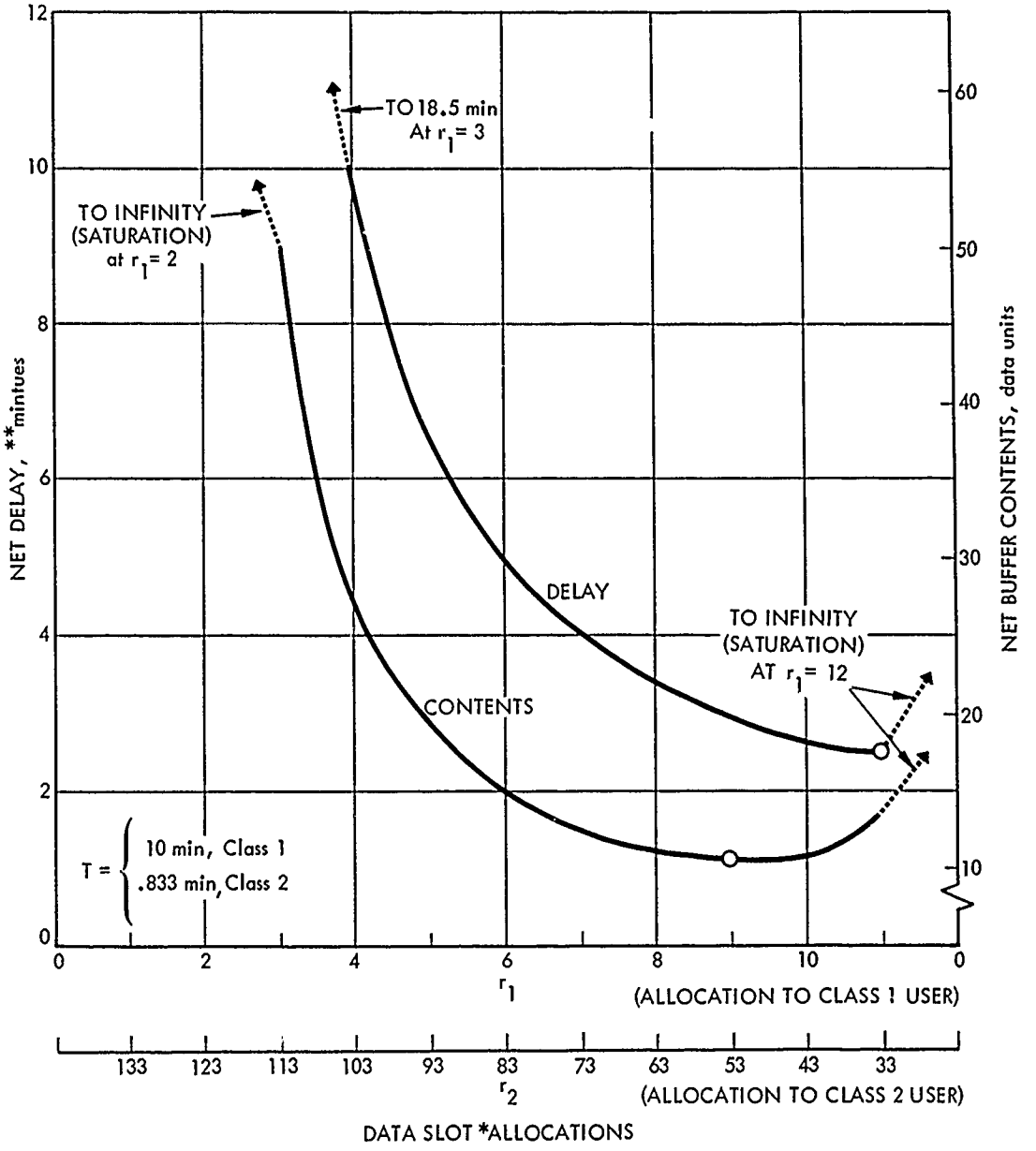
TABLE II-3. ASSUMED OVERHEAD VALUES

Case	Transmission Overhead[a]	New Entry Request Time
I	0.39 seconds[b]	Proportional $\left(\frac{15}{120}\right)$ to cycle length
II	0.0	
III	0.0	0.0

[a]Walking time (preamble transmission time plus guard time).
[b]Transmission rate R = 2.4 kb/s.

TABLE II-4. ASSUMED PARAMETER VALUES FOR SATELLITE COMMUNICATION NET

Transmission Rate (R):	2.4 kb/s
Number of Users (N):	10 users in class 1 1 user in class 2
Average Interarrival Times (T):	27 (10 minutes for class 1) or 2.25 (0.83 minutes for class 2)
Data Unit Length:	608 bits
Average Message Length ($\bar{m}$):	26 data units
Virtual Message Length (m):	26 data units



*A data slot in the transmission time of a data unit.

**The net queueing times are equal to the net delay values plus 0.11 minutes for this case.

3-31-76-49

FIGURE II-5. System Performance Measures with Fixed Assignments under a Cycle Length Constraint of 183 Slots for Overhead Case I

TABLE II-5. MINIMUM DELAY OF FIXED ASSIGNMENTS FOR THE OVERHEAD CASES

Overhead Case	Cycle Length, data-unit slots[a]	User Slot[a] Allocations		Net Delay, min
		Class 1	Class 2	
I	183	11	33	2.5
II	76	5	16	1.99
III	13	1	3	1.48

[a]A slot is the transmission time of a data unit.

Now that a limited discussion has been provided on the effects of the user allocation and overhead on the performance of a net using fixed assignments, a comparative evaluation of the techniques can be performed. Comparative results are given for the case of one user class (user message processes are identical), because only in this case are the analytical results valid for polling and for the system in which the net overhead is reduced. With these assumptions, the optimum user allocation in fixed assignments is one data slot to each user.

With one user class, the net queueing time is equal to the average queueing time for a user. A queueing-time comparison of the techniques for a sample net is presented in Table II-6. Recall that the net capacity allocation R is subdivided under reservations into R_1, the orderwire capacity, and R_2, the data channel capacity. The channel utilization ρ is equal to the average number of data-unit arrivals per time slot in the net divided by the number of data units that can be transmitted during a slot. The dynamic allocation schemes, polling and reservations, provide a grade of service that is approximately an order of magnitude better than that for fixed assignments. Figure II-6 presents the effects on queueing time of message arrival rate or channel utilization with a constant average message length. Polling and fixed assignments saturate at a utilization of one, while the net under reservations saturates at a reduced utilization (0.917) on the data channel because of the orderwire capacity. Finally, in Table II-7 a comparison of the techniques is presented with respect to the capacity that must be assigned to the net to achieve

TABLE II-6. SAMPLE[a] COMPARISON OF TECHNIQUES

	Queueing Time, min
Fixed Assignments[b]	3.00
Polling[c]	0.35
Reservations[b,d]	0.44

[a] One user class, N = 11, T = 2 min, m = 26 data units, R = 2.4 kb/s, and channel utilization = 0.6.

[b] Zero preamble.

[c] 96-bit preamble.

[d] R_1 = 200 b/s, and R_2 = 2.2 kb/s.

a grade-of-service (queueing-time) objective. Again, polling and reservations are preferred over fixed assignments and provide a capacity saving of at least 8 kb/s, enough to support three additional nets with the same parameters.

Because analytical expressions were obtained for the performance measures, the regions in the parameter space where one technique is preferred over another can be established for overhead case III. For this overhead case and one user class, the optimum allocations can be determined analytically (one data slot for each user) for fixed assignments. The analytical results of Chapter IV for reservations are used in the comparisons with polling and fixed assignments. Therefore, the comparisons only establish the regions where fixed assignments or polling are definitely preferred. Tables II-8 and II-9 respectively present the crossover walking time for polling (time to transfer control) and the critical turnaround time for reservations (the time to effect a capacity assignment) in the buffer-contents comparison for fixed assignments. The corresponding transmission rate is also presented. Fixed assignments yields smaller buffer contents and is therefore preferred if the transmission rate is larger than the listed rate. Polling is preferred over reservations when the turnaround time is larger than half the average time the channel is not available to a user in polling (half the average time between a user's transmissions).

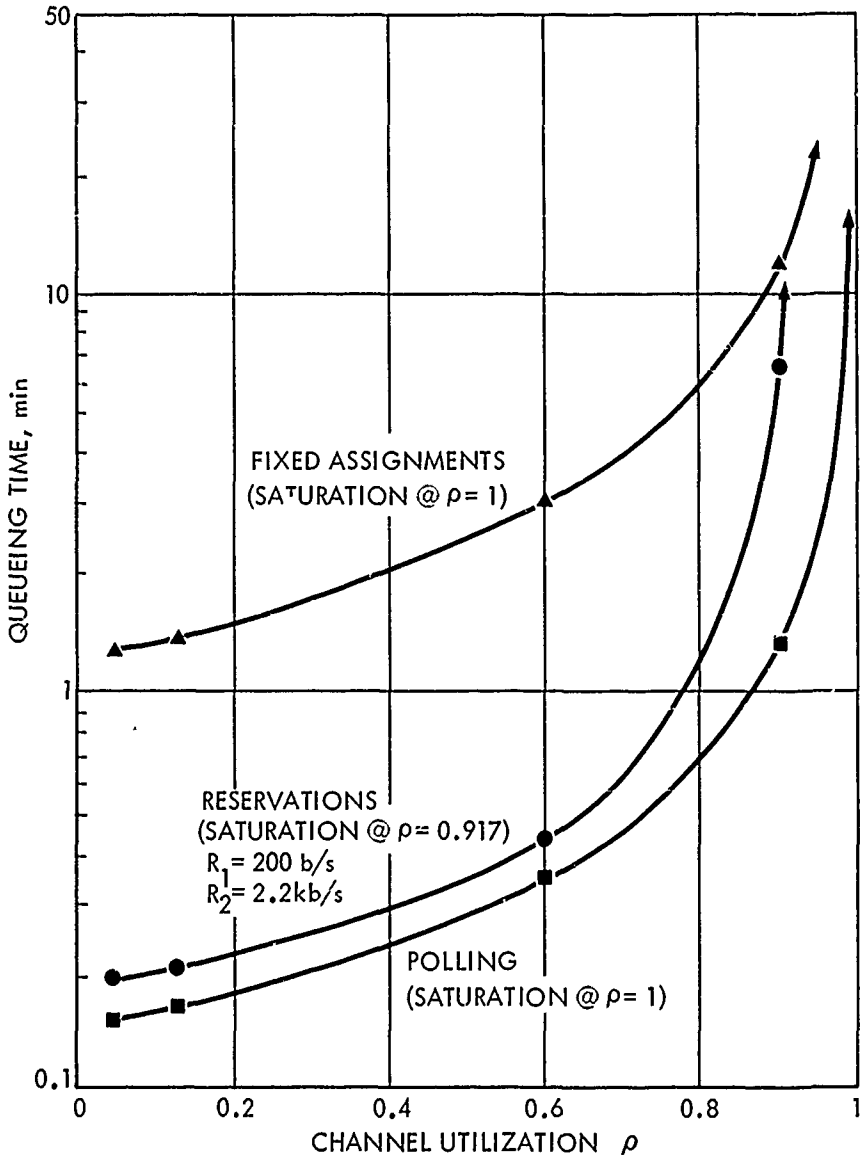


FIGURE II-6. Effect of Channel Utilization on Queueing Time (N = 11, R = 2.4 kb/s, and $\overline{m}$ = 26)

3-19-76-35

TABLE II-7. CAPACITY TO ACHIEVE A 0.3-MINUTE QUEUEING TIME WITH N = 11 AND T = 2 MINUTES

Technique	Transmission Rate, kb/s
Fixed Assignments	11
Polling	2.6
Reservations	3.0[a]

[a]A 2.4-kb/s data channel plus a 600-b/s orderwire.

TABLE II-8. CROSSOVER WALKING TIME AND CORRESPONDING TRANSMISSION RATE IN THE COMPARISON BETWEEN FIXED ASSIGNMENT AND POLLING WITH DISTRIBUTED CONTROL

Average Message Length, data units	N	Utilization	r_c, slots	Transmission Rate,[a] kb/s
5	11	0.1	9.2	20
		0.9	9.0	20
26	11	0.1	47.7	107
		0.9	50.6	114
26	100	0.1	51.4	115
		0.9	51.0	115

[a]Data unit = 608 bits; preamble length = 96 bits for polling.

TABLE II-9. CRITICAL TURNAROUND TIMES AND TRANSMISSION RATES IN THE COMPARISON BETWEEN FIXED ASSIGNMENT AND RESERVATIONS WITH CENTRALIZED CONTROL

Average Message Length, data units	N	Utilization	a, slots	Transmission Rate, b/s
5	11	0.1	55	61×10^3
		0.9	455	511×10^3
26	11	0.1	288	323×10^3
		0.9	2555	2.9×10^6
26	100	0.1	2855	3.2×10^6
		0.9	25,295	28.5×10^6

Finally, the effects are examined of combining nets and sharing the pooled capacity instead of dedicating a channel to each net (for example, C nets, each with a channel capacity R). There are two distinct ways of sharing the pooled capacity:

1. The users dynamically *share the channels* (C) by transmitting at the individual channel capacities (R) (C servers, each working at rate R with 1 queue).
2. By transmitting at the total capacity (CR) obtained by *merging channels* (1 server working at a rate CR with 1 queue).

It is well known in queueing theory that C servers--1 queue yields smaller queueing times than C separate "1 server--1 queue" systems and, further, that 1 server working at a rate CR yields smaller queueing times than C servers working at a rate R. In a satellite system, overhead (walking time in polling and order-wire capacity in reservations) are associated with the manner in which the techniques allocate capacity to the users and are not accounted for in the models used in queueing theory, and hence the results are not simply transferable from queueing theory to the store-and-forward system. Using reservations allows the combining of nets and the sharing of channels (C servers--1 queue) because a controller (common) buffer is essentially formed. On the other hand, sharing of channels is not possible under fixed assignments and polling because the net operation is centered around one channel. Merging of channels is possible under all of the considered techniques. Table II-10 presents the queueing times for an example for each technique of sharing the pooled capacity, with the results of Table II-6 reproduced as a baseline (one net--one channel). The results for reservations and fixed assignments agree with the previously mentioned queueing theory results, but under polling the merging of the channels produces an increase (of about twice) in the queueing time.

TABLE II-10. ARCHITECTURAL COMPARISON

	Queueing Time[a], min		
Number of Users	11	44	110
Transmission Rate	2.4 kb/s	2.4 kb/s	24 kb/s
Number of Channels	1	4	1
Fixed Assignments[b]	3.00	--	2.99
Polling[c]	0.35	--	0.66
Reservations[b]	0.44	0.25[d]	0.07[e]

[a]One user class, T = 2 minutes, channel utilization = 0.6, average and virtual message length = 26 data units.

[b]Zero preamble.

[c]96-bit preamble length.

[d]One channel is used for the orderwire.

[e]R_1 = 4 kb/s and R_2 = 20 kb/s.

Table II-11 gives the number of channels that must be provided to the combined net (N = 110) of Table II-10 to achieve specified queueing time objectives. Two objectives are used to demonstrate the sensitivity of the required capacity to the grade of service provided. For the 1-minute queueing time objective, the total capacity required with reservations cannot be reduced significantly, even though the resultant queueing time is substantially smaller than the objective, because the utilization on the data channel is about 0.7 already.

TABLE II-11. NUMBER OF CHANNELS REQUIRED TO ACHIEVE AVERAGE QUEUEING TIME CRITERIA WITH N = 110 AND T = 2 MINUTES

	Number of 24-kb/s Channels		
Technique	A[a]	B[b]	Queueing Time, min
Fixed Assignments	30	2	0.85
Polling	10	1	0.66
Reservations	2[c]	1[d]	0.07

[a]A: 0.04 minute average queueing time.

[b]B: 1 minute queueing time.

[c]One data channel plus an orderwire channel.

[d]Data channel capacity = 20 kb/s plus a 4-kb/s orderwire.

3. Conclusions

In a net operating under fixed assignments, the user allocations must be predetermined and should be selected carefully because both system performance measures are sensitive to the allocations. In the examples considered where more than one user class exists, the optimum allocations are not only different from the allocations obtained by an intuitive approach but also reduce the resultant delay by about 50 percent below that obtained by using intuitive allocations. In either polling or reservation systems, the user allocations are dynamic and intrinsic to the technique and therefore require no optimization. On the other hand, the (minimum) capacity allocation to the net in any technique must be determined to ensure stable operation. This requires an estimate of the expected average message traffic load in the net.

Overhead can be characterized as of two types. One type, represented by synchronization preambles and transmission guard times, is independent of allocated transmission times. The other type, represented by the time reserved for new entry requests in the cyclic techniques and the orderwire capacity in reservations, is dependent upon allocated transmission times. Both types of overhead reduce the effective data transmission rate and hence reduce the grade of service provided; the second type also reduces the message traffic load, resulting in saturation. In the cyclic techniques, the inefficiencies induced by allocation-independent overhead can be reduced by allocating longer transmission times, and the effect on the saturation level can be made small (actually, there is no saturation effect in polling). In reservations, because the allocated transmission times are determined by message lengths, the effect of the preamble overhead is not reduced but increases the channel utilization and also reduces the saturation level. The reservation technique is more sensitive to this overhead than the cyclic techniques.

In the sample comparisons for a fixed capacity allocation to the net, polling and reservations both show significant improvements over fixed assignments in the grade of service provided. The performance measures are shown to increase with utilization, and at higher levels of utilization they are tremendously sensitive to small changes in utilization. This behavior is typical of a queueing system and should be taken into account in the choice of an operating point (in channel utilization or in capacity allocation to the net).

The reservation technique increases flexibility and further improves the grade of service provided because it permits the pooling of users and the sharing of available channels among pooled users when the users cannot transmit at a higher rate than the individual channel rate. Further, if the users can transmit at the total capacity of the channels (sharing the merged channels), the grade of service is improved significantly only in reservations, while it degrades in polling.

As would be expected from the above results, polling and reservations in the examples considered require significantly less capacity than fixed assignments to achieve a grade of service.

Finally, the general conditions under which the techniques are preferred with respect to buffer-contents performance measures (i.e., the conditions under which the techniques yield lower values) are obtained by analytical comparisons of the respective buffer-contents equations. In comparing polling and fixed assignments, which are both cyclic time-sharing techniques, smaller buffer contents are obtained by using polling, except when the time to transfer control (walking time) from one user to another in polling is about twice the average message transmission time. The only parameters that determine the preferred regions are the transmission rate and the average message length. On the other hand, in comparisons of cyclic techniques with reservations, all of the parameters--number of

users, channel utilization, average message length (except in polling), and transmission rate--determine the preferred regions. (The orderwire capacity required for reservations is not included in the comparisons.) Polling is preferred over reservations for those systems where the time to effect a capacity assignment with reservations is larger than half of the total average time the channel is not available to the user in polling. Fixed assignments is preferred over reservations for those systems where the time to effect a capacity assignment is larger than the product of the number of users and the average message transmission time divided by one minus the channel utilization. The cyclic techniques are at a disadvantage with respect to reservations for large numbers of users. The cyclic techniques are definitely preferred over reservations when the number of users in the net and the average message transmission time are small.

D. GENERAL CONCLUSIONS

An analytical appraoch is shown to exist which relates user traffic to satellite capacity needs with regard to a variety of system access, sharing, and transmission strategies. The ability to optimize a particular strategy and then compare it to other strategies is a direct consequence. In general, no one strategy or sharing arrangement is uniformly best. Traffic levels and various operational constraints can favor one sharing method over another, depending upon the operating conditions existing in the system.

It is important to be able to delineate the key descriptors of the system and the usage factors together with the proper quantitative criteria for determining when one mode is superior to another and by how much. In sample calculations for digital circuit communication service, it is shown that, above a certain minimal level of traffic in each net of a group of nets, there is no significant penalty in the overall capacity required if one chooses to partition capacity to each of the nets individually

along dedicated lines, as opposed to amalgamating all the users into a super net and fully sharing all the available capacity. This result holds only if technical means are implemented to ensure a flexible allocation in time of capacity to each net in order to match the level of traffic existent in the nets. Otherwise, very serious losses in capacity utilization can occur.

On the other hand, the fully shared approach reduces the need to individually balance each net's capacity need against its traffic level. However, in this situation the grade of service provided to each net becomes interdependent among the nets, as opposed to net independent, when capacity is dedicated.

For store-and-forward data transmission services, the more time-dynamic strategies (such as polling or reservations) for assigning transmission times to user terminals are usually more capacity efficient than the less dynamic fixed assignments technique. They provide savings in required capacity to achieve a specified delay in message delivery time for a given level of message traffic. Further, the reservation technique facilitates a pooling of the users for making greater use of the available capacity. However, the fixed assignment method is more attractive for those systems where the transmission bit rates (speed) coupled with the long satellite path delays are such that the time taken to effect assignment of transmission time to terminals is large compared to the average time to transmit a message.

In addition to predicting capacity utilization and comparing sharing strategies, analytical results also have collateral benefits with regard to system design and control. They indicate network operational sensitivity to variations in traffic levels and establish stability conditions. They can also provide alternative measures for assessing the network status needed for system control. For example, monitoring delay or blocking in order to estimate operating traffic levels may be easier to implement than direct measurement of traffic activity at each terminal in the system. Finally, simulations

of systems are always benefited by a theoretical base that not only provides interpretive insight into the results of a simulation but also aids in the "design" of the simulation itself.

The results presented in this report are based on analytical models containing many simplifying assumptions. Many areas of theoretical and computational interest still remain to be investigated for practical systems. This chapter ends by identifying some study areas of further interest.

E. AREAS FOR FURTHER RESEARCH

Identified below are areas for further research in traffic and resource utilization relevant to DoD satcom programs. They are listed in no priority of order.

General

- Nonlinear capacity usage and multiple access systems
- Priority/preempt effects
- Control transmission system interrelationships with allocation
- More complex user-to-terminal-to-satellite interfaces
- Hybrid systems; store-and-forward (packet) transmission intermingled with data circuits.

Traffic Models

- Collection and processing of real traffic statistics
- Correlated traffic sources--overlapping user nets
- Nonexponential-type traffic statistics.

Systems

- Blocked Calls Held
- Call signaling and control
- Alternative store-and-forward techniques
- Interconnected systems--terrestrial/satcom.

Numerical Techniques/Computer Science

- Computational algorithms
- Simulation technology.

III. SATELLITE CAPACITY ALLOCATION FOR CIRCUIT SWITCHED SYSTEMS

A. INTRODUCTION

1. Conceptual Framework

In this chapter consideration is given to developing a mathematical model to evaluate certain nodal capacity allocation strategies for supporting communication traffic of the circuit type. In Chapter IV, store-and-forward message traffic will be addressed. Interest here centers on the problem of organizing circuit-type connections and allotting capacity for supporting circuits between sources and sinks connected to terrestrial terminals of differing capabilities (e.g., antenna size, power, information bandwidth) through a common communication satellite relay.

Shown in Fig. III-1 is a typical satcom configuration of terminals and their (base-band) traffic sources/sinks interconnected through and sharing the capacity of the satellite. In Fig. III-1 the traffic sources/sinks are represented by circles of differing size to emphasize their potentially differing information bandwidth requirements (e.g., bit rate). The terrestrial terminals are shown as having different-sized antennas to represent the various receiver sensitivities, transmitter powers, and (if appropriate) RF bandwidths.

The terminals each have a single-node interconnect through the satellite relay utilizing suitable signal modulation on

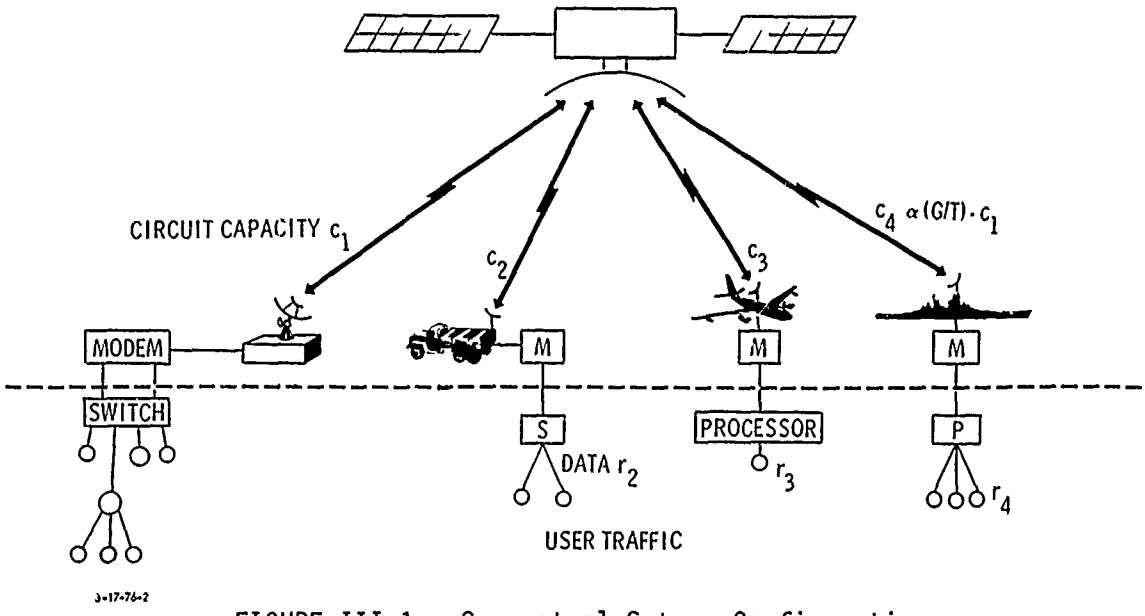
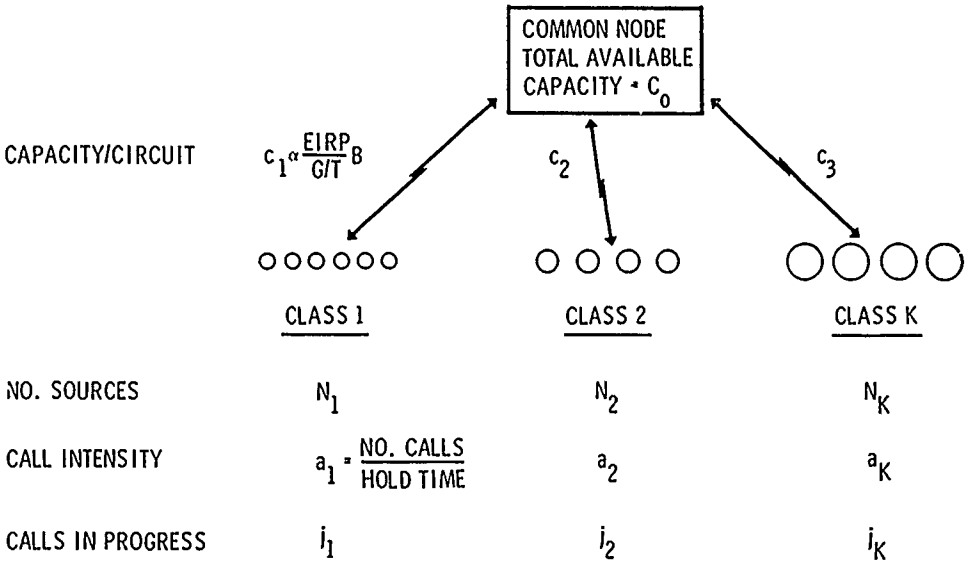


FIGURE III-1. Conceptual Satcom Configuration

their radio carriers to enable the desired link interconnections to be made. The composite problem of signal modulation, spacecraft transponder configuration, and RF system operation has been commonly referred to as satellite multiple access,* with three generic classes of multiple-access modulation identified as frequency-division multiple access (FDMA), time-division multiple access (TDMA), and spread-spectrum multiple access (SSMA). The multiple-access technique chosen defines and partitions available satellite capacity and provides constraints on sharing this capacity among the terminals.

Conventionally, multiple access has not addressed the issues of interconnection between data sources and sinks that access a terminal. These issues, designated as base-band multiplexing, to date have received less attention. Especially lacking has been an overall systems understanding of the interrelation between base-band multiplexing anf RF multiple access.

*For an introduction to this subject, the interested reader is referred to Ref. 10. Further detail is provided in IDA Study S-268 (Ref. 11).



3-17-76-4

FIGURE III-2. Conceptual Model

For the purposes of developing an initial mathematical model, the traffic source, terminal, and satellite relay configuration are idealized as shown in Fig. III-2. Each potential traffic source has access to a common node of total capacity C_o. The sources are aggregated into classes defined by parameters of population traffic intensity (source activity) and capacity required to support a circuit. A number of key simplifications are made in the abstract model here proposed. Outstanding among these is limiting the analysis to linear subdivision of available capacity and the "Blocked Calls Cleared" mode of operation. Less limiting operating modes can be easily implemented, but their analysis becomes exceedingly complex. Further discussion of the concepts and limitations imposed are presented in Section III-A-4 following a description in Section III-A-3 of the model used.

The elementary form of the model used allowed a meaningful and representative exploration of the problem which was tractable enough to be dealt with within the time and resources available. In this regard, the approach parallels in spirit the early phases of congestion theory developed for telephonic traffic. The discussion of the limitations in the model in Section III-A-4 is intended to stimulate interest in these problems and to encourage research activity directed towards building a body of theory extending beyond the simple model examined here.

It must be emphasized that there are important considerations that influence network aggregation other than the theoretical modeling studied here. Any decision to aggregate users must be made with cognizance of and in response to such factors (aside from cost) as:

- Chain of command/mission context
- Addressing/interconnection
- Interoperability/standardization
- Control responsibilities/accountability
- Security.

Perhaps the key point to note is that there is no a priori reason that networks cannot be made flexible enough to be reorganizable in time, according to need. Limitations and constraints have to be studied, and management mechanisms created. Central to these issues are the commonality and standardization of components and a concept of network hierarchy and control. Independent of specific network organization(s), there is the need to develop analytical methods to judge in *quantitative terms* the interrelationship of traffic loads and aggregations (network source/sink configurations) with "system" capacity allocation and control (network transmission configuration). It is the purpose here to demonstrate by examples the feasibility and desirability of acquiring such analytical tools.

2. General Example

As part of an ongoing TDMA study performed by Comsat Laboratories (Ref. 12) for the United States Army Satellite Communication Agency, the number of circuits necessary to support various levels of traffic associated with specified networks was estimated in order to size terminal equipment requirements (e.g., "channel units") which interface terrestrial communication lines. Data were taken from Ref. 12 and plotted in Fig. III-3. Figure III-3 is a conventional "Erlang B" curve and shows the number of available circuits needed to support a level of traffic offered, measured in Erlangs,* for a 0.05 probability of call blocking (a grade of service of 5 percent). Also shown is the average fraction of capacity used (circuit utilization), the average number of circuits in use divided by the total number of circuits available. From specified network requirements developed by the TRITAC office of the Signal Corps for DoD, Comsat Laboratories derived the associated traffic levels that are delineated on the abscissa. The right-hand ordinate shows the circuit capacity required for a given Erlang traffic volume, and the left-hand ordinate shows the circuit utilization.

The results from Ref. 12 utilize classic Erlang theory and are predicated on all circuits' requiring the same per-circuit capacity (e.g., identical earth terminals with identical bit rate) and having the same traffic intensity parameters. Thus, through link budgets the number of circuits needed can be translated into the satellite capacity necessary to support each

---

*An Erlang is a dimensionless unit measuring the level of offered traffic. For the infinite source model ($N_i = \infty$) it is defined as the average number of new call requests occurring during an average call holding time. [One Erlang unit is equal to 3600 call-seconds per hour.] Clearly, the number of available circuits must exceed the offered traffic to obtain a blocking probability less than one.

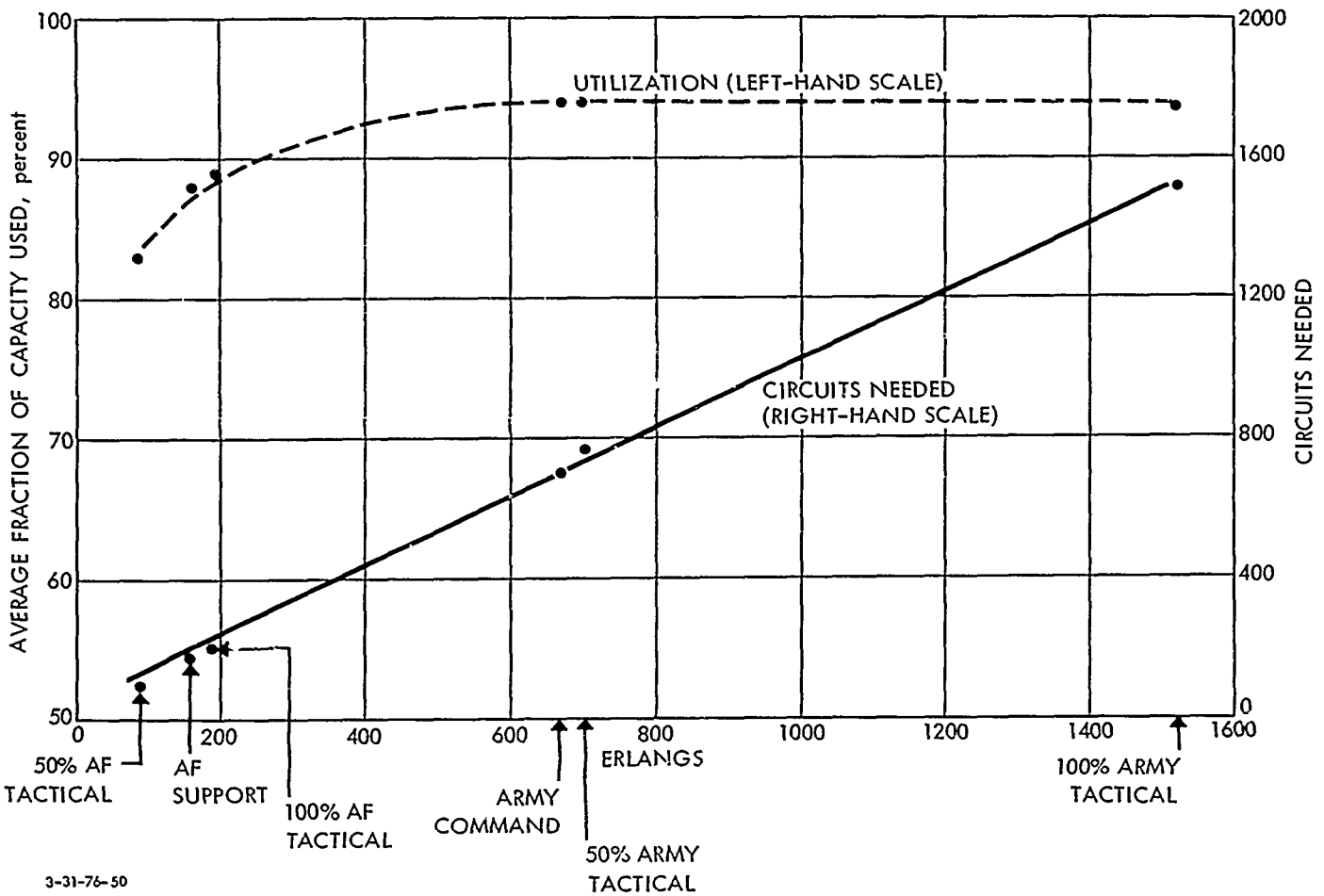


FIGURE III-3. TRITAC Circuit Requirements Data as Reduced by Comsat Laboratories. "Blocked Calls Cleared," Conventional Erlang B Formula, GOS = 0.05. (Data Source: Ref. 12)

network. An important issue is whether the traffic sources associated with each network should be pooled together in a very large common network (with common control) or whether satellite capacity should be flexibly subdivided and allocated to support each network individually (with hierarchical control). For the example studied in Ref. 12, Fig. III-3 shows that, from the point of view of efficient utilization of satellite capacity, there is no strong incentive to aggregate the networks provided capacity is flexibly matched to network needs. There is at most a 13 percent gain in capacity efficiency in going from 50 Erlangs of traffic (50 percent of the Air Force tactical network requirement) to 700 Erlangs (50 percent of Army tactical). Note, however, that a 50-circuit satellite network is a relatively large bundle of circuits. There will be advantages to network aggregation of low traffic loads.

3. Basic Mathematical Model

The (traffic) source population is subdivided into K (finite) classes of statistically independent traffic sources. When a source makes a call, he requests service via an order-wire, or calling channel, and is allotted a portion of nodal (i.e., satellite) capacity necessary to support his call if such capacity is available. If no capacity is available, the source is blocked and is modeled to return to a status "statistically" identical to not having attempted the call.* That is to say, the effect (frustration) of being blocked is not modeled, and there is an immediate call reattempt. Further, all users within a class are assumed to be equivalent. User needs for service, the likelihood of calling, and the capacity used per active user are identical for all users in the same class.

*This model is referred to as "Blocked Calls Cleared." It is analogous to a queueing model with no buffer or waiting room.

Circuits are assigned on a first come, first served basis. A discussion of such models is given in Ref. 13. Reference 14 provides a fundamental probabilistic framework.

Each class has the following characteristics, (for class $i = 1, 2, \ldots K$):

Traffic Model (Birth-Death)

- Exponentially distributed interarrival time with parameter λ_i (λ_i = average number of call requests/second from a source in class i).
- Exponentially distributed call holding time with parameter u_i ($1/u_i$ = average call duration from class i).
- Number of total sources N_i in class i, N_i may be infinite,* in which case $\lambda_i \to 0$, so that $N_i \lambda_i \to \tilde{\lambda}_i$.

It is assumed that any source with a call in progress uses nodal capacity as follows:

Capacity Model (Linear)

- Any source of class i uses c_i units of capacity from the common node (i.e., satellite).
- Capacity is measured in units normalized to class 1 so that $c_i = 1$. The classes are ordered so that $1 \leq c_2 \leq c_3 \leq \ldots \leq c_K$.
- Capacity is used linearly. That is to say, if there are $0 \leq j_i \leq N_i$ active sources from each user class $i = 1, 2, \ldots K$, the capacity in use is $j_1 + c_2 j_2 + \ldots c_K j_K \leq C_o$.

*When N_i is finite, the source arrivals are said to be quasi-random. This model was studied by Engset in 1918. With infinite N_i, the source arrivals are Poisson. This model was used by Erlang in 1917.

The number of active calls in progress, j_i, from each of the K traffic classes specifies a state, j, of the "system." Mathematically the state j is a K-dimensional vector with coordinates having nonnegative integer values, $j \triangleq (j_1, j_2, j_3 \ldots j_K)$. Due to the statistical nature of the calls, interest centers on finding the probability of occurrence $P(j)$ for each state j--that is to say, the steady-state probability of there being in progress j_1 calls from class 1, j_2 from class 2, and so on, up to j_K calls from class K.

Given that certain general conditions are met (Ref. 14), a steady-state probability distribution* $P(j)$ exists that is time independent. $P(j)$ is the solution of a K-dimensional difference equation, known as the equation of state. The mathematical problem is to formulate A-sets of allowable** states j (access strategies), determine the associated equation of state, and solve for $P(j)$. From knowledge of $P(j)$, system performance measures such as blocking probability and average utilization can be calculated.

Within the above framework, the mathematical model can be characterized as a K-dimensional Birth-Death process of quasi-random (N_i finite) or Poisson (N_i infinite) type. The Birth-Death nomenclature implies that (in probability) only one call at a time can arrive (Birth) or terminate (Death). The quasi-random or Poisson process refers to the exponential nature of the probability distribution for call arrival and holding time, predicated upon the mutual independence of all sources and pure-chance origination and termination of calls. An elegant,

---

*$P(j)$ is a (scalar-valued) probability function over the state vectors j in a K-dimensional space.

**For example, at the very least, every allowable set of states must satisfy the two conditions, (1) for each $i = 1, 2, \ldots K$, $0 \leq j_i \leq N_i$ and (2) $j_1 + c_2 j_2 + \ldots c_K j_K \leq C_o$. Additional constraints can be imposed to further restrict the set of allowable states.

rigorous, and mathematically advanced development of the multi-dimensional stochastic theory for the traffic model used here is provided in Refs. 15 and 16. A more heuristic approach, the one used here, is provided in Refs. 13 and 17.

4. Discussion

One of the key postulates in the elementary model used is that capacity used from the common node is the *linear* sum of the circuit capacity needed to support each of the source sink pairs with a "call in progress."

The abstract approach used here is not tied to any particular method of multiple access. Thus, to a first approximation, they are equally valid for FDMA, TDMA, and SSMA techniques. Whichever multiple-access technique is used, it is modeled as providing a pool of capacity that can be linearly subdivided and assigned to terminals individually or to nets of terminals, which, in turn, could further share their allotted capacity. On a per-user basis, the terminals and/or nets take different amounts of capacity, depending on terminal sensitivity, power, and bandwidth. Strategies for sharing out the capacity and methods for evaluation are then examined in the context of this abstract, simplified (e.g., linear) model of a sharing mechanization.

Ideally, the capacity drawn from a satellite by a terminal with a receive sensitivity G/T,* for a downlink with information bandwidth B, would be proportional to $B(G/T)^{-1}$. The maximum capacity available would depend on the satellite power, the largest terminal G/T value in the system, and the available RF bandwidth. In principle, TDMA comes closest to realizing this

*G/T is the ratio of terminal antenna gain to receiver noise temperature.

ideal. Capacity allotment is controlled by the amount of time allotted to each accessing terminal in proportion to $B(G/T)^{-1}$.

SSMA introduces the uplink impact on transponder power-sharing among active carriers. As long as the sum of the power of all active uplink carriers remains constant, capacity will divide linearly. Should the total carrier power arriving at the satellite depend on active traffic, capacity might not subdivide linearly but algebraically according to the ratio of desired to total uplink transmitter power. In this case, however, the system could be linearized to the "worst-case" maximum total uplink power. Interestingly enough, should one carrier (or, more likely, a jammer of constant power) dominate the uplink, the residual capacity left to the remaining uplink carriers would again divide linearly, independent of the summed power of the desired carriers.

For FDMA, capacity division is not only dependent on the nature of the uplink power levels but also on the intermodulation product spectrum generated. Should a radically uneven intermodulation spectrum result, capacity division would be not only algebraically nonlinear (which can be linearized) but also frequency dependent.

Thus, extension of the linear capacity division studied here to nonlinear capacity division, as determined by multiple-access modulation techniques, is an important avenue for future investigation. Certainly, for TDMA and linearized versions of SSMA and FDMA, the models used here apply. Moreover, it is believed that the models can be generalized to at least include algebraically nonlinear capacity division.

For the purpose of the analytical methods used, it is assumed that the interface between the terminal-associated baseband digital system (e.g., teletype, printers, secure voice codes) and the terminal modem(s) can be such that the satellite transmission system (from the transmitting terminal through the

satellite transponder to the receiving terminal) can be made transparent to the operation of the users' base-band digital system. This assumption represents a desirable practical goal. Space-related technology and user digital-device (i.e., computer) technology are independently evolving, each at its own rapid rate. It is desirable in satcom system development and application to preserve flexibility between satellite transmission technology and user data-system technology.

Postulation of linear use of capacity avoids the complex relationships (not yet fully understood) between terminal multiple access and base-band multiplexing of sources. Efforts to avoid dependence on specific system design for quantifying this interaction, together with a desire for mathematical tractability, lead to the convenient theoretical artifice of relegating practical complexities to the category of engineering problems associated with implementation.* A major need in satcom system engineering is to develop a better quantitative understanding of the relationship between the multiple-access transmission and the user-multiplexing interface. A promising direction for extension of the theory discussed here is towards inclusion of nonlinear capacity dependence on the number of circuits in use.

A major limitation placed upon the model was its restriction to a first-come, first-served and "Blocked Calls Cleared" mode of operation. Thus, call priorities, preempts, and call request queueing are not addressed. It is not necessarily difficult to implement more elaborate operating modes in handling call requests that cannot be immediately assigned satellite

*This dodge is not altogether vacuous. There do exist multiple-access base-band interconnections (such as TDMA with digital data circuit sources) which should lead to linear capacity use.

capacity. Except for the case of dedicated capacity allocations to each user class, the mathematical problems of describing and analyzing the performance of the call request strategy or the protocol for managing the queues of held calls become exceptionally difficult.

The difficulty in analyzing more sophisticated call request strategies centers on the precise mathematical statement of the equation of state and its solution. The equations must specify how calls of different user classes are queued, how queues are managed, how calls can be preempted, how waiting calls are given priority, and how capacity freed by termination of a call is allotted to calls in queue. For example, when the system is saturated, capacity from a call termination can be allotted only to calls in queue from user classes requiring less per-circuit capacity than the departing call. For those calls awaiting service that do require less per-circuit capacity than the departing call, any number of rules could be written for the allocation of the capacity freed by the departing call. This is discussed further in Section III-C.

Circuit connections between user classes are not considered. It is tacitly assumed that this problem can be avoided in the definition of traffic source/sink pairs by defining more user classes, that is, through the creation of additional classes of sources. This may not adequately model the situation where traffic sources might simultaneously belong to several user classes. More generally, source activity between classes could be correlated. In the simple model analyzed, it is assumed that all traffic sources are uncorrelated. This assumption is least accurate when the source population N_i is small.

Another simplification is the assumption that the sole limitation on circuit availability is satellite capacity. A terminal serving many sources can potentially run out of capacity through shortage of either source transmission handling equipments

(e.g., channel units) or bandwidth in the multiple-access carrier. Referring back to Fig. III-1, it is seen that in effect blocking is at least a two-step affair. A traffic call can be blocked on entering a terminal or at the satellite. The conventional method for dealing with this complication is to note that overall blocking must be a concatenation of the blocking events at the terminal and the satellite. If the probability of blocking of the terminal is made an order of magnitude smaller than that of the satellite, it is usually true that overall blocking* is then dominated by available satellite capacity.

Finally, a whole set of simplifications is implied concerning network control (satellite and terminals), circuit signaling/monitoring (call setup, take-down, busy signal), and switching (between sources and between terminals). All of these involve very important mechanizations of system control. They include the important areas of data formats, control protocols, terminal interfaces, and modulation parameters, as well as assignment of control signaling channels and nodal network management responsibilities (e.g., master/slave stations and satellite terminal and terrestrial switch interconnect). An especially important problem is the mechanization of a call request circuit, orderwire, reserved channel, and in-band signal. The transmission and processing of an asynchronous, randomly arriving call request signal can differ considerably from supporting an established call in progress.

It will be important in the future to develop more refined models that will incorporate the complexities discussed above. By historical analogy, this would parallel the theoretical

*A rule of thumb for simplified calculation treats terminal blocking and satellite blocking as independent events. Although, strictly speaking, these events are not independent, treating them as such gives an estimate of blocking probability.

development of classical telephonic traffic theory. For example, in 1917-1918 local lines accessing a central telephone office to outbound long-haul trunks did not have full access to all available trunks due to the prohibitive cost of providing the number of mechanical relays (selectors) necessary to provide full trunk access to each local line. Erlang's fundamental paper in 1917 (and Engset's in 1918) assumed an idealized connection system with full capacity availability.

The limitation of trunk availability to local circuits is called a "grading." Ironically, preceding Erlang's work, the first grading patent responding to the practical problem was awarded to E.A. Gray of the United States in 1908, U.S. Patent Specification 1002388.* Subsequently, analytical work on grading complexity was done by Erlang in 1920 and by E.C. Molina in 1921. Further effort continued at a dwindling rate until about 1931 with R.I. Wilkinson's paper. Following this, there was a hiatus of any interest until the late 1940s, when interest was rekindled in the Scandinavian countries. With the advent of electronic switching and advanced computational capabilities, there has been an accelerating interest in exploring the complexities of the theory (Ref. 18). One hopes that theoretical interest in capacity connection and switching theory will receive more rapid development for satellite communication than it did for telephony.

5. Presentation of Results

The results of the study of capacity allocation for c cuit service are presented as follows. Section III-B pre a summary of the general theoretical development. Section discusses the application of the theory to the dedicated and

*Reference 16 provides a concise mathematical review of this history.

fully shared strategies. Section III-D presents some numerical results. The principal mathematical development of the material is presented in Appendix C, together with a description of the numerical techniques used. In addition, Appendix C provides a short discussion of previous and current work on traffic theory, as well as related activities in computer science.

B. ANALYTICAL RESULTS

1. A-Set Geometry

Appendix C presents a careful mathematical review of the blocked-calls-cleared state equations, boundary equations and solutions, and several important theoretical properties. The previous section presented the basic traffic and capacity models. The parameters of the problem are the total available satellite capacity C_o and, for each user class i, the per-circuit capacity c_i required by a traffic source and the traffic parameters* $a_i \triangleq \lambda_i/\mu_i$ and N_i.

Interest centers on determining and solving the equation of state for the scalar-valued steady-steady probability $P(j)$ of a state vector $j \equiv (j_1, j_2, \ldots j_K)$, each of whose nonnegative integer-valued coordinates represents the number of calls in progress from user class i. The collection of allowable states, j, forms a set A which specifies an access strategy. The two A-sets of special interest are "dedicated capacity"** and "fully shared capacity." For any A-set, surely the capacity

---

*Recall that λ_i is the call request rate from a single source, N_i is the source population, and $1/\mu_i$ is the average call duration. Appendix C shows that only the ratio $\lambda_i/\mu_i = a_i$ is important to the solution of the problem.

**Although the term "dedicated" will be used here, there still is "demand access" operation, but it is limited to operate only within the capacity dedicated to a user class.

in use, $c_1j_1 + c_2j_2 + \ldots c_Kj_K$, must not exceed C_o. If a new call request were to arrive such that the capacity in use would exceed C_o, then it could not be accepted and that call would be blocked. Consequently, any A-set must be contained in the set Ω, defined as the collection of state vectors, j, which satisfy the inequality

$$c_1j_1 + c_2j_2 + \ldots c_Kj_K \leq C_o.$$

The set Ω is itself an A-set, the "largest" such, and represents the fully shared strategy.

Additional constraints can be imposed on the manner in which states are allowed to occur. Partition the available capacity C_o into K separate capacity pools, C_1, C_2, ... C_K, such that $C_1 + C_2 + \ldots C_K = C_o$. Then assign each pool of capacity to its correspondingly numbered user class for dedicated use by it and no other class. This can be represented as an allowable class of states A by j vectors whose j_i coordinates satisfy the following inequalities:

$$c_1j_1 \leq C_1$$

$$c_2j_2 \leq C_2$$

$$c_Kj_K \leq C_K .$$

This A-set represents the dedicated strategy.

In every case, the A-set can be represented as a geometrical object in K dimensions. This provides important theoretical insight. With reference to Fig. III-4, the fully shared A-set (i.e., Ω) is the volume enclosed between the positive coordinate

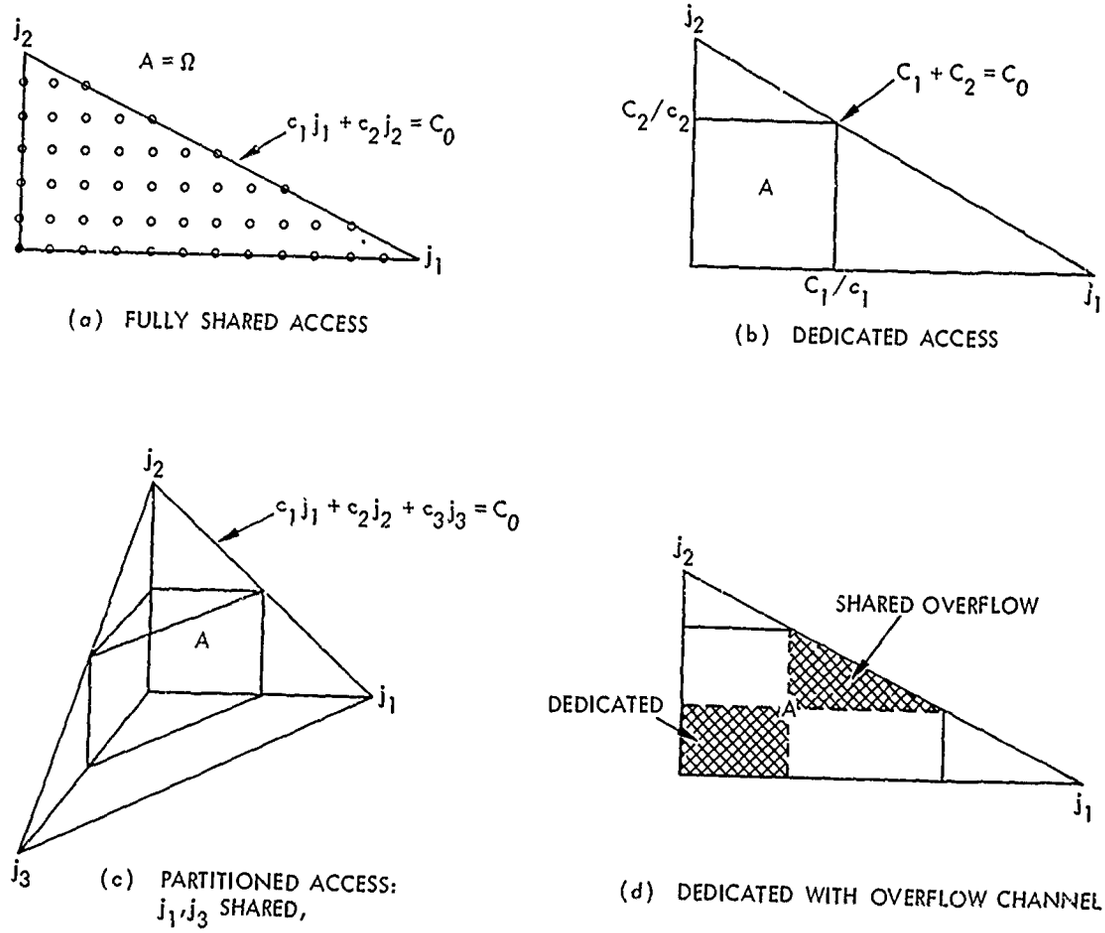


FIGURE III-4. Access Strategies and A-Sets

axes and a tilted plane* offset from the origin by a distance proportional to the total capacity C_o. (Thus, more available capacity permits more allowable states.) The dedicated A-set is a K-dimensional rectangular box resting on the positive coordinate axes with dimensions $C_1 \times C_2 \times \ldots C_K$. Note that the rectangular box of the dedicated A-set fits inside the fully shared volume Ω, with its furthermost corner point touching the upper bounding plane of Ω.

Additional A-sets are shown in Fig. III-4. Of potential interest is the "imbedded capacity" A-set in Fig. III-4(d), where most of the available capacity is dedicated to the user classes, but some is held in reserve as an "overflow channel." Those calls that are blocked on the dedicated channel overflow to the reserve channel, where they fully share the reserve capacity with all other calls that have overflowed. The mathematical statements for the A-sets of Fig. III-4 are given below:

1. $A = \Omega \equiv \{j_i | \Sigma\, c_i j_i \leq C_o,\ j_i \leq N_i\}$

   Fully Available Capacity. All calls treated equally, with access to any user class of any available capacity in system.

2. $A = \prod_{i=1}^{K} \{j_i | c_i j_i \leq C_i;\ \text{where } \Sigma C_i = C_o\}$

   Capacity Dedicated to Each User Class. Capacity apportioned to each user class, with demand access only within that capacity apportioned to each class.

*The equation for this plane is $\Sigma c_i j_i = C_o$. This is a direct property of the linear model for capacity use. It is believed that the theory can be extended to algebraic nonlinearities which can be expressed geometrically as smooth bounding curved linear surfaces in the K-dimensional state space.

3. $A = \{j \mid \sum_{i=1}^{M} c_i j_i \leq C_o - \sum_{i=M+1}^{K} C_i\} \times \prod_{i=M+1}^{K} \{j_i \mid c_i j_i \leq C_i\}$

Partitioned Access. First M-K classes fully available access per item 1 above, and last K-M classes fully dedicated per item 2 above.

4. Imbedded Capacity. Choose a point j* with coordinates j_1^*, j_2^*, etc., interior to Ω. Inscribe a right rectangle resting on the positive coordinate axes with vertices at the origin and j*. Then project forward each face of this rectangle not contained in some coordinate plane until it touches the plane $\Sigma c_i j_i = C_o$. Let the set A be the resulting inscribed complex rectangular solid. This A-set provides a level of dedicated capacity to each user class defined by the point j* which uses up $\Sigma c_i j_i^*$ capacity units. The residual capacity $C_o - \Sigma c_i j_i^*$ is then fully shared.

From the examples it can be seen that the system architecture for capacity sharing is expressed in the specification of the admissible set of states A. It should be noted that there are limitations on specifying A-sets. Not only must they lie within Ω, but, because of the Birth-Death model, any j state in A must be reachable by passing through other j states within A (i.e., A must be fully connected). A performance objective can be specified by determining blocking sets B which will be contained in A by computing P(B).

The subset B of states which cause newly arrived calls to be blocked lies on the outer upper edges of the A-set. Performance of grade of service will be determined by summing the probability of being in an edge state of the A-set. Thus, the transmission capacity, the strategy for sharing it, and the blocking conditions are mathematically characterized by the choice of A-set. The connection between the user traffic and

the transmission system is made by determining the steady-state probability $P(j)$ that a particular state j in A is occupied. $P(j)$ is physically the probability of circuit occupancy by the K user classes.

The probability values $P(j)$ are found by solving a second-order K-dimensional difference equation (i.e., the "equation of state") whose coefficients are determined by the traffic parameters and whose boundary conditions are determined by the edge surfaces of the A-set. Because the traffic is statistically modeled as purely random with exponential distributions, the equation of state is of the Birth-Death type Markov process. When the dedicated A-set is chosen, it is physically clear that one is dealing with K noninteracting channels in parallel.* This, then, reduces to the classical (one-dimensional) Erlang model (which must be solved K times, once for each channel).

For the fully shared A-sets (and all others), one seeks to "separate" the equations of state in a manner analogous to the dedicated A-set. The separating property is of great practical as well as theoretical value, and its study was one of the major efforts as shown in Appendix C.

Once the basic probabilities $P(j)$ of each allowed state or call occupancy state are found, key system performance measures can be calculated. The grades of service or blocking probabilities are computed by summing $P(j)$ along the appropriate boundaries of the A-set. The average, in any moment, of the number of calls in service can be found. This, in turn, determines system utilization, defined as the average capacity in use divided by the total available capacity.

*Note, however, that one still is at liberty to choose the separate capacity allotments to each channel.

2. Form of P(*j*) Solution

The principal theoretical result is a characterization of the general validity of the solution form* for P(*j*) independent of the A-set, provided the A-set satisfies a very desirable system objective. Namely, the A-set should ensure that the capacity used by an active circuit is immediately returned for reuse when a call is completed. This ensures the folliwing geometrically important property. Take any state j* in A. Drop K perpendiculars from j* to the orthogonal lower boundary planes of A formed by the coordinate axes. Then, requi.. that *all* of the *j* points that these perpendiculars pass through also be in the A-set. As shown in Appendix C, this property (denoted as "coordinate convexity") is equivalent physically to returning terminated call capacity and permits the recursive-type solution method which "separates" the K-dimensional equation of state.

Given that an A-set meets the conditions of being contained in Ω, connected, and "coordinate convex" (as do the examples given in Fig. III-4), then the steady-state probability P(*j*) of being in state *j*, that is to say, the probability that there are precisely $j_1, j_2, \ldots j_i \ldots j_K$ users in service from each user class, has the following form:

$$P(j) = P_A(0) \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i}$$

$$a_i = \lambda_i/\mu_i \qquad \text{(III-1)}$$

$$1/P_A(0) = \sum_{j \in A} \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i}$$

*Remember that the boundary conditions for the state equation depend on the A-set, so that the solution form for P(*j*) could depend on the A-set.

where, for each user class, i = 1, 2, ... K

$\underline{j} = (j_1, j_2, \ldots j_K)$ = state vector

j_i = number of calls in progress from user class i

N_i = number of sources in class i

$\binom{N_i}{j_i}$ = number of combinations of N_i objects taken j_i at a time

$a_i = \lambda_i/\mu_i$ = the source call intensity given by the average number of call generations per source during average call holding time for class i users

$P_A(0)$ = a normalization constant, independent of $\underline{j}$.

$$1/P_A(0) = \sum_{\underline{j} \in A} \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i} \qquad \text{(III-1')}$$

For those user classes with $N_i = \infty$, replace the term $\binom{N_i}{j_i} a_i^{j_i}$ with $\tilde{a}_i^{j_i}/(j_i!)$. Note that the constant $P_A(0)$ is the probability that there are no calls in progress. Thus, $P_A(0)$ is the fractional amount of time none of the capacity is in use.

Equation III-1 has an extremely important form. It is a product of two quantities. Notice that the K-fold product of combinatorial factors on the far right (the traffic factor) depends only on the traffic and is totally independent of the transmission system and the sharing strategy (A-set). These numbers can be computed directly from projected traffic levels ("requirements") and the independent variables j_i, the number of calls in progress from class i, with no reference to the transmission system structure.

The available capacity C_o and the access strategy (A-set) are only manifested in the calculation of $P_A(0)$. $P_A(0)$ in

Eq. III-1 is completely independent of any particular state j in the A-set. For a given specification of traffic level and per-circuit capacity draw of each user class, the system factor $P_A(0)$ can be calculated for different A-sets (which logically includes changes in available capacity). Thus, the form of $P(j)$ does not change as the A-set or the traffic level is varied. Note that the simplicity in the formal expression for $P(j)$ in Eq. III-1 is misleading in that the calculation of $P_A(0)$, in particular the sum over all j states in A, can easily get out of hand.

Another feature of Eq. III-1 is the product form, which suggests that the calls in progress (thus capacity usage) from each user class might be statistically independent. In Appendix C, it is shown that this is generally false; the calls in progress from the user classes are interdependent. However, since $P(j)$ is in a product form, the mathematical manipulation of the probabilities enjoys some of the advantage of statistical independence. Only in the case where capacity is dedicated to a user class are the calls from that class statistically independent of the remaining classes.

3. Blocking and Performance Measures

From Eq. III-1, performance measures can be calculated. For each i = 1, 2, ... K, then, denote i^{th} user class blocking sets as B_i, i.e., those subsets of states j in A such that inadequate capacity is left to service one more call from user class i. [More precisely, $B_i \equiv B_i(A)$. It remains an analytical exercise to find the B_i for a given A-set.] Since the sets B_i are those states from which additional call requests from user class i will be blocked, the probability of blocking on user class i is $P(B_i)$:

$$P(B_i) = \sum_{j \in B_i} P(j) , \qquad \text{III-2)}$$

where $P(j)$ is given in Eq. III-1.

The grade-of-service (GOS) objective refers to a specification to users that the probability that their traffic will be blocked from service is less than an agreed percentage of time. For finite population users ($N_i < \infty$), there is a slight difference between the probability $P(B_i)$ that the system is in a blocked state ("time congestion") and that a user requesting service is blocked ("call congestion"). This is discussed in Appendix C. For one-dimensional user class problems, the distinction is not practically significant. For general multi-user class problems, there may be a significant difference. However, for the "Blocked Calls Cleared" strategy and the exponential Birth-Death probability models used here there is no practical difference between call and time congestion. Thus, although the users are interested in "call congestion," the GOS objective can be satisfied by constraining the "time congestion" to be less than a GOS objective, typically 5 percent:

$$P(B_i) \leq \text{GOS}.$$

For convenience, order the user classes according to their ascending capacity need per circuit, so that $1 = c_1 \leq c_2 \leq \cdot a \cdot \leq c_K$. In Appendix C it is shown that for fully shared capacity access the blocking states are "nested" $B_1 \subseteq B_2 \subseteq \ldots \subseteq B_K$. Thus, it follows from fundamental principles that the blocking probabilities are in ascending order $P(B_1) \leq P(B_2) \leq \ldots P(B_K)$, independent of the offered traffic level given by the a_i and N_i.

The marginal probability that there are K calls in progress from the i^{th} user class, $j_i = K \leq 0, 1, 2, \ldots \hat{j}_i$, is found by summing $P(j)$ over all j states whose i^{th} coordinate is equal to K. The largest value K can have for the i^{th} user class, denoted $\hat{j}_i$, is found by taking the maximum of the i^{th} coordinate of j as j varies over A. Thus, the marginal probability $Q_i(k)$ that $j_i = k$ is given by

$$Q_i(k) = \sum_{j \in \sigma_i(k)} P(j); \quad \text{for } 0 \le k \le \hat{j}_i, \tag{III-3}$$

where the set $\sigma_i(k)$ is all those j states of A whose i^{th} coordinate equals k.

$$\sigma_i(k) \equiv \{j \in A \mid j_i = k\} \tag{III-3'}$$

The $Q_i(k)$ are the probabilities of interest to the i^{th} user class. Note that, except for dedicated capacity A-sets, the user classes are not statistically independent, so that

$$P(j) \neq \prod_{i=1}^{K} Q_i(j_i).$$

The ν^{th} ($\nu > 0$) moment* of j_i is

$$\langle j_i^{\nu} \rangle = \sum_{k=0}^{\hat{j}_i} k^{\nu} \, Q_i(k). \tag{III-4}$$

Since the capacity used is linear** on the number of calls in progress, the average capacity used $\langle C \rangle$ is the linear sum of the average capacity drawn by each user class, given by

$$\langle C \rangle = \sum_{i=1}^{K} c_i \langle j_i \rangle, \tag{III-5}$$

---

*The principal moments of interest are $\nu = 1$ and 2.

**Note that this has the result that Eqs. III-5 and III-6 do not hold for nonlinear capacity usage. In that event, one must use the fundamental definition of an average $P(j)$

$$\langle C \rangle \equiv \sum_{j \in A} C(j) \cdot P(j).$$

where, from Eq. III-4,

$$\langle j_i \rangle = \sum_{k=0}^{\hat{j}_i} k\, Q_i(k) \quad . \tag{III-6}$$

One measure of the efficiency with which the capacity C_o is used by an access strategy is given by the average capacity utilization U_A:

$$U_A \triangleq \langle C \rangle / C_o \quad . \tag{III-7}$$

U is subscripted by A to emphasize the dependence on the A-set through Eq. III-1.

In summary, Eqs. III-1 through III-5 provide the theoretical basis for relating traffic activity and capacity allocation strategy to a grade-of-service objective and efficient use of capacity. There still remains the problem of developing the computational tools to implement the indicated calculations, especially the arithmetic formulation and the indicated summation over the sets B_i and A.

DEDICATED AND FULLY SHARED CAPACITY ALLOCATIONS

In this section the theoretical results are specialized to the dedicated and fully shared A-sets. The dedicated strategy is the simpler. It replaces a K-dimensional problem with K one-dimensional problems, each of which is equivalent to the classical Erlang theory of traffic congestion. The commonly used classical results are introduced to provide the reader with a basis to relate the fully shared results.

1. Dedicated Capacity

In this case, the total available capacity C_o is split into K separate noninteracting pools C_i, each of which is accessed independently by users on a demand basis from the assigned user class. This is equivalent to conventional teletraffic theory

(see, for example, Ref. 12). If C_i, i = 1, 2, ... K, is the capacity assigned to user class i, the maximum number of circuits $\hat{j}_i$ available to user class i is given by

$$\hat{j}_i = [C_i/c_i],$$

where [x] indicates the largest integer contained in the number x. Since the user classes are independent, $P(j)$ becomes for each i = 1, 2, ... K and $0 \le j_i \le \hat{j}_i$

$$P(j) = \prod_{i=1}^{K} Q_i(j_i) \qquad \text{(III-8)}$$

$$Q_i(j_i) = K(j_i)\, a_i^{j_i} \Big/ \sum_{j_i=0}^{\hat{j}_i} K(j_i) a_i^{j_i} ,$$

where*

$$K(j_i) = 1/j_i! \text{ for } N_i = \infty \text{ (Erlang formula)}$$

$$K(j_i) = \binom{N_i}{j_i} \text{ for } N_i < \infty \text{ (Engset formula)} .$$

Equation III-8 is the classic B Blocked Calls Cleared formula of Erlang in 1917 (and Engset in 1918). Several key results of this one-dimensional theory are reviewed here for comparison purposes. An excellent discussion of this theory is given in Ref. 13. The probability of blocking for users in class i is given by evaluating Eq. III-8 at $j_i = \hat{j}_i$

$$P(B_i) = Q_i(\hat{j}_i) \qquad \text{(III-9)}$$

$$\hat{j}_i = [C_i/c_i]$$

*Recall that for the Erlang case $N_i \to \infty$, $\lambda_i \to 0$ such that $N_i\lambda_i \to \tilde{\lambda}_i$, so that $\tilde{\lambda}_i$ can exceed 1, while in the Engset case $\lambda_i < 1$.

For the Erlang case, the offered load is taken to be $\tilde{a} = \tilde{\lambda}/\mu$, the average number of call arrivals during mean call duration, and is measured in "Erlangs." In the Engset case, specification of N_i and a_i completely determines $P(j)$. However, conceptual difficulty arises in specifying "offered load." This is caused by the coupling between the finite number N of traffic sources and the available circuits. The instantaneous source population for placing call requests is N minus the calls in progress. For the Erlang model, there is an infinite source population, and new calls are generated independent of the number of calls in progress. Consequently, in the Erlang model the call source and the circuits in use are decoupled, while for the Engset model (finite source population) the two are coupled. This difficulty in defining "offered traffic" is important principally for drawing fair comparisons between Erlang and Engset models. This is discussed further in Ref. 13. A rather accurate approximation in equating "offered traffic" between the two models is given by defining Engset offered traffic as follows:

$$\text{Engset offered traffic} = \frac{N_i a_i}{1+a_i} \approx N_i a_i \text{ for small } a_i, \quad \text{(III-10)}$$

which compares to Erlang offered traffic = $\tilde{a} = \lim \tilde{N}^a = \lim[N^a/(1+a)]$. The Engset offered traffic is also measured in "Erlangs." Another conventional traffic parameter, known as server occupancy, is defined as the ratio of the average number of circuits in use to the total available:

$$\text{Server occupancy} \equiv \langle j_i \rangle / \hat{j}_i = \langle j_i \rangle c_i / C_i \ . \quad \text{(III-11)}$$

Note that the utilization measure of Eq. III-7 is a linear generalization of Eq. III-11:

$$\langle C\rangle/C_o = \sum_{i=1}^{K} (c_i/C_o)\,\langle j_i\rangle$$

$$= \Sigma \langle j_i\rangle\, c_i/\Sigma C_i\,. \qquad \text{(III-7')}$$

For the classical one-dimensional Erlang model, the average number of circuits in use $\langle j_i\rangle$ is simply expressed (Ref. 13) in terms of the i^{th} user class blocking probability $P(B_i)$, where $P(B_i)$ is the Erlang version of Eq. III-8 evaluated at $\hat{j}_i$:

$$\langle j_i\rangle = \tilde{a}_i\,(1 - P(B_i)). \qquad \text{(III-12)}$$

For the Engset model, Ref. 13 gives $\langle j_i\rangle$ in terms of the blocking probability $P(B_i)$, where $P(B_i)$ is the Engset version of Eq. III-8 evaluated at $\hat{j}_i$.

$$\langle j_i\rangle = \frac{a_i N_i\,(1 - P(B_i)) + \hat{j}_i a_i P(B_i)}{1 + a_i} \qquad \text{(III-13)}$$

Equations III-8 through III-13 apply only to the dedicated capacity A-set strategy. However, the concepts of offered load, intensity, blocking (congestion), and utilization do pertain to other A-set strategies. It is of theoretical interest to note (Ref. 13) that Eq. III-8 describes the state probabilities $P(j)$ for general call holding time distribution and not just for exponential distributions. For general A-sets, other than dedicated capacity, this generalization in the call duration probability distribution (nonexponentially distributed calls) need not hold.

Provided separate independent queues are employed for each user class, the restrictions to "Blocked Calls Cleared" can be removed without any additional mathematical complexity (Ref. 13).

This results in the well-known "Blocked Calls Held" (with finite or infinite holding space, i.e., queue length) formulae. With a very slight increase in complexity, the "Blocked Calls Delayed" model can be treated. Here, when an incoming call finds all circuits busy, he waits in queue no more than a fixed length of time T. If he is not provided a circuit in time T, he departs from the system as a "lost" call.

2. Fully Shared Access

The opposite extreme to partitioning the available capacity C_o into K separate pieces, each dedicated to a user class, is common sharing of C_o among all the users, irrespective of their class. In this situation the set of admissible states, A, is equal to Ω, the set of all available states bounded by the constraint $\Sigma c_i j_i \leq C_o$. For this case the system state probabilities are given by Eq. III-1, with A replaced by Ω. No further simplification is possible.

Furthermore, there is no simple extension of the theory to deal with strategies other than "Blocked Calls Cleared." The mathematically trivial extension to "Blocked Calls Held" modes of operation for the dedicated capacity allocation fail for the more general A-sets, including the fully shared strategy. The basic reason for this failure resides in the way in which a rule or set of rules (protocol) must be mathematically implemented in the state equations (i.e., Eq. C-1 in Appendix C) in order to structure the waiting room of blocked calls of different user classes and the manner and order in which newly available circuits are assigned to waiting calls. To illustrate this point, consider three user classes, small, medium, and large in their need of circuit capacity. Suppose further that the system is blocked, a mix of medium and small calls are being held but no large call is waiting, and a large call in progress is completed. To what mix of waiting medium and small calls should the capacity of the newly terminated large call be assigned? The protocol rules of held-call assignment must be complete and consistent (i.e., they must cover all situations and never be in conflict).

The analysis of such protocols in predicting system behavior depends on two mathematical requirements. The first is the specification of a set B of j states (analogous to A), which structures the "waiting room" and a means of translating the call protocol into a specification of Birth-Death coefficients in the equations of state. The Birth-Death coefficients (μ,λ) can also depend on the state location in B. Thus, the equations of state can become exceedingly complex. A salient feature of the state equations (Eq. C-1) for the "Blocked Calls Cleared" case is their linearity (although they do not have constant coefficients). It is possible that some protocols for "Blocked Calls Held" may cause the equations of state to become non-linear.

Two reports in the technical literature (Refs. 19 and 20) have addressed this problem. In both studies only two classes of users, wide band and narrow band, were considered but with differing protocols. In Ref. 19, only wide-band calls are held; narrow-band calls are cleared. In both studies the resulting equations of state, although linear, are quite complex even for such conceptually simple systems. Reference 19 requires eight different statements for the state equation, one for each of eight regions of state space, while Ref. 20 requires nine different statements for the state equation.

Evaluation of Eqs. III-1 with a digital computer is required. This immediately takes the problem* into the realm of computer science in order to alleviate several computational limits. Central to this issue is the growth in computational complexity as driven by a large number of user classes K and/or

*This problem is not limited to the fully-shared-capacity case but is ubiquitous to the theory. For the case of dedicated capacity, the issues are made less pressing due to the separation of the problem into K independent problems of smaller degree.

capacity C_o, both yielding algebraic growth in the number of states j in Ω. This, in turn, leads to exponential growth in computing complexity. Specific problems arise with regard to

1. Core memory space
2. Arithmetic overflow/underflow
3. Execution time.

If a direct approach is taken to solving Eq. III-1, one first evaluates the product terms and then sums them over all j states in A to normalize the probability. Thus, items 1 and 3 follow directly with growth in the number of possible j states. But, additionally, item 2 occurs in that very large numbers can develop when $N_i a_i$ (or $\tilde{a}_i$) and j_i (note $j_i \leq C_o/c_i$) are large. Straightforward summation over these large values (prior to normalization) will cause computation overflow. Another item of concern is the determination of the blocking sets of states B_i, as defined in Appendix C. Application of a brute-force computational test, as implied in the definition of the blocking set, yields a very large increase in the number of computational steps and long execution times.

In order to evaluate the probability of blocking and utilization for some relatively simple cases, a computer program was developed (Appendix C) capable of evaluating Eqs. III-1 for either dedicated capacity (Eq. III-6) or fully shared A-sets with either the Engset (finite) or Erlang (infinite) source models. This program is written in Fortran IV and can be executed on a remote time-sharing computer system. It does not address the inherent computational problems and as a consequence is limited to systems of relatively small scale. No more than three user classes are permitted ($K \leq 3$, small-, medium-, and large-capacity traffic sources).

The determination of the blocking states B_i is approximated with an upper bound derived in Appendix C. That is to say, a

number, denoted $\hat{P}_i$, is found such that for i = 1, 2, 3, $P(B_i) \leq \hat{P}_i \leq GOS$. (This technique is also used in Ref. 21.) Its attractiveness is in its ease of programming implementation. However, since the true probability of blocking is less than or equal to the bound, the capacity required for a given GOS objective will be somewhat overestimated. Consequently, when making comparisons between fully shared and dedicated capacity allocation, the fully shared approach will be slightly penalized relative to the dedicated approach for equal GOS objectives.

D. COMPARISON OF DEDICATED CAPACITY AND FULLY SHARED ACCESS

1. Comparison Method

The computer program presented in Appendix C compares dedicated and fully shared capacity allocation strategies for up to three user classes, K = 3. The basis selected for the comparison is the resulting total capacity C_o needed for the fully shared strategy, as opposed to that needed for the dedicated strategy for a common grade-of-service (GOS) objective. Also calculated is the average utilization efficiency $\langle C \rangle / C_o$ for each strategy.

Conceptually, the program is given the traffic parameters N_i, $a_i = \lambda_i/\mu_i$, the per-circuit capacity c_i needed, and a GOS_i specification (i.e., probability of blocking $\leq GOS_i$) for each user class i = 1, 2, 3. It then calculates, using the classical results of Eq. III-8, the dedicated capacity C_i needed for each class to achieve the GOS_i objective. The total capacity C_o is then the sum of the C_i.

Using the C_o calculated for the dedicated strategy as an initial value, the program calculates an upper bound approximate to the blocking probabilities that then result with the fully shared strategy. In all cases, this will result in a blocking probability that is better than the GOS objective. The initial value of C_o is then decremented a selectable fixed

amount (i.e., $C_o - c_i$; i = 1, 2, or 3), and the process is reiterated until the blocking objective is just exceeded. The total capacity calculated on the iteration just prior to the one that exceeded the GOS objective is then the C_o needed for the fully shared strategy.

There are several points to be kept in mind with regard to this comparison. First, although this comparison is meaningful for capacity, there are other bases* for comparison that could be used. The one chosen is convenient and serves to show by example the utility of the theory. Second, the comparison assumes that traffic parameters are known (or are measurable) and that capacity is allotted in a matched manner to the level of traffic offered. In this regard, the dedicated strategy is less robust to capacity/traffic mismatch. Consequently, the comparison tends to show the dedicated strategy in a somewhat** more favorable light. As the sample results will show, although there is no overwhelming capacity penalty for choosing to organize the allocation of system capacity along dedicated lines, such organization must be flexible*** enough to match each user class to its level of traffic. Otherwise, considerably poorer performance (increased blocking) and/or excessively idle capacity will result. [Note that while these opposing effects cannot simultaneously occur within the same user class, without flexible capacity (re)assignment, they can occur simultaneously in the overall system.]

---

*For example, for a given capacity and equipment cost curves, compare achievable blocking probability as a function of available expenditures.

**As the traffic volume $N_i a_i$ gets larger, system traffic and capacity uncertainties tend to have less overall impact on performance provided the system is not traffic saturated (i.e., traffic does not exceed allocated capacity).

***Specifically, each user class should be able to technically operate with differing levels of allocated capacity C_i.

2. Sample Comparsions

The program in Appendix C was first tested for accuracy. For example, the probability tables were checked against hand calculations (since the tables are recursively generated starting with $j_i = 0, 1, 2$, exact checks at small j_i values ensure accuracy over the full range of j_i). This verifies the accuracy of the dedicated allocation mode of the program. The shared mode was checked by making the three user classes all the same. Then, symmetry dictates that $\langle j_1 \rangle = \langle j_2 \rangle = \langle j_3 \rangle$; such was the case. Further, three equal user classes with 10 Erlangs of offered traffic should have a shared capacity requirement equal to one dedicated user class with 30 Erlangs of offered traffic. The program was also checked for this property.

First considered were comparisons between dedicated and shared capacity usage as a function of offered traffic for two different GOS levels (0.05 and 0.01) and two mixes of user characteristics. In all of these cases the user class circuit capacity was taken to be $c_1 = 1$, $c_2 = 2$, $c_3 = 4$ (representative of bit rates in the ratios 1:1, 2:1, and 4:1, e.g., 1.2 kb/s, 2.4 kb/s, and 4.8 kb/s). The activity for the Engset finite-source model of any source, independent of user class, was taken to be $a_1 = a_2 = a_3 = 0.1$ (a 10% user activity factor). The offered traffic for the Engset model was varied by changing the number of sources N_i in the user class. The effective offered traffic is thus $a_i N_i/(1 + a_i) = N_i/11$.

For this limited comparison, the two mixes of user sources considered were:

1. Balanced Traffic:* the product $N_i c_i$ = constant, thus $N_3 = N_1/4$, $N_2 = N_1/2$
2. Equal Traffic:* N_i = constant, thus $N_3 = N_2 = N_1$.

*More generally, N_i should be replaced by $a_i N_i/(1+a_i)$. However, since all a_i were set equal, the simpler expressions were used.

The results of GOS 0.05 are shown in Figs. III-5 and III-6. Figure III-6 extends Fig. III-5 to lower levels of offered traffic. [Note the granularity in the fixed-equal capacity curve; the circuits needed must be divisible by three, and consequently marked jumps occur (see Appendix C).]

Figures III-5 and III-6 are qualitatively very revealing and indicate the following trends.

1. Above moderate levels of offered traffic, needed capacity increases linearly with offered traffic and increases at approximately the same rate, whether dedicated or fully shared strategies are used. The capacity difference between the two is constant. Thus, the percentage increase in needed capacity for the dedicated strategy tends to zero at the higher levels of offered traffic.

2. At low levels of traffic, the increase in capacity needed by the dedicated allocation can be as high as 50 percent.

3. There can be as large or even larger an effect on capacity utilization by organizing user class traffic (connectivity needs permitting) as in dedicated versus shared strategies. The spread in needed capacity between balanced and equal source mixes is greater than between dedicated and shared strategies for a given mix of user class sources. For the case computed, the rate of needed capacity increase with offered traffic is 25 percent faster for the equal traffic user classes than for the balanced traffic case. Balanced sources $\approx$1.8 capacity units/Erlang, while equal sources $\approx$2.4 capacity units/Erlang.

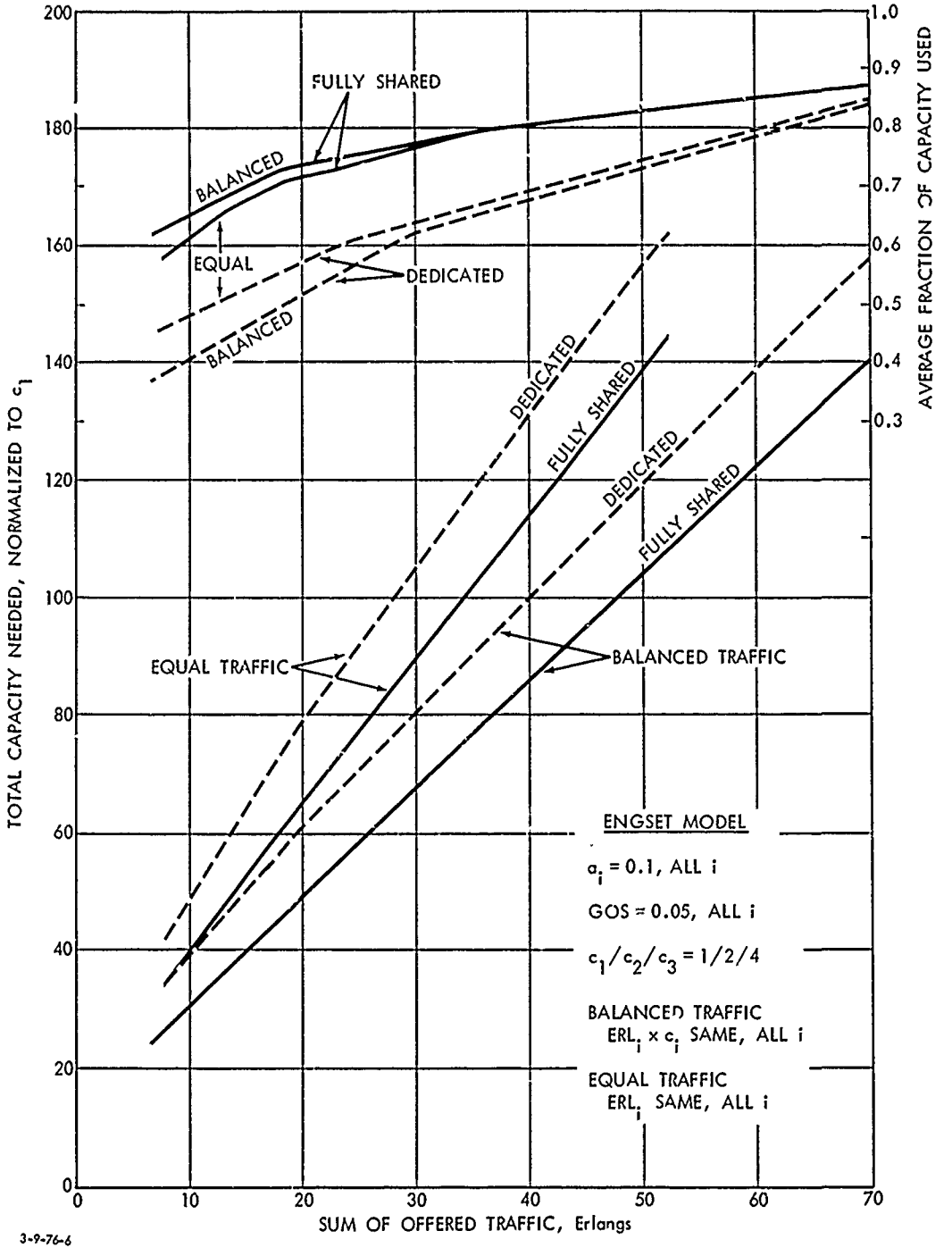


FIGURE III-5. Capacity Comparison, Dedicated versus Fully Shared for Two Traffic Mixes: Balanced and Equal

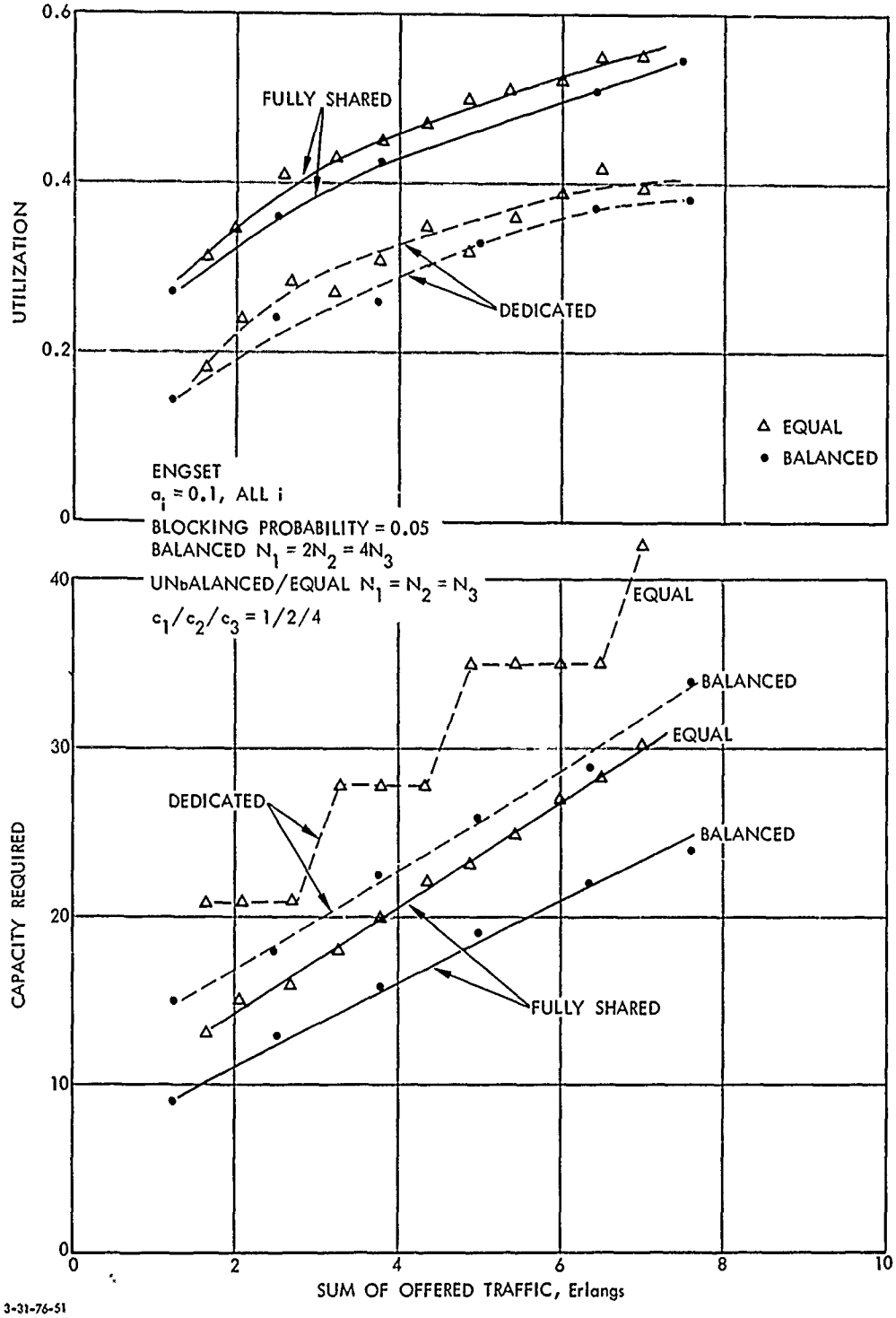


FIGURE III-6. Comparison at Low Traffic Level

4. At approximately 70 Erlangs of offered traffic, the program used ran into computational overflow.

In Fig. III-7 a comparison is made between a 0.01 and a 0.05 grade-of-service blocking objective for the case of balanced user classes. This replicates the same trends of Fig. III-5 with almost the same asymmetric capacity per Erlang slope. But note the following:

5. The almost constant difference in required capacity between fixed and shared strategies increases with increased grade of service, (i.e., reduced blocking probabilities).

In Figs. III-8 and III-9 a comparison is made between the Engset (finite source) model and the simpler Erlang (infinite source) model for 0.01 and 0.05 blocking objectives with dedicated and shared strategies. These results indicate that:

6. For balanced user classes with total offered traffic in excess of 5 Erlangs, the two finite and infinite source models predict results in very close agreement. Using the somewhat simpler Erlang model as an approximation to the Engset model provides a conservative estimate for required capacity to meet a grade-of-service objective.

As a last example, a hypothetical mix of users is considered (e.g., Naval task force). Three user classes are modeled* as follows:

*This model is exceptionally simple as it does not address source-sink pairs of circuits. The model considers only the platform activity and capacity rather than all pairings of transmit activity receiver sensitivity. It does not, for example, consider detailed source-sink pairings on capacity requirements, such as flag-to-flag, flag-to-element, flag-to-leader, leader-to-element, etc. Theoretically, this would generate more classes of users than the program could handle.

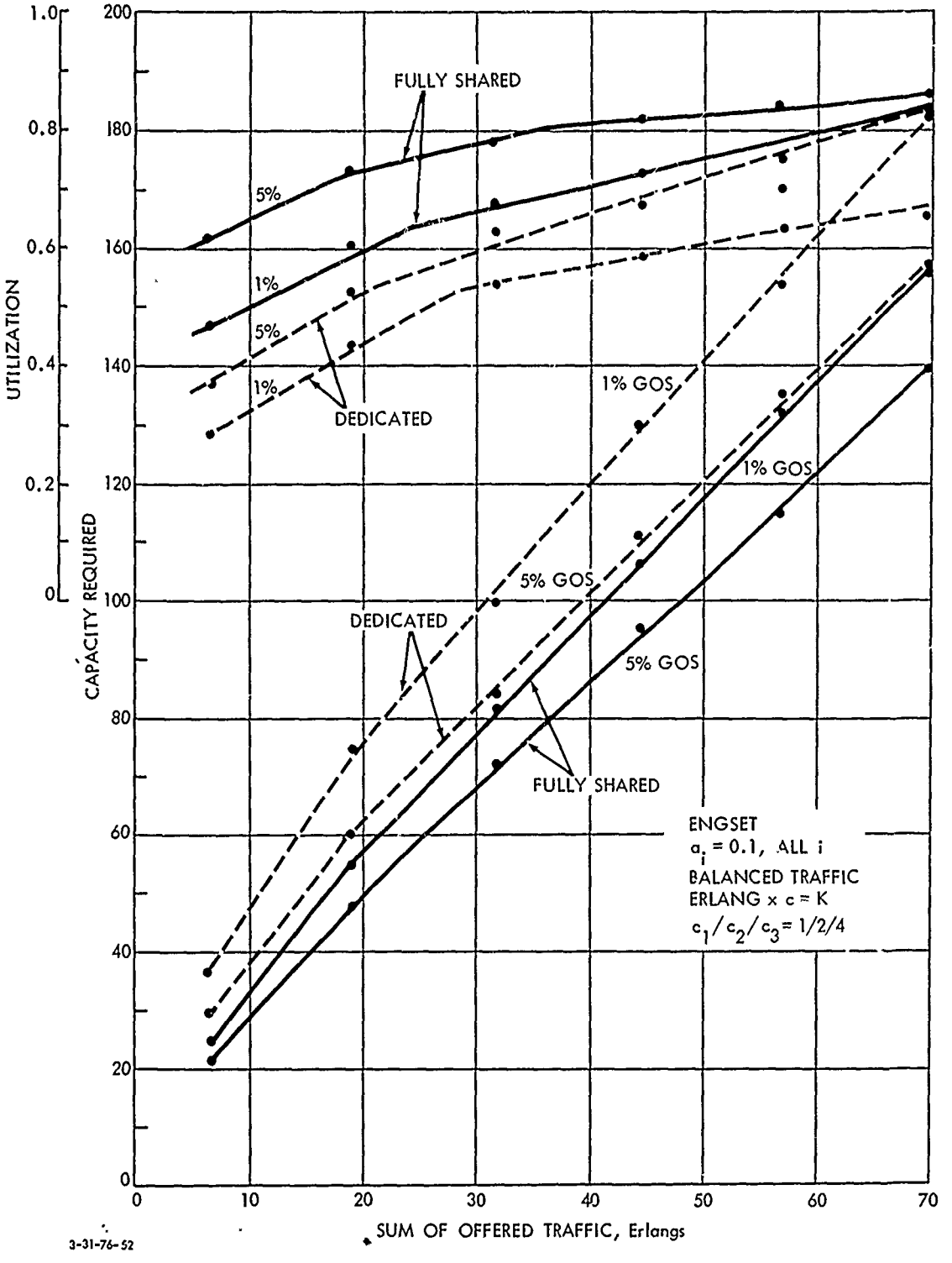
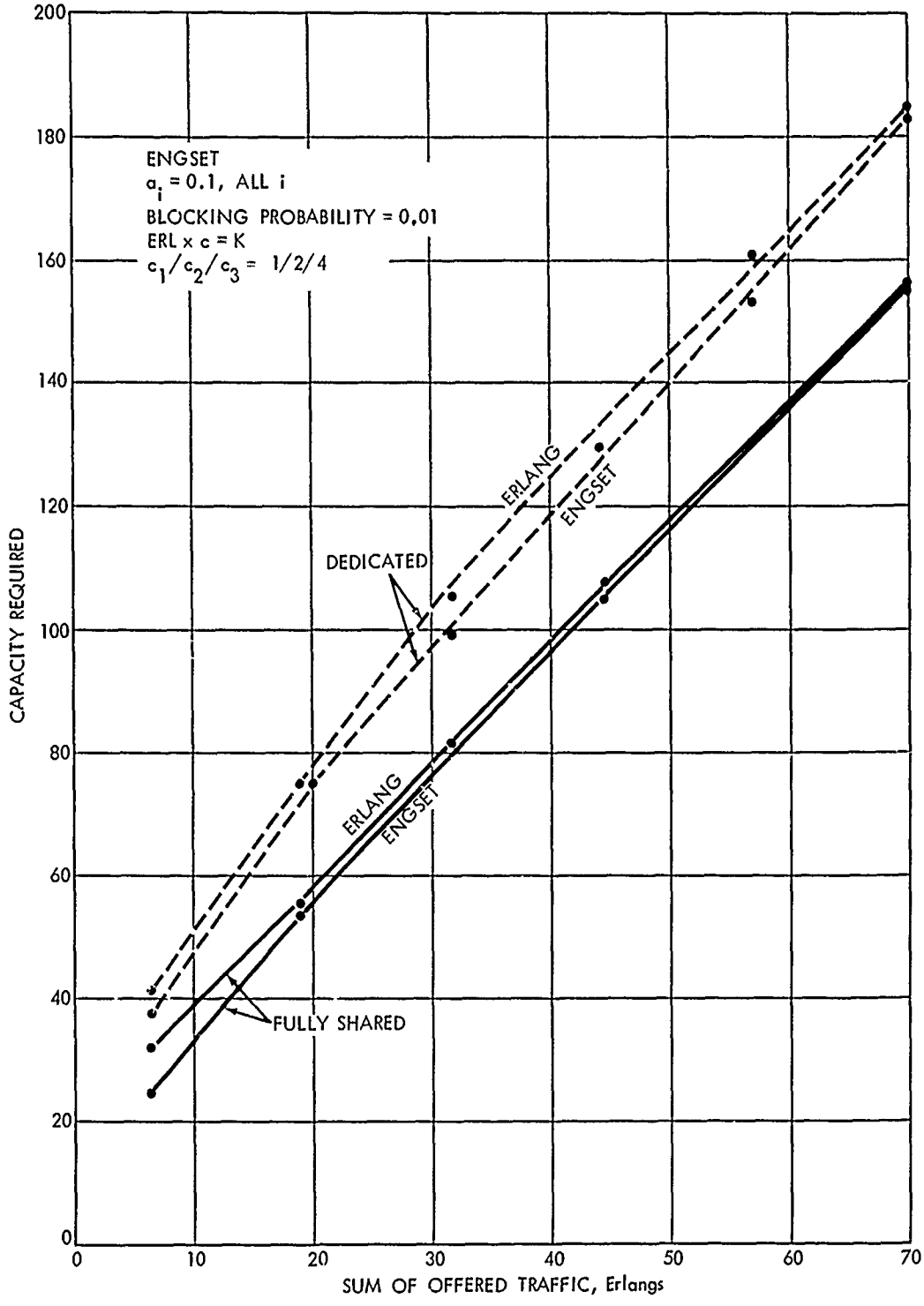


FIGURE III-7. Capacity Comparison for Different Grades of Service



3-31-76-53

FIGURE III-8. Erlang/Engset Model Comparison at 1 Percent GOS, Balanced Traffic

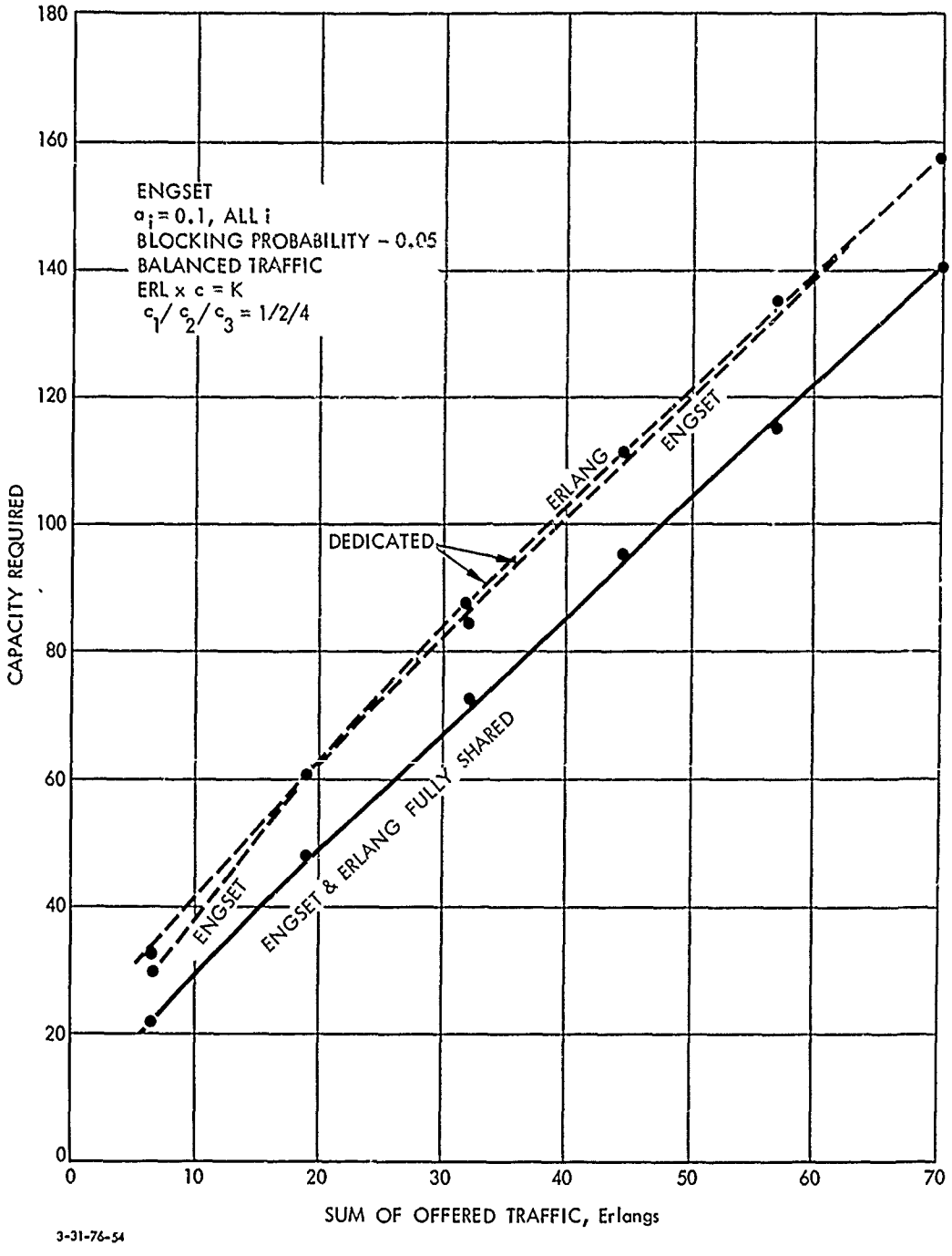


FIGURE III-9. Erlang/Engset Model Comparison at 5 Percent GOS, Balanced Traffic

Class 1. Leader Class

Typical of circuits off medium flagships [e.g., destroyer leader (DL)] with moderate traffic source parameter $a_1 = 0.1$, moderate bit rate/circuit, and moderate terminal EIRP and G/T. Thus, satellite capacity* needed, c_1, is taken as unity for this class.

Class 2. Force Element Class

Typical of circuits off force element ships [e.g., destroyer (DD, DE)] with low traffic source parameter $a_2 = 0.01$ and low bit rate/circuit, but even lower G/T. Hence, capacity/circuit $c_2 = 4$.

Class 3. Major Flagship Class

Typical of circuits off major flagships (e.g., carrier or cruiser) with high traffic source parameter $a_3 = 0.5$, but, although high G/T, even higher bit rate, so $c_3 = 4$.

In a task force, one might expect that overall user traffic would tend to correlate with the number of sources on a major flagship.** Consequently, required satellite capacity for dedicated versus shared strategies is plotted in Fig. III-10 as a function of the number of major flagship sources N_3 (rather than offered traffic, as in the previous figures). Two different mixes of N_1 and N_2 to N_3 sources were used. The solid lines show the case where there are eight times as many force element sources as major flagship sources and half as many leader sources

*Capacity is taken proportional to (bit rate)$^{-1}$ x G/T.

**There may be more than one user circuit per flagship and/or more than one flagship per task force.

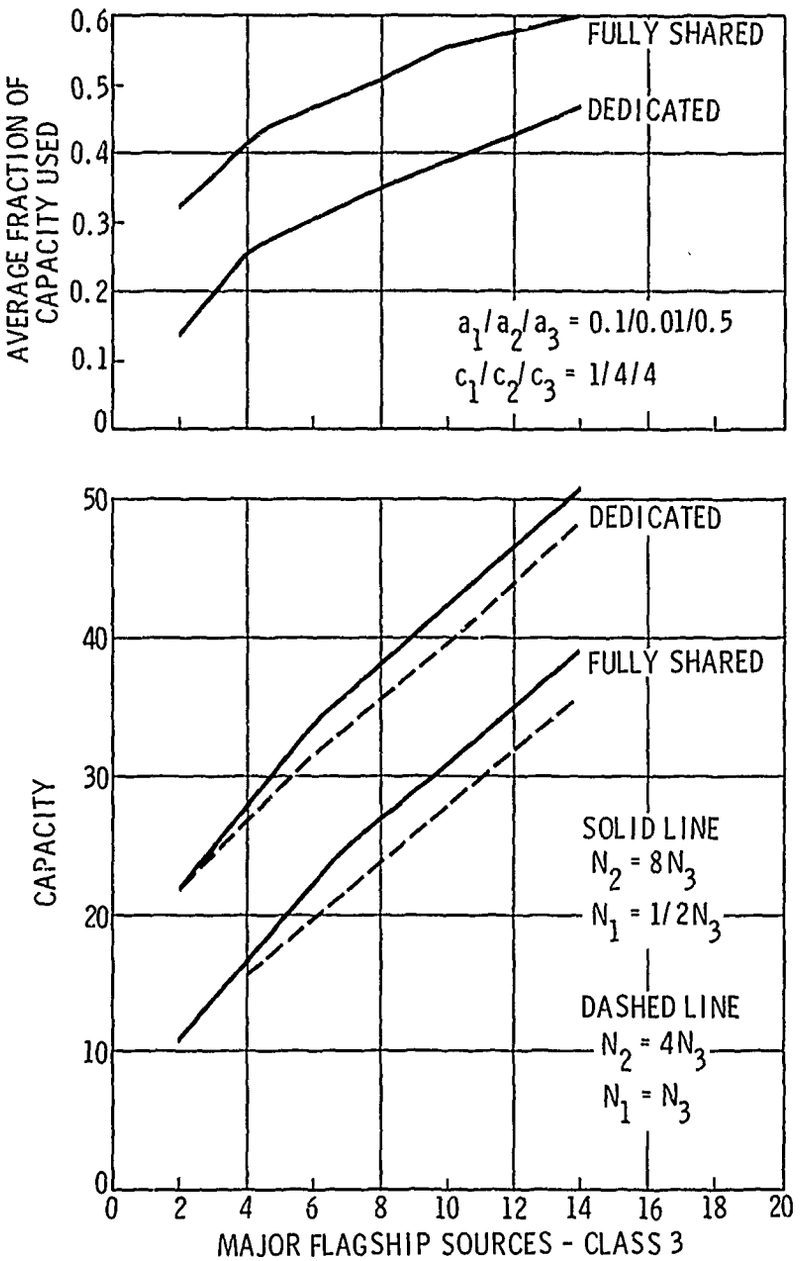


FIGURE III-10. Hypothetical Naval Task Force Example

as major flagship sources. The dashed line shows the case where there are four times as many force element sources as major flagship sources and the number of leader sources is equal to the major flagship sources.

7. The results are qualitatively similar to those of the fixed and shared allocation schemes in the previous examples, namely, there is a relatively constant capacity difference between the two allocation strategies.

8. At a moderate number of major flagship sources (eight), the percentage capacity gain for the fully shared strategy is about 50 percent. This diminishes percentagewise with higher traffic level.

9. The difference in the required satellite capacity is small between the two cases of source mix chosen.

The above results lead to the following conjecture:

10. When dealing with many user classes ($K > 3$), it may be possible, for the purpose of reducing the number of classes to be analyzed, to computationally treat as a single compound class those individual classes with a close balance in traffic circuit-capacity product. That is to say, merge into larger equivalent classes those user classes having $a_i N_i c_i/(1+a_i)$ products approximately the same.

A final observation would be that since the capacity savings are not great in the cases studied between the extremes of dedicated and fully shared capacity allocations:

11. Further insight as to differences between alternative capacity allocations will derive from:

    a. More sophisticated traffic models and call priority/holding strategies

b. Extremes in mixes of traffic sources

c. Interconnection with more complex systems of resources (e.g., complex terrestrial switching nodes and multiple satellites)

d. Special connectivity/addressability features of military operations.

IV. SATELLITE CAPACITY ALLOCATION FOR STORE-AND-FORWARD SYSTEMS

A. INTRODUCTION

The problem of allocating satellite capacity to users for store-and-forward data communication ("message" switched) services is examined in this chapter. It is assumed that a communication satellite system exists from which communication channels have been allocated to provide service for store-and-forward data messages. The specific manner in which these communication channels are derived (e.g., the satellite and user communication hardware and the multiple-access technique) is not manifested in the theoretical development pursued here.* The communication channels are to be shared among a group of users. The users are organized into nets, each of which is assigned a communication channel with sufficient capacity, represented as an equivalent line speed or bit rate, to handle the generated store-and-forward message flow. The net members time-share the assigned communication channel. Methods of allocating transmission time to the users and the means of evaluating various allocation techniques are investigated in this chapter.

A significant amount of work has been performed in conjunction with FLEETSATCOM in studying and developing the hardware and software needed to implement various allocation strategies for store-and-forward message communication systems, i.e., the Information Exchange Systems (Appendix B). It was deemed

*Chapter III contains a limited discussion.

that more theoretical work was required in establishing quantitative performance measures by which the various allocation techniques can be compared. The work reported in this chapter is also applicable to certain message communications to be supported by AFSATCOM (e.g., force-element report-back and general-purpose communications) and potentially to future DSCS systems.

Representative time-division techniques from computer data communications are investigated for application to store-and-forward message communication systems with satellites. The objective of this section is to develop quantitative performance measures for comparative purposes and for performing the trade-off between capacity and traffic intensity. The analytical approach used can provide insight into the behavior of the techniques under various loading and their sensitivity to overload. The insight provided can also aid in determining the choice of system simulations and in interpreting the results thereby obtained.

The performance measures chosen and the methodology used for determining those measures parallel the work that has been done for computer communication systems (Ref. 22), which has its basis in queueing theory. The communication system is viewed as a service system. The data messages are the customers, the communication channel is the server, and the transmission of data from one user to another is the service performed. The method of allocating transmission time to the net members is the service discipline. Each user subdivides his allocated transmission time to the stored messages according to some queueing discipline (e.g., first in, first out; last in, first out; or a priority system). Under certain disciplines, it is possible to apply the queue discipline to all the messages generated in the net, and not just separately at each user.

Some of the system performance measures used in queueing theory are:

- The queue length, which is the number of customers in the system being served or waiting on the queue. The queue length is independent of the chosen queueing discipline.

- The waiting time, which is the time the customer waits in the queue for service. In this chapter, the term "queueing time" refers to the total time a customer spends in the system and is equal to the waiting time plus the service time. Both waiting and queueing times are dependent upon the queueing disciplines.

These queueing performance measures are usually evaluated in the steady state. The queue lengths and waiting times are also natural measures for users of the satellite communication system because each user would like to determine the buffer size needed to store his messages and would like to know how long he must wait to transmit a message. The waiting time is also a natural measure of the "grade of service" that is provided. The service discipline chosen (the method of time-dividing the channel among the users in the net) affects the queue length and waiting time and therefore will also affect the net capacity allocation needed to achieve a "grade-of-service" (waiting time) objective.

Messages originate at a user, are buffered or stored at a terminal, and then are transmitted via satellite (hence, the designation "store and forward") to the destination user terminal. The user traffic model used assumes that the message arrivals and lengths are stationary random processes. The message arrival process is Poisson, and therefore the related interarrival time between messages is exponentially distributed. The message length is also assumed to be exponentially distributed and independent of the arrival process. In our development, it is more appropriate, as will be seen, to describe a

message in terms of data units that are fixed data blocks of bits. The distribution of the number of data units contained in a message is geometric, as a consequence of the exponentially distributed length (Ref. 23). The arrival process for data units is classified as compound Poisson.*

It is assumed that all the terminals in the net transmit and receive at the same data rate, R bits/second, although different bit rates can be incorporated. To ensure that the net operation is stable (i.e., that the user buffer contents and waiting times remain finite), the data rate allocated to the net must be such that the average number of message arrivals in the net does not exceed the average number of messages that can be transmitted. Depending upon the allocation technique chosen, other parameters such as the overhead associated with a transmission and the "walking" time** can affect system stability and performance. The "walking" time is the time needed to transfer access to the satellite channel from one user to another (to be discussed later).

The system performance measures used in comparing the effects of the service discipline on the net operation are the queueing performance measures discussed earlier but applied to the net. The measures are:

- The average over the net of the mean values of the terminal buffer queue lengths, given by

$$L = \frac{1}{N} \sum_{i=1}^{N} E\left\{L^{(i)}\right\} ,$$

*The results of a study of traffic characteristics in certain time-sharing computer systems (Ref. 24) indicate that the interarrival time between messages can be approximated by the exponential distribution and that the length of messages is approximated by the geometrical distribution.

**For convenience, in Chapter II the term "walking time" also included the overhead associated with a transmission.

where $L^{(i)}$ is the buffer contents of the i^{th} terminal in steady state, and N is the number of terminals in the net.

- The net waiting time, given by

$$W_q = \frac{1}{N} \sum_{i=1}^{N} W_q^{(i)}$$

where $W_q^{(i)}$ is the average waiting time for a message buffered at the i^{th} terminal.

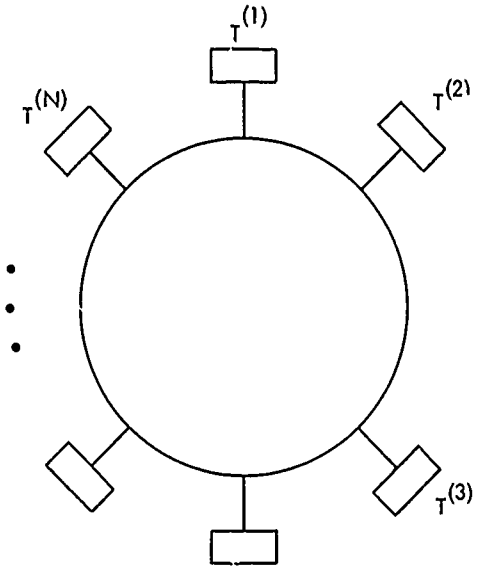
The chosen performance measures equally weight all net members. For certain situations, measures with unequal weighting (e.g., a weight proportional to the average terminal traffic load or a priority weighting of the users) may be more appropriate.

The above system performance measures determine the effects of the time-sharing techniques on the user for a specified net size and capacity allocation. The system designer still must determine the capacity (bit rate) that will be allocated to the net. One method of sizing the amount of capacity is to determine the capacity that just satisfies a grade-of-service objective for the net. In circuit switching, the grade-of-service is typically the probability of blocking (in the "Blocked Calls Cleared" case), i.e., the probability of being denied a circuit at the time of request. Because messages are buffered and not cleared, a more natural grade-of-service measure is either the net average waiting time or the queueing time of a message. Therefore, another method that is also used to compare time-sharing techniques is to determine and compare the capacities required by the techniques to achieve a specified grade-of-service objective.

In a computer communication loop system, a group of terminals is arranged (Fig. IV-1) in an orderly fashion by terminal position around a communication circuit or line which they share in transmitting data to a central computer. In this context, various service disciplines (Refs. 22, 25-28) of allocating transmission time on the circuit have been proposed and investigated; the corresponding terminal buffer queue lengths and waiting times have been developed. There are differences between this loop system and the satellite communication net that impact upon the applicability of some of the proposed service disciplines.

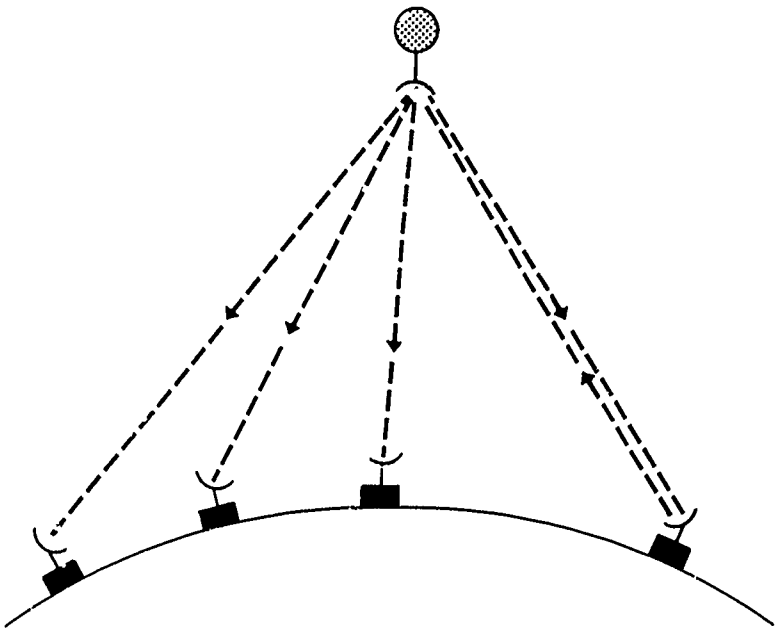
In the satellite net (Fig. IV-2), the users can be dispersed over a vast geographical area and are not positioned in any orderly fashion. There is a nonnegligible propagation delay from the time a user transmits a message to the satellite to the time the message is received. This propagation delay prevents a user from ascertaining transmission activity in the present. A user's perspective of the activity is what has occurred a propagation-time delay ago. Further, a transmission can, in principle, be received by all terminals included in the communication coverage and not just by the terminals assigned to the net (Fig. IV-2). This property is referred to as the broadcast feature. Therefore, with modifications to the terminal communication hardware, users assigned to one net can transmit and receive over other communication channels. Due to the flexibility of the satellite net, other service disciplines not applicable in the loop system are possibly alternatives. Wherever possible, we make use of the results developed for the computer communication loop system.

Some service disciplines under which net members can share an assigned communication channel that are attractive for a satellite net are:



3-31-76-56

FIGURE IV-1. A Communication Loop System



3-31-76-57

FIGURE IV-2. A Satellite Communication Net

- Contention or Random Access. A message arrives at a particular user, who then transmits all or part of the message on the available link without coordination with the other users in the net. If two or more transmissions overlap, each message or parts of each message may be blocked, and the affected users try again.

- Access with Reservations. At the time of a message arrival, the user requests capacity from the net controller to transmit the message. The controller allocates transmission time to the requesting users according to the chosen queueing discipline. At the appropriate time, the user transmits the message.

- Access with Polling. Users are arranged in a predetermined polling order. Each member is sequentially polled according to that order by a central or distributed net controller. If messages are buffered, the polled user transmits all or part of the buffer contents (this is also predetermined). At the end of transmission, the next member is polled. The polling order can be periodically changed.

- Fixed Assignments. Each member of the group is assigned a specified transmission time that occurs periodically and that is his and only his to use. The users know when and how long they can transmit.

- Various Combinations of the Above. For example, some members have fixed assignments and the other members use access with reservations.

The random-access method is the time-sharing technique based on contention among the users for the communication channel. Examples of systems which utilize the random-access technique are the ALOHA system (Ref. 29) and the computer communication loop with random slot* seizure discipline (Refs. 22-26,** 30). Because terminals in a loop system can ascertain

*A slot is the transmission time of a data unit.

**References 22-26 identify the technique as asynchronous time-division multiplexing with the terminals prioritized by position on the loop rather than random access.

in "real time" whether a data unit has been transmitted in a particular slot, the terminals contend (nondestructively) only for unused slots, which establishes a terminal priority structure according to terminal position on the loop.

In an ALOHA-type system, the user transmissions can overlap in the satellite, resulting in possible errors in the received messages because the propagation delay prevents ascertaining the current message traffic on the channel. The affected messages are considered blocked and are retransmitted. The interference (blocking) errors limit the number of users and the amount of data that can be transmitted. It has been shown (Ref. 29) that for Poisson data-unit/packet arrivals, the average number of transmissions for a successful reception caused by the interference errors becomes unbounded (the system saturates) when the channel utilization* exceeds $1/2e = 0.184$ for unsynchronized access (pure ALOHA) or $1/e = 0.368$ for synchronized access** (slotted ALOHA). For the arrival process of data units/packets to be Poisson, the message arrival process must be Poisson, and each message can contain only one data unit.

Recent results (Refs. 31-33) have indicated that the ALOHA-type system may become unstable even when the channel utilization is smaller than the saturation value due to statistical fluctuations in the arrival process and therefore may require control (Refs. 31, 34, 35) to ensure stable operation. The message throughput of an unstable (uncontrolled) ALOHA-type system decreases with time and eventually becomes zero

---

*Channel utilization is the average arrival rate multiplied by the transmission time of a message.

**Synchronized access is random access starting at specified times.

as the average number of retransmissions tends to infinite, unlike fixed assignments and polling, where the throughput tends to one data unit/slot (further indicating the need of control).

The theoretical results obtained for the ALOHA-type system are not applicable to our assumed message process, which is compound Poisson with respect to data units, nor was it clear how the results* could be extended without resorting to a simulation. Therefore, a quantitative evaluation and comparison of the random-access technique with the other disciplines could not be performed, but the following comments are germane if the utilization resulting in saturation in a slotted system remains at 1/e.

Saturation occurs at a utilization of unity for fixed assignments and polling. A slotted contention system could require up to three times as much capacity as the other techniques. In situations where the cost of communication channels is high or where the channels are a limited resource, the random-access technique does not seem to be attractive because of the maximum allowable utilization. If rather inexpensive communication channels are available, utilization less than 1/e is reasonable. In this context, the ALOHA-type system has been demonstrated to be effective for situations where the average utilization of the channel by a user is very small (long average interarrival times with small message lengths). The random-access technique is also attractive for support of the other

*An abstract of a paper by M.J. Ferguson supported by the ALOHA System at the University of Hawaii became available to the authors at the end of work on this study. Ferguson investigated the performance of an unslotted ALOHA System with exponentially distributed packet lengths and indicated that saturation occurs at a lower utilization (0.136). The result was stated to be optimistic.

techniques (e.g., the technique by which a new user (nonmember) obtains an assignment in a fixed-assignment net and the access technique for the orderwire associated with reservations). Further work (theoretical and/or simulation) is encouraged in establishing the average waiting times and average buffer contents for an ALOHA-type system with compound-Poisson data-unit arrivals to permit a comparison with the other technique.

Three allocation methods are further investigated in this chapter: fixed assignments, polling, and access with reservations. A common characteristic of these disciplines is that the terminal transmissions are coordinated in some manner. This coordination allows the net operation to remain stable for larger traffic intensities than is possible with random access. The fixed-assignments and polling disciplines impose the structure of the communication loop system (Fig. IV-1) on the satellite net by organizing the users in time.

With fixed assignments, the net members are organized in a predetermined time order (Fig. IV-1). The channel is made available in a sequential manner to the net users for data transmission. A user can transmit data for a *specified* time duration when the channel is made available. For example, the first user $T^{(1)}$ can transmit for his preassigned time, say t_1 seconds, and then the channel is made available to the second user $T^{(2)}$ for t_2 seconds, and so on, until the channel is again made available to $T^{(1)}$ for t_1 seconds. The transmission durations for the terminals are predetermined and constant. If a terminal empties the buffered messages during the assigned time, the remaining time is not available to another terminal. The net operates in a periodic manner, the period, or net cycle time, being given by the time difference between the beginnings of two consecutive channel availabilities to one user. The net cycle time is constant. The net operates in a "synchronous" manner, each user knowing when and how long he can transmit.

The "walking" time for fixed allocations is the guard time allocated between consecutive terminal transmissions to ensure that signals from different users do not overlap in the satellite. Terminal ranging to the satellite in conjunction with a time reference will reduce the required guard time. In fixed assignments, the system designer must determine equitable time allocations for the users. The time allocations that will be used are those which minimize either the net buffer contents or the net waiting time. The synchronous time-division multiplexed (STDM) communication loop (Refs. 22, 25) is an example of a net using fixed assignments.

In access with polling, the channel is sequentially made available to the net members, as in fixed assignments, but the channel is assigned to the polled member until his buffer is emptied, and then the process continues with a poll of the next user. The terminal transmission duration is a random process that is dependent upon the buffer contents. The service discipline is still periodic, but the period or cycle time is random. The technique is classified as asynchronous time-division multiplexing. Hub polling (Refs. 22, 27) is an example of the access-with-polling technique. The walking time in polling is proportional to the propagation delay. The next user or the controller must receive an end-of-transmission message from the transmitting terminal before the process continues. Under heavy loading because each buffer is emptied, the terminal transmission durations will be long, resulting in a long average cycle time. Restrictions on the transmission durations could be imposed to reduce the cycle length, yielding other polling service disciplines. The restriction could be random, for example, allowing a terminal to transmit only one message per poll (Refs. 36, 37). More than one poll per net cycle may be required for a terminal with a substantially higher message arrival rate. A deterministic limit could also be used

that is similar to fixed assignments, but with control passing to the next terminal either at the end of a transmission or when the limit is reached. One then has to determine acceptable limit values for the users in a manner similar to finding the time allocations in fixed assignments. The chaining discipline (Ref. 28) for a computer communication loop system is an example of polling with deterministic limits. The previously mentioned walking time may make pure polling with deterministic limits unattractive for a satellite net. The discipline, proposed for the Common User Digital Information Exchange Subsystem (CUDIXS) net, which will use one channel on FLEETSAT, reduces the walking-time effect by using a combination of chaining and reservations. The sequential position of the terminals and the time allocations are changed from cycle to cycle by a net controller using information transmitted by the users during the previous cycle. At the beginning of each cycle, the users are informed when and how long they can transmit. Due to limited resources, we have restricted our attention to the case where the buffer is emptied.

The visualization in Fig. IV-1 is not appropriate for a net operating under access with reservations. The terminal transmissions do not occur in a predetermined manner but are determined by message arrivals and the queueing discipline. When a message originates at a user, the user requests transmission time via an orderwire from a controller. The controller processes these requests and allocates transmission times to the requesting users according to a queueing discipline. At the assigned time, the user transmits the message for which time was requested. The orderwire permits messages to queue in a hypothetical common buffer, from which the controller removes messages by allocating transmission time. Access with reservations is another example of asynchronous time-division multiplexing. Various proposed implementations of this technique are packet reservations (Ref. 38) and the time-division

multiple access with demand access (TDMA/DA) being considered for FLEETSAT (Ref. 39). Access with reservations is a technique that can be used to pool the data nets into a supernet with access to the pooled communication channels on a demand basis. This is equivalent to having multiple servers with a single queue. It is well known from queueing theory (Refs. 17, 40, 41) that a system with a single queue and C servers is more efficient with respect to queue lengths and waiting times than C separate systems, each with one server and one queue. Fixed assignments and access with polling are inherently schemes for time-sharing one channel. Access with reservations conceptually treats messages as entities for the purpose of allocating capacity, while the other two techniques treat the user as an entity.

To summarize, the user in a net with fixed assignments is allocated a set time duration for transmission of data every period, independently of his instantaneous traffic load. This may be inefficient, because sometimes the allocated time duration is more than the user requires, and sometimes it is less than he requires. In access with polling, the periodic channel availability to a user is retained, but the time allocation is random and dependent upon the integrated message load from his lass access. In access with reservations, the user is allocated time according to each generated message.

In a net operating under either fixed assignments or polling, a queue discipline can be applied only at a user and not over the complete net, due to the fact that the channel is assumed to be made available in a fixed, orderly fashion. In access with reservations, a queue discipline can be used for the complete net because messages are treated as entities. In our development we assume that the first-in, first-out discipline is used either in the net for access with reservations or at the terminals for the other service disciplines.

Before proceeding, let us examine the similarities and differences between the queue models generated by the fixed-assignments, polling, and access-with-reservations service disciplines and a typical queue model. In queueing theory, customers requiring some service arrive, join one queue, and are served according to the queue discipline--for our case, first-in, first-out. A customer is completely served before service is provided to the next customer. The model for access with reservations is very similar to the typical model. Customers (messages) from several population classes (terminals) arrive, join the queue at the hypothetical common buffer, and are served (transmitted). In the models for fixed assignments and polling, customers arrive at a user and join the queue in the user buffer, resulting in N distinct queues. The server sequentially moves from queue to queue, providing service (the server takes a vacation with respect to each queue). In fixed assignments, service to each queue is provided for a specified time, during which sometimes a customer (message) may be only partially served and at other times several customers may be served. In polling, service is provided until the polled queue is emptied, and then the server proceeds to the next queue.

In the following sections, the service disciplines are explained in more detail. The equations for the terminal buffer queue lengths and waiting times are presented for each discipline and are evaluated for a specific example. In the examples, the net size and capacity (bit rate) is fixed, while the message arrival process is varied to determine the behavior of the techniques under various channel utilizations. The development of the buffer queue lengths and waiting times is presented in the appendixes. Section IV-B is devoted to a net operating under fixed assignments, and the optimum time allocations are numerically determined. Polling and access with reservations are treated, respectively, in Sections IV-C and

IV-D. A comparison of the disciplines for a fixed net size and capacity is presented in Section IV-E. Section IV-F presents a limited discussion of the problem of capacity allocation to a net. Conclusions are presented in Section IV-G.

B. FIXED ASSIGNMENTS

In a net operating under fixed assignments, the data transmission time is subdivided and allocated to the users in a deterministic manner. For example, the net members could be organized in a cyclic manner similar to Fig. IV-1, the positions indicating the time sequence in which the users have access to the communication channel. The first user $T^{(1)}$ would be allowed to transmit for his preassigned time t_1, then the channel would be made available to $T^{(2)}$ for t_2 seconds, and so on, until the channel is made available to $T^{(1)}$ again. The terminal transmission durations are constant and occur periodically with a fixed period or cycle time. This time allocation is depicted in Fig. IV-3 with a guard time, $t_g = t_w$, allocated to ensure that transmissions from different terminals do not interfere. The cycle time M is the time difference between two consecutive transmission durations for a terminal and, in this example, is identical for all terminals. From a user's viewpoint, the i^{th} terminal is allocated a transmission duration of t_i seconds every period M. These parameters, the time allocation per cycle, t_i, and the time between allocations to a terminal, M, characterize the service discipline and determine the behavior of the buffer and the resulting waiting time for the i^{th} terminal. These parameters are the variables over which the system designer has control. In this example, each terminal has access to the channel once and only once in a net cycle, and the duration of the access is the quantity we have to determine. Note that all the terminals are operating under the same period, which is obviously dependent upon

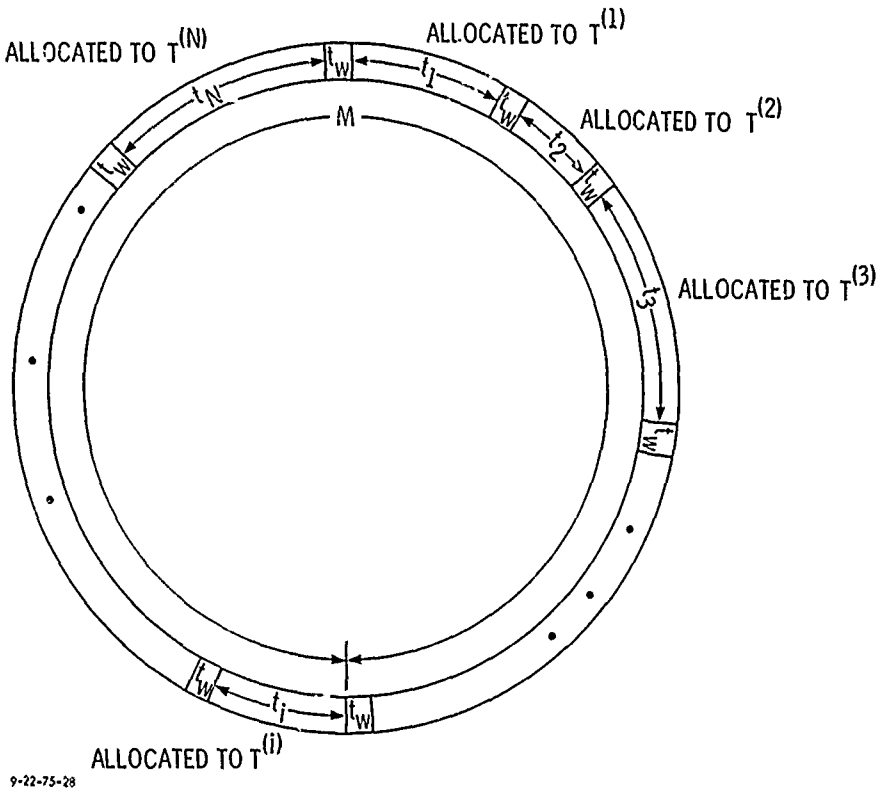


FIGURE IV-3. Time Diagram of Net Operating with an Implementation of Fixed Assignments (Contiguous)

the time allocations. This case is referred to as "fixed contiguous allocations." Another possible implementation of fixed assignments is to allocate equal time for each terminal access but to permit more than one access per net cycle. From the user viewpoint, the i^{th} user is allocated a transmission duration of t_o seconds (the duration is the same for all terminals) n_i times in a net cycle. If the accesses to a terminal are equally spaced in the net cycle, the i^{th} terminal cycle time, which is the time difference between consecutive accesses for the i^{th} terminal, is $M_i = M/n_i$. The terminal cycle times can now be different. Figure IV-4 depicts this implementation, $T^{(1)}$ having access four times in the net cycle with a terminal cycle time of M_1, $T^{(2)}$ having three accesses, and $T^{(N)}$ having

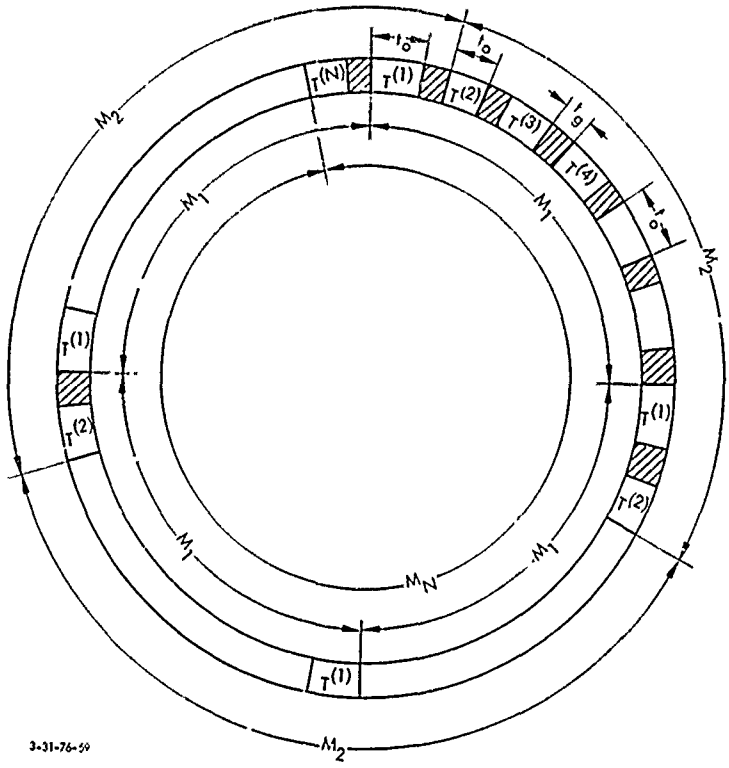


FIGURE IV-4. Time Diagram of Fixed Assignments (Distributed)

only one access. This implementation is labeled "fixed distributed assignments." If the net cycle time and the total time allocated a terminal (i.e., $t_i = n_i t_o$) are the same for both implementations, the terminal is allocated the same amount of time per net cycle, the time allocation occurring in one access (hence contiguous) in the first implementation or occurring in n_i accesses (hence distributed) in the other implementation. In the distributed assignments, the time distribution of the accesses within the net cycle need not be limited to a uniform distribution (equally spaced). In the contiguous implementation, the set of time allocations t_i have to be determined, while in the distributed implementation, the access duration t_o, the number of accesses n_i and the time distribution of accesses must be determined. A terminal in

his access time transmits the buffered message, if any, but always maintains control for his preassigned time (i.e., the unused portion of the allocated time is not available to another terminal). The net operates in a synchronous manner, each user knowing when and how long he can transmit.

Messages arrive randomly at a terminal during the net cycle time, are stored, but can be transmitted only during the terminal's access time. The positions and access times of the terminals are identical from cycle to cycle.

Due to the facts that terminals are allocated a fixed portion of time and that messages are randomly distributed in length, a terminal may transmit part of a message during one access, a complete message at another access, and possibly several short messages at yet another. This implies that treating a message as a unit and calculating the number of messages waiting for transmission is not appropriate. Therefore, messages will be subdivided into data units that are fixed blocks of bits. Each data unit contains the necessary information (e.g., address and signature) to allow the message to be reconstructed at the destination. The number of data units in a message, m, is geometrically distributed with parameter q, i.e., $P(m = k) = (1-q)\, q^{k-1}, 0 \leq q < 1$, and mean length $\bar{m} = 1/(1-q)$. This is a consequence of the assumption that the message length in bits was exponentially distributed. The transmission time for a data unit is defined as a data-unit time slot, $\Delta = B/R$, where B is the number of bits in a data unit, and R is the transmission rate, assumed equal for all terminals. At this point, it is clear that allocated time must be integer multiples of the time slot to ensure complete service to a data unit.

Thus, the arrival process in data units at a terminal is compound Poisson, the times at which messages arrive being

determined by a standard Poisson process [with rate λ_i (messages/slot interval) for the i^{th} terminal], while the message length in data units is geometrically distributed with parameter q_i for the i^{th} terminal. The probability that j data units arrive at a terminal during a time slot is given by (Ref. 22):

$$P_r\left\{X^{(i)} = j\right\} = \begin{cases} e^{-\lambda_i} & j = 0 \\ e^{-\lambda_i} \sum_{k=1}^{j} \binom{j-1}{k-1} q_i^{j-k} \frac{\left(\lambda_i(1-q_i)\right)^k}{k!} & 1 \le j < \infty \end{cases}$$

where $X^{(i)}$ is the random number of data-unit arrivals during a time slot for the i^{th} terminal. The mean μ_i and the variance σ_i^2 of $X^{(i)}$ are $\lambda_i/(1-q_i)$ and $\lambda_i(1 + q_i)/(1 - q_i)^2$, respectively.

In the following sections, both implementations are further investigated. The equations for the terminal buffer queue length and waiting time for an arbitrary allocation vector are presented, but their development is deferred to Appendix D. The allocation vectors which minimize the system performance measures are numerically determined for contiguous allocations. Finally, both implementations are compared.

1. Contiguous Allocations

The behavior of a net operating under fixed contiguous allocations is depicted in Fig. IV-3, the time allocations being integer multiples of the data unit time slot. Recall that the terminals are arranged in a predetermined order (Fig. IV-1), each terminal is allocated one access per net cycle, and the access durations are the variables that the system designer allocates.

The cycle length M must also be expressed as an integer multiple of the basic time slot. In Fig. IV-3, a "walking"

guard time t_w was allocated to ensure that consecutive terminal transmissions did not overlap in time at the satellite. This guard time depends on many factors--for example, on the geographic dispersion of the net users and the availability of a reference time in conjunction with user ranging to the satellite. In addition, time slots could be allocated for new users to indicate the desire to join the net. Depending upon the situation, each terminal transmission may have to be prefaced by a synchronization preamble to allow the destination terminal to receive the message properly. All of the above--guard times, new user slots, and synchronization preambles--will be lumped together into a cycle overhead and will be expressed in time slots. The cycle overhead ϕ is given by

$$\phi = \left[\frac{N}{B}\,(t_w R + P)\right]_+ + n_{new} \ ,$$

where $[x]_+$ is the smallest integer larger than or equal to X, P is the number of bits in the preamble, and n_{new} is the number of slots allocated for new users to request service. By including the synchronization preambles in the cycle overhead, the terminal transmission durations are separated into overhead slots, which are necessary but are not used for removing buffered data, and the slots specifically used for transmitting message data units.

In the development of the equations which relate the buffer contents and waiting time to the data transmission durations, it is necessary to assume specific but arbitrary allocations to the net member. The data allocation vector is assumed to be $(r_1, r_2, \ldots r_N)$, where r_i is the number of data slots per net cycle allocated to the i^{th} user for transmission of the buffered data units. The net cycle length M is given by

$$M = \sum_{i=1}^{N} r_i + \phi \ ,$$

where ϕ is the previously defined cycle overhead. From the i^{th} user's viewpoint, at most r_i data units can be transmitted every M slot. Further, the i^{th} terminal's buffer behavior is independent of the behavior of the other buffers and is only dependent upon his allocation r_i and the cycle time M, which is determined by the cycle overhead and the other allocations. The effective data transmission rate for the i^{th} user is $R_i = r_i R/M$, where R is the bit transmission rate. The sum of the effective data rates, $\sum_{i=1}^{N} R_i$, is less than R because of the overhead. The effective data rate can be viewed as the rate at which messages are served. If the overhead is large, the effective data rate will be small if the number of slots allocated is small, and therefore the tendency will be to force larger slot allocations.

Let $L_{jM+k}^{(i)}$ be the buffer queue length in data units at the i^{th} terminal at the beginning of the k slot in the j^{th} cycle. The equations which relate the behavior of the buffer contents from slot to slot are

$$L_{jM+k+1}^{(i)} = \begin{cases} (L_{jM+k}^{(i)} - 1)^{+} + X_{jM+k}^{(i)} & \text{if the } i^{th} \text{ terminal is transmitting} \\ L_{jM+k}^{(i)} + X_{jM+k}^{(i)} & \text{otherwise} \end{cases} \qquad \text{(IV-1)}$$

where $X_{jM+k}^{(i)}$ is the number of data-unit arrivals during the $(jM+k)^{th}$ slot and $a^{+} = \max(a,0)$, which indicates removal of a data unit only if at least one is stored in the buffer. The set of equations (Eq. IV-1) describes a Markov process that does not have a steady-state solution due to the time-varying transitional probabilities. Because we were not able to obtain a solution for any imbedded process, another assumption was made: Messages that arrive during the terminal's transmission period are gated and prevented from being transmitted during

this transmission period. With this assumption and examining the buffer behavior at the times when control is given to the terminal, we find that the generated process is an imbedded Markov process for which a steady-state solution exists.

Assume that the allocated transmission time for the i^{th} terminal starts at the beginning of the k slot each cycle. The behavior of the buffer queue lengths at these imbedded points (recall that r_i slots are allocated to the i^{th} terminal) is given by

$$L^{(i)}_{(j+1)M+k} = (L^{(i)}_{jM+k} - r_i)^+ + \sum_{n=1}^{M} Y^{(i)}_{jM+k+n-1} \qquad i = 1,2,\ldots,N \quad \text{(IV-2)}$$

These equations are of the same form as those in Ref. 26 developed for a communication loop system with capacity to transmit r_i data units per slot, and hence some of the results are applicable. The other results do not apply because the communication loop assumed a priority structure with respect to the users (random access).

The random variables $\left\{ L^{(i)}_m : 0 \le m < \infty \right\}$ with $m = jM+k$ have been shown to converge in probability provided

$$E\left\{ \sum_{n=1}^{M} X_{jM+k+n-1} \right\} = M\, E\left\{ X^{(i)} \right\} = M\mu_i < r_i$$

for all i (Ref. 26). This stability condition states that the average arrivals per cycle at a terminal must be less than the number of data units that can be transmitted each cycle by that terminal. The derivation of the following results is presented in Appendix C. The average value and the variance of the steady-state buffer queue length $L^{(i)*} = \lim_{m\to\infty} L_m^{(i)}$ are given (Ref. 26) by

$$E\left\{L^{(i)*}\right\} = \frac{1}{2}\frac{M\sigma_i^2}{r_i - M\mu_i} + \frac{1}{2}M\mu_i + \sum_{t=2}^{r_i}\left(\frac{1}{1-\theta_t^{(i)}} - \frac{1}{2}\right) \quad \text{(IV-3)}$$

$$\text{Var}\left\{L^{(i)*}\right\} = M\sigma_i^2 + \frac{1}{12} + \frac{M\mu_{i_3} - r_i^3}{3(r_i - M\mu_i)} + \left\{\frac{M\sigma_i^2 - (r_i^2 - (M\mu_1)^2)}{2(r_i - M\mu_i)}\right\}^2$$

$$+ \sum_{t=2}^{r_i}\left\{\frac{1}{1-\theta_t^{(i)}} - \frac{1}{(1-\theta_t^{(i)})^2}\right\}, \quad \text{(IV-4)}$$

where the $\theta_t^{(i)}$ are the solutions of

$$z^{r_i} - P^{(i)}(z)^M = 0, \quad \text{(IV-5)}$$

with the solutions ordered so that $\theta_1^{(i)} = 1$ and $\theta_t^{(i)} \neq 1$ for $2 \leq t \leq r_2$; $P^{(i)}(z)$ is the generating function for the arrival process at the i^{th} terminal, and $\mu_{i_3} = E\,(X^{(i)3})$. For an allocation of one time slot ($r_i = 1$), the equations are simplified; in Eqs. IV-3 and IV-4, the summation terms are zero. As is evident from Eq. IV-3, the allocation of more than one slot increases the complexity substantially, and one first has to solve for the roots of Eq. IV-5 before obtaining numerical values for the stationary expected buffer queue length.

The equations presented for the buffer queue lengths are independent of the queueing discipline provided that the selection of the next data unit to be transmitted does not depend upon the transmission time. However, the waiting time is dependent upon the queueing discipline, which was assumed to be first-in, first-out. The buffer queue length is expressed in data units and not messages, and, if Little's theorem* were used, the average waiting time for a data unit would be obtained. Because messages can consist of many data

*Relates the waiting time to the average queue length by the average arrival rate.

units, the average data-unit waiting time is not appropriate as a measure of the waiting time for a message. It is natural in a queueing system with batch arrivals to employ the notion of a virtual customer to measure the waiting time. This is the virtual delay and will be the waiting-time measure used for messages. The methodology used in determining the virtual delay parallels that in Ref. 25.

The cycle with respect to the i^{th} user is initiated at the beginning of his allocated transmission duration. In other words, the cycle starts with the i^{th} user able to transmit r_i data units, and he then must wait $M-r_i$ slots for the channel to be made available again. The virtual message, consisting of m data units, arrives randomly in the j^{th} cycle and joins the queue at the i^{th} user. The virtual delay $D_m^{(i)}$ for this message is the difference between the average queueing time (total time in the system) and the message transmission time.

The queueing time consists of:

1. Waiting for the start of the next transmission allocation to the i^{th} terminal.
2. Waiting for the number of the slots need to transmit the already buffered data units.
3. The slots needed to transmit the virtual message.

Unfortunately, the steady-state virtual delay could not be obtained but the following upper and lower bounds are developed (Appendix D);

$$D_\ell^{(i)} \leq D_m^{(i)} \leq D_u^{(i)} \qquad \text{(IV-6)}$$

with

$$D_u^{(i)} = \frac{M+1}{2} + \frac{M}{r_i}\,\bar{Q}^{(i)} + \left(\frac{M}{r_i} - 1\right) m \qquad \text{(IV-7)}$$

and

$$D_\ell^{(i)} = \frac{M+1}{2} + \frac{M}{r_i}\,\bar{Q}^{(i)} + \left(\frac{M}{r_i} - 1\right)\left(m - r_i\right)^+ \qquad \text{(IV-8)}$$

where $\bar{Q}^{(i)} = E\left\{L^{(i)*}\right\} - \frac{M+1}{2}\,\mu_i$ is the average number of data units buffered at the arrival of the virtual message. The first term of both the upper and lower bound is the average waiting time for service to begin again from the virtual message arrival, the second term is the time to transmit the already buffered data units, and the third time is a measure of the waiting time to transmit the complete virtual message due to the periodic availability of the channel.

The difference between the upper and lower bound values, which is a measure of how tig the bounds are, is given by

$$D_u^{(i)} - D_\ell^{(i)} = \left(\frac{M}{r_i} - 1\right) m - \left(\frac{M}{r_i} - 1\right)\left(m - r_i\right)^+ = \begin{cases} \left(\frac{M}{r_i} - 1\right) m & \text{if } m \le r_i \\ M - r_i & \text{if } m > r_i \end{cases}$$

It is dependent upon the relative length of the virtual message with respect to the allocation, but the largest difference is $M - r_i$, which is exactly the waiting portion of the cycle.

For the case of allocating one data slot (i.e., $r_i = 1$), the virtual delay can be exactly determined. The virtual delay with $r_i = 1$ is equal to the lower bound, Eq. IV-8, r_i being set to unity.

Having obtained expressions for the terminal average buffer contents and virtual delays, one can now write expressions for the chosen system performance measures of Section

IV-A, replacing the average net waiting time by the net average of the *upper bound* on the virtual delay

$$D = \frac{1}{N} \sum_{i=1}^{N} D_u^{(i)} , \qquad \text{(IV-9)}$$

$D_u^{(i)}$ being given by Eq. IV-7. The average net buffer length is

$$L = \frac{1}{N} \sum_{i=1}^{N} E\left\{ L^{(i)} \right\} , \qquad \text{(IV-10)}$$

the average value of the terminal buffer queue length $E\left\{ L^{(i)} \right\}$ being given by Eq. IV-3. Unfortunately, analytic expressions for the allocation vectors that minimize either performance measure, Eqs. IV-9 or IV-10, could not be obtained. The optimum allocation vectors depend upon the assumed values for the net overhead and the parameters of the message statistics. Because of this, the optimum allocations are numerically determined for several cases. The assumptions and results are presented in the following section. The allocations which minimize the delay upper bound are not necessarily equal to the optimum allocations for the delay.

Numerical Results. This section presents the numerical evaluation of the allocation vectors which minimize either system performance measure. The parameter values used in the numerical examples are suggested by a Common User Digital Information Exchange Subsystem (CUDIXS) net time-sharing one channel of FLEETSATCOM. The average message length is assumed to be identical for all users. Further, it is assumed that two classes of users exist, and that the message arrival rate is identical for all users in a class. The parameter values are presented in Table IV-1.

TABLE IV-1. ASSUMED PARAMETER VALUES FOR THE SATELLITE COMMUNICATION NET

Transmission Rate (R):	2.4 kb/s
Number of Users (N):	10 users in class 1 1 user in class 2
Average Interarrival Times (T):	27 — 10 minutes for class 1 or 2.25 — 0.83 minutes for class 2
Data Unit Length:	608 bits
Average Message Length ($\bar{m}$):	26 data units
Virtual Message Length (m):	26 data units

The class 1 users represent the ships in the net, while the class 2 user is the shore station (NAVCOMSTA) that acts as the net controller. The average interarrival time between messages at the class 2 user is twelve times that of a class 1 user, which is representative of the larger volume of shore-to-ship message traffic. The optimizations are performed for two overhead cases (Table IV-2) and for each case with both sets of interarrival times. The transmission overhead is the overhead associated with each separate transmission and consists

TABLE IV-2. OVERHEAD VALUES

Case	Transmission Overhead[a]	New Entry Request Time
I	0.39 seconds[b]	Proportional to cycle length (15/120)
II	0.0	

[a]Preamble transmission time plus guard time.
[b]Transmission rate R = 2.4 kb/s.

of the synchronization preambles which preface a terminal transmission to provide synchronization information to the receiving terminal and the guard time allocated between consecutive transmissions from different terminals to ensure that the transmissions do not interfere. The new entry request time is the time allocated for requests from users who are not members of the net to participate in the net by sharing the communication channel. The request time is utilized by nonmember users on a random-access basis and is proportional to the total cycle time, with 15 seconds of every 2 minutes of cycle time set aside. The duration of the "request to join" message and the corresponding blocking probability were not investigated. The request time should be chosen on the basis of the expected number of new users. The second overhead case is based on a net operating in a synchronous mode with reference timing provided by the controller and with user ranging to the satellite. The net average of the stationary expected terminal buffer lengths, given by Eq. IV-9, is referred to simply as the net buffer length, and the net average of the upper bound on the terminal virtual delays, given by Eq. IV-10, is referred to as the net delay. The results are expressed in data units for the net buffer contents and in minutes for the net virtual delay.

It should be pointed out that we are not trying to simulate or evaluate the performance of a CUDIXS net but are merely using the statistics instead of inventing statistics.

Two utilizations can be defined for a net operating under fixed assignments:

1. A terminal utilization $M\mu_i/r_i$ that must be less than unity if the terminal buffer contents are to remain finite.

2. A channel utilization that is the sum of the average arrivals during a slot in the net, $\sum_{i=1}^{N} \mu_i$, divided by the number of data units that can be served in a slot (in this case, one). The channel utilization is independent of the slot allocations and cycle length. Table IV-3 presents the channel utilization for the parameter values of Table IV-1.

TABLE IV-3. CHANNEL UTILIZATION

Interarrival Times, min	Channel Utilization
$T_1{}^a = 27$ $T_2 = 2.25$	0.09
$T_1 = 10$ $T_2 = 0.833$	0.24

[a] T_i is the average interarrival time for a class i user.

The numerical technique used for finding the solutions of Eq. IV-5 was Newton's method for determining the complex zeroes of a function. It became evident that the summation term of Eq. IV-3, involving the solutions θ_t, had a negligible effect on the terminal and net buffer lengths in most instances. Therefore, the optimizations were performed neglecting the summation term, which tremendously simplified the numerical procedure. Analytical solutions for the allocation vectors still could not be determined even when the summation term was neglected. The summation term was evaluated occasionally to ensure that the approximation was still valid.

The numerical search for the optima was restricted to the case where allocations to a class were equal. Although this

restriction may exclude the global optimum, it was not clear how one decided which selected users in a class should be allocated more time slots. Therefore, the optimization was conducted in two dimensions (r_1, r_2), the respective allocations to the terminals in class 1 and 2, instead of performing an eleven-dimensional search. We quickly discovered that both system measures, net buffer length and delay, could have several, if not many, local minima, and hence numerical search procedures were ruled out. The optimization was performed in two stages: (1) the optimum allocations were determined for a constrained cycle length and (2) the constrained minimum values were then compared to determine the global optima.

Overhead Case I. The overhead values assumed for this case are presented in Table IV-2. The numerical evaluation is performed for both sets of interarrival times and the other parameter values of Table IV-1 with the channel utilizations presented in Table IV-3. The slot allocation is r_1 slots per cycle for a class 1 user and r_2 slots per cycle for a class 2 user. With a cycle-length M constraint, the class 2 user allocation is related to the class 1 user allocation by $r_2 = M - \phi - 10r_1$. Examples of the behavior of the system performance measures under various cycle-length constraints are presented in Figs. IV-5, IV-6, and IV-8. The behavior of the net virtual delay (upper bound) and net buffer contents at the respective constrained optimum allocations is exemplified in Fig. IV-7 and Table IV-4, respectively. Finally, the optimum allocations for the net buffer contents and net delay are presented in Tables IV-5 and IV-6, respectively.

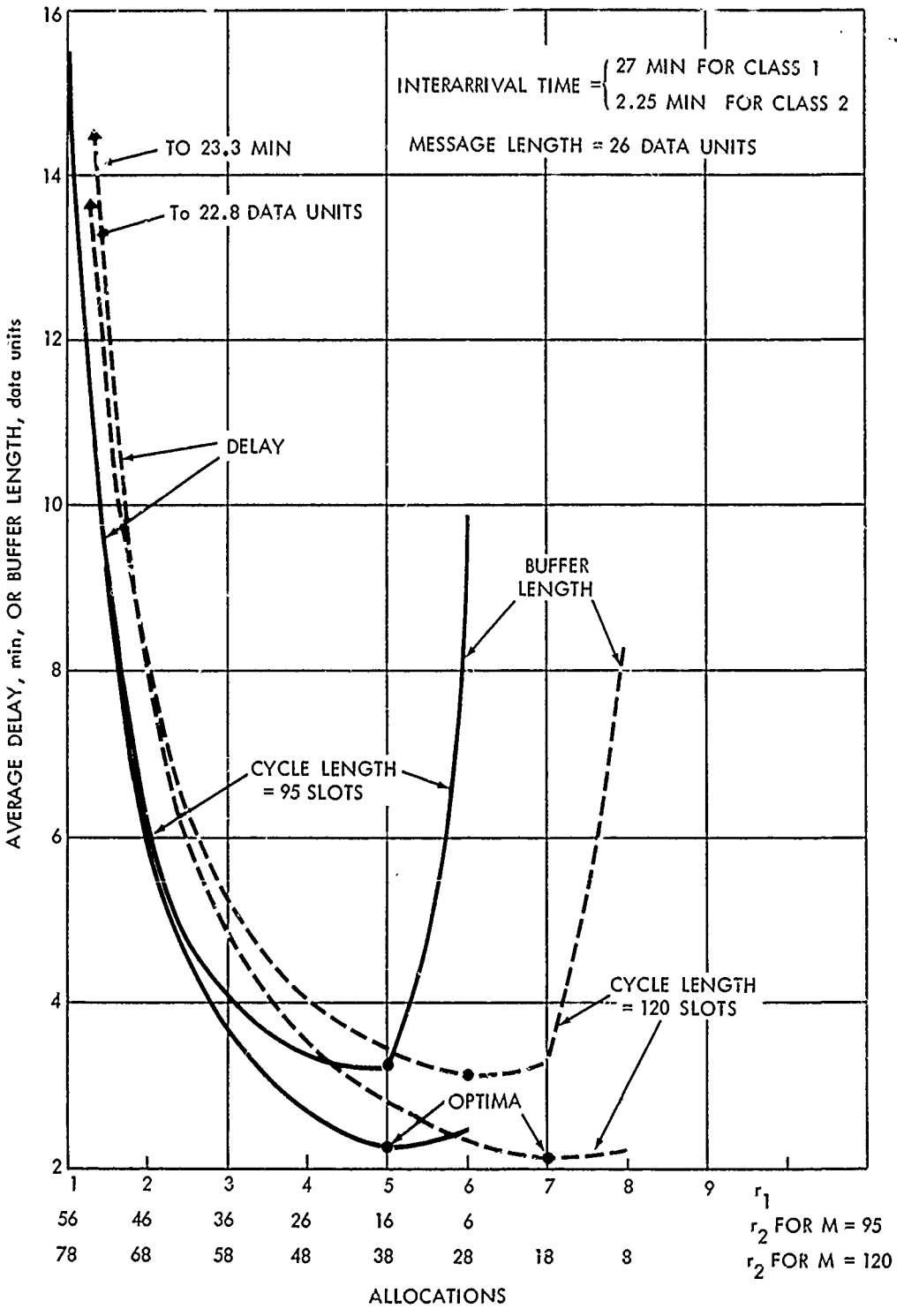


FIGURE IV-5. Optimization with Equal Allocations to a Terminal Class for Fixed Cycle Duration

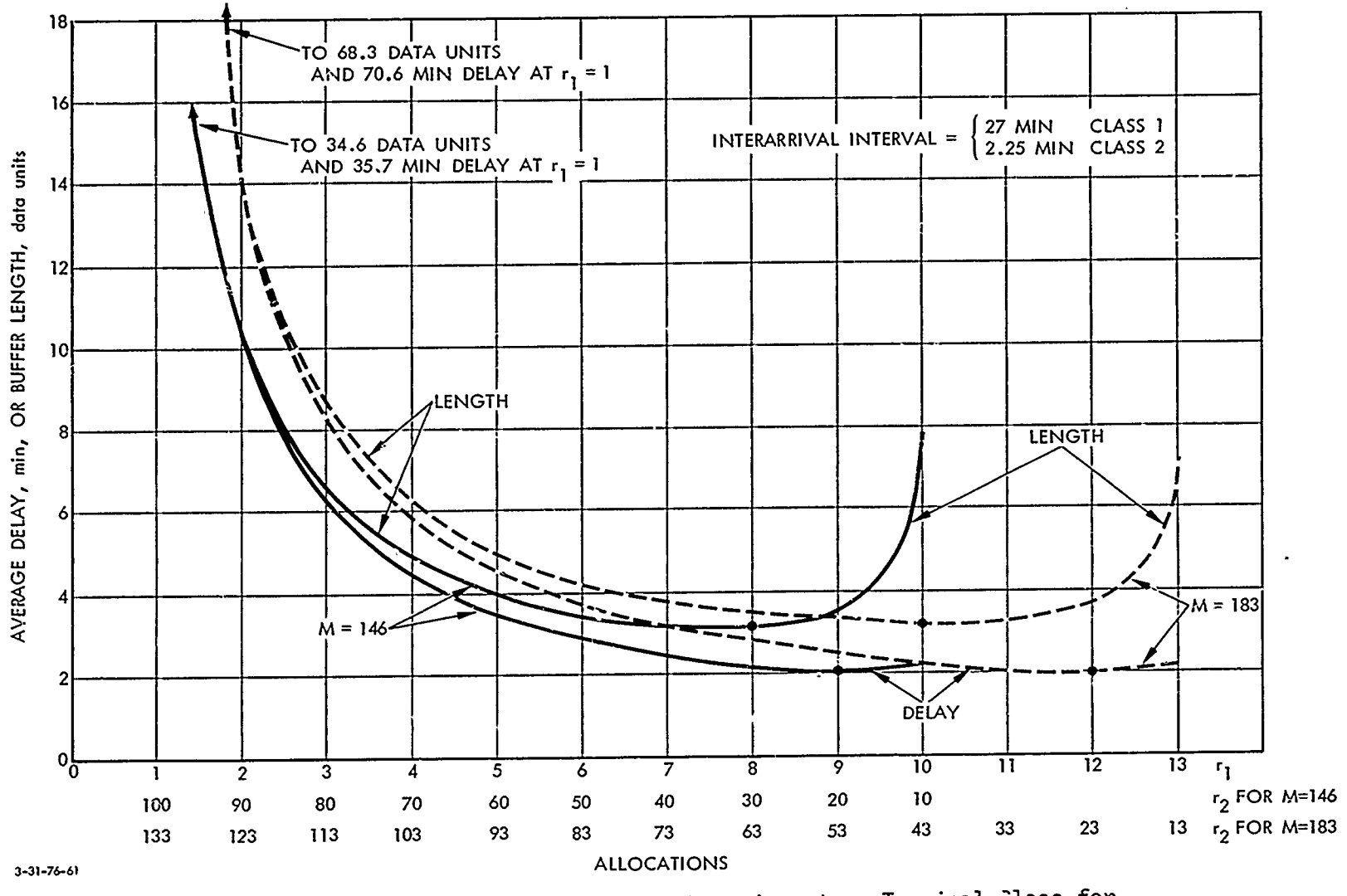


FIGURE IV-6. Optimization with Equal Allocations to a Terminal Class for Fixed Cycle Length

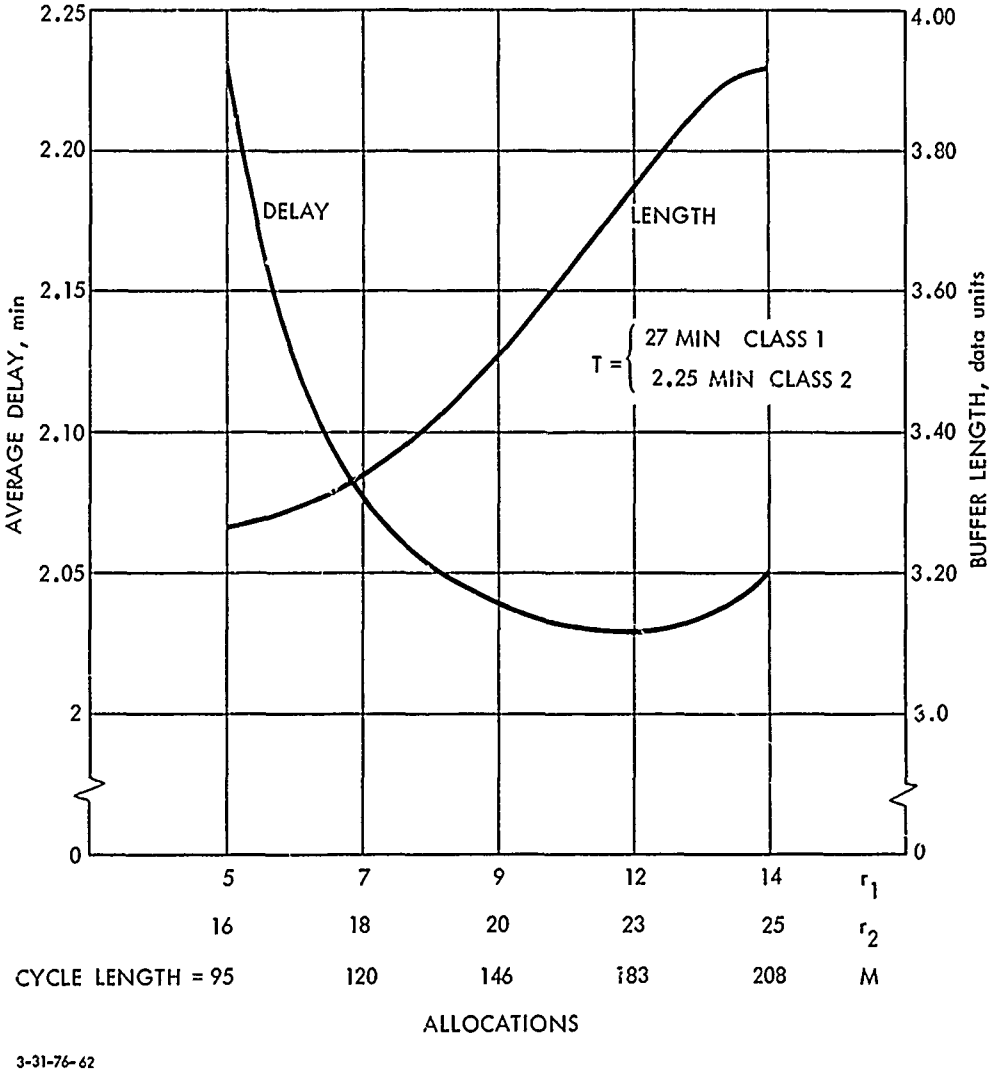


FIGURE IV-7. Optimization with Equal Allocations to a Terminal Class

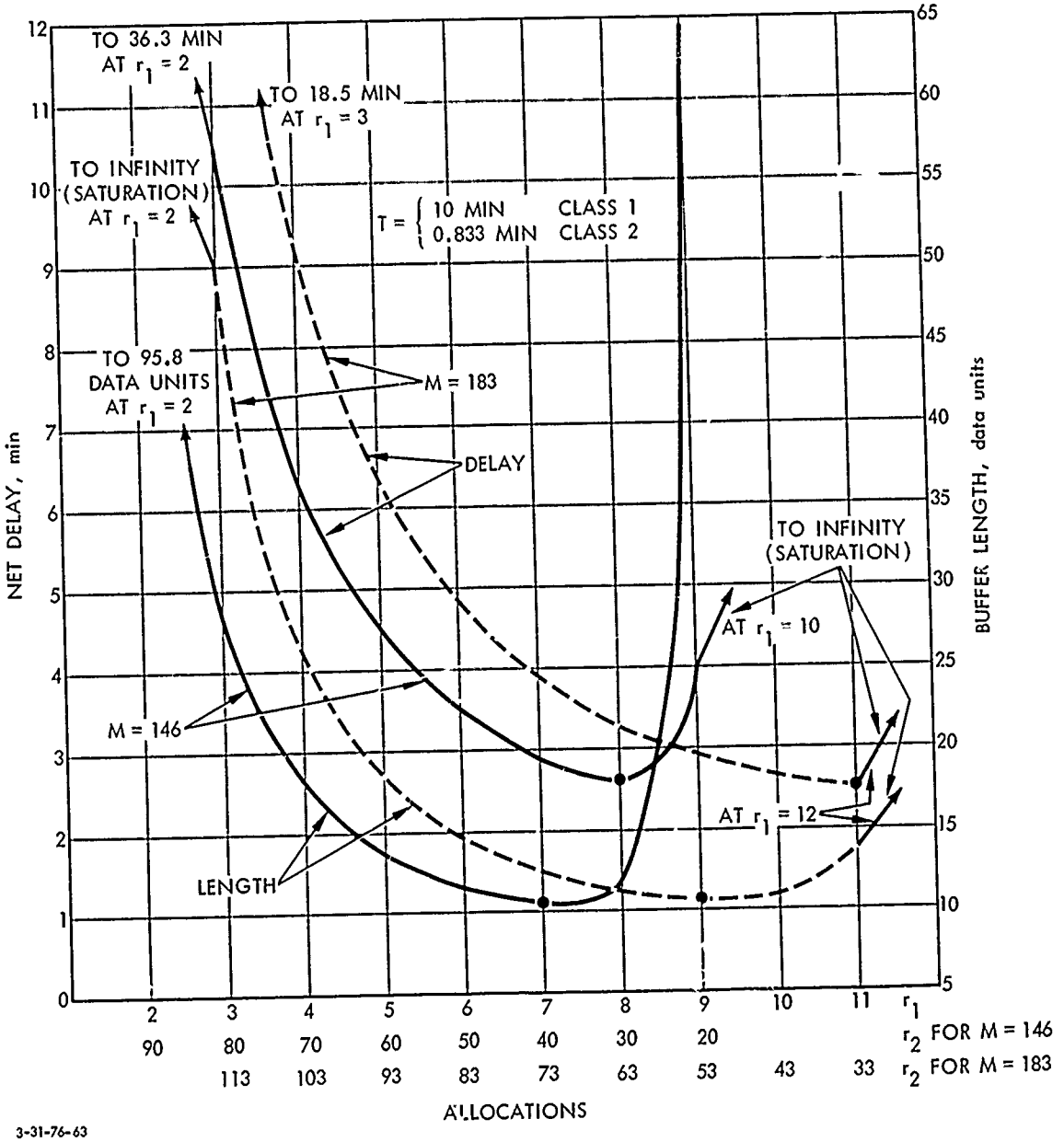


FIGURE IV-8. Optimization with Equal Allocations to a Terminal Class for Fixed Cycle Length

TABLE IV-4. CONSTRAINED OPTIMUM ALLOCATIONS FOR NET BUFFER CONTENTS, T_1 = 10 MINUTES AND T_2 = 0.83 MINUTES

Allocations, slots			Net Buffer Contents, data units	
M	r_1	r_2	A[a]	B[b]
120	6	28	10.84	> 10.84
146	7	40	10.63	10.83
183	9	53	10.60	10.93
208	11	55	10.66	> 10.93

[a] A: neglecting summation term.

[b] B: including summation term.

TABLE IV-5. OPTIMUM ALLOCATIONS FOR NET BUFFER CONTENTS

Interarrival Time, min		Allocation, slots			Net Buffer Contents, data units	Net Delay, min
T_1	T_2	r_1	r_2	M		
27	2.25	6	28	120	3.15	2.37
10	0.83	7	40	146	10.83	2.96

TABLE IV-6. OPTIMUM ALLOCATIONS FOR THE NET DELAY UPPER BOUND

Interarrival Time, min		Allocation, slots			Net Buffer Contents, data units	Net Delay, min
T_1	T_2	r_1	r_2	M		
27	2.25	12	23	183	3.75	2.03
10	0.83	11	33	183	13.5	2.5

The following observations and conclusions are obtained from the numerical computations:

1. The respective minima for the net buffer contents and net delay bound occur at different allocations. In comparing the minimum value for one measure to the corresponding value with the optimum allocations for the other measure, the differences are not large (limited to a 25 percent increase). The effects on the individual terminals are more pronounced (Table IV-7), the optimum allocation tending to equalize that performance measure for the class 1 and 2 users.

TABLE IV-7. USER BUFFER CONTENTS AND VIRTUAL DELAYS, T_1 = 10 MINUTES AND T_2 = 0.83 MINUTES

User	Buffer Contents, data units		Virtual Delay, min	
	A[a]	B[b]	A[a]	B[b]
Class 1	8.4	6.7	3.2	2.5
Class 2	35.5	81.2	1.0	2.5

[a]A: with optimum allocations for net buffer contents.

[b]B: with optimum allocations for net delay bound.

2. The system performance measures are sensitive to the allocations (Fig. IV-8). Substantial reductions in the performance measure values are possible by determining the optimum allocations. For example, an intuitive approach would allocate twelve times the class 1 user allocation to the class 2 user ($r_1 = 5$, $r_2 = 60$, and $M = 146$) and would result in a net delay of 4.5 minutes, which is about twice the optimum delay value.

3. The behavior of the system performance measures when examined at the constrained optima reveals a broad minima region (Fig. IV-7 and Table IV-4), but when it is examined for a fixed cycle length it is similar to a valley with high rising sides that broadens with increasing cycle length (Figs. IV-5 and IV-6). This demonstrates the sensitivity of the performance measures to the allocations.

4. The virtual delay appears to be dominated by the virtual message length (26 data units). In Table IV-6, the message load increased from the lower arrival rate to the higher rate by about 170 percent, resulting in a 260 percent increase in the buffer contents but only a 23 percent increase in the net delay.

5. The approximation to the system performance measures by neglecting the summation term $\frac{1}{2}\sum_{k=2}^{r_i} \frac{1+\theta_k^{(i)}}{1-\theta_k^{(i)}}$ is valid for a large range of parameter values (Table IV-8), while the effect on the buffer contents of the class 2 user is more pronounced, and for large allocations the approximation is invalid with respect to the class 2 user. Even when the

approximation is valid, the summation term can affect the optimum allocation with respect to buffer contents (Table IV-4).

TABLE IV-8. EFFECTS OF THE CLASS 2 USER SUMMATION TERM ON THE CLASS 2 USER AND NET BUFFER CONTENTS, T_2 = 0.83 MINUTES

Class 2 User Allocation, slots	Summation Value for Class 2 User, data units	% Increase in Buffer Contents	
		Class 2 User	Net
33	2.4	3	2
53	3.6	10	3
113	6.4	34	1

Overhead Case II. In the previous overhead case, the slot allocations (Table IV-6) that result in the minimum net delay for both arrival rates occur at a cycle length of 183 slots. The net overhead is equal to 40 slots or 22 percent of the cycle length, leaving 78 percent of the slots for data transmission. The net overhead with different cycle lengths consumes 27 percent and 25 percent of the slots for M = 120 and 146, respectively. Because the overhead slots consume a larger percentage of the smaller cycle lengths, the optimum occurs with a longer cycle length. The synchronization preambles could be reduced by providing common bit timing, and the guard time could be reduced by having the users determine their respective ranges to the satellite. The effects of reducing the overhead are examined by considering overhead case II (Table IV-2), which assumes that the users operate in a synchronous manner requiring no preamble or guard time. New entry request time is still provided in a manner identical to that in the previous case.

The values of the system performance measures with a fixed cycle length are plotted in Fig. IV-9 for the lower arrival rates. Figures IV-10 and IV-11 are the plots of the net buffer contents and net delay at the allocations which result in the constrained minima for the net delay bound for the two sets of arrival rates, respectively. The behavior of both performance measures in Fig. IV-10 is erratic. The erratic behavior is probably caused by rounding the new entry request time to integer values of data slots so that the cycle lengths in slots are integers. For the lower arrival rate (Fig. IV-10), the optimum allocation could not be determined because the delay value with $r_1 = r_2 = 1$, which is exact, is less than the upper bound values but greater than the lower bound values for larger allocations.

The optimum allocations and the corresponding values for net buffer length and delay are tabulated in Table IV-9 for both overhead cases and interarrival rates. The reduction in the net buffer contents or delay is at most 27 percent. This is probably a consequence of the selected arrival rates, and it is felt that the effects will be more dramatic for higher arrival rates, where the average number of arrivals during the overhead slots will be significantly larger. As was expected, the allocation size and cycle length for optimal performance are significantly reduced (e.g., the minimal reduction was about half, the cycle length decreasing from 120 slots to 51 slots). For the higher arrival rate of Table IV-9 and the allocations which minimize the net delay, the effective data transmission rate was increased to 2.1 kb/s from 1.9 kb/s by reducing the transmission overhead.

2. Distributed Allocations

In the previous section, the channel was made available to each user only once per net cycle. The users were allowed

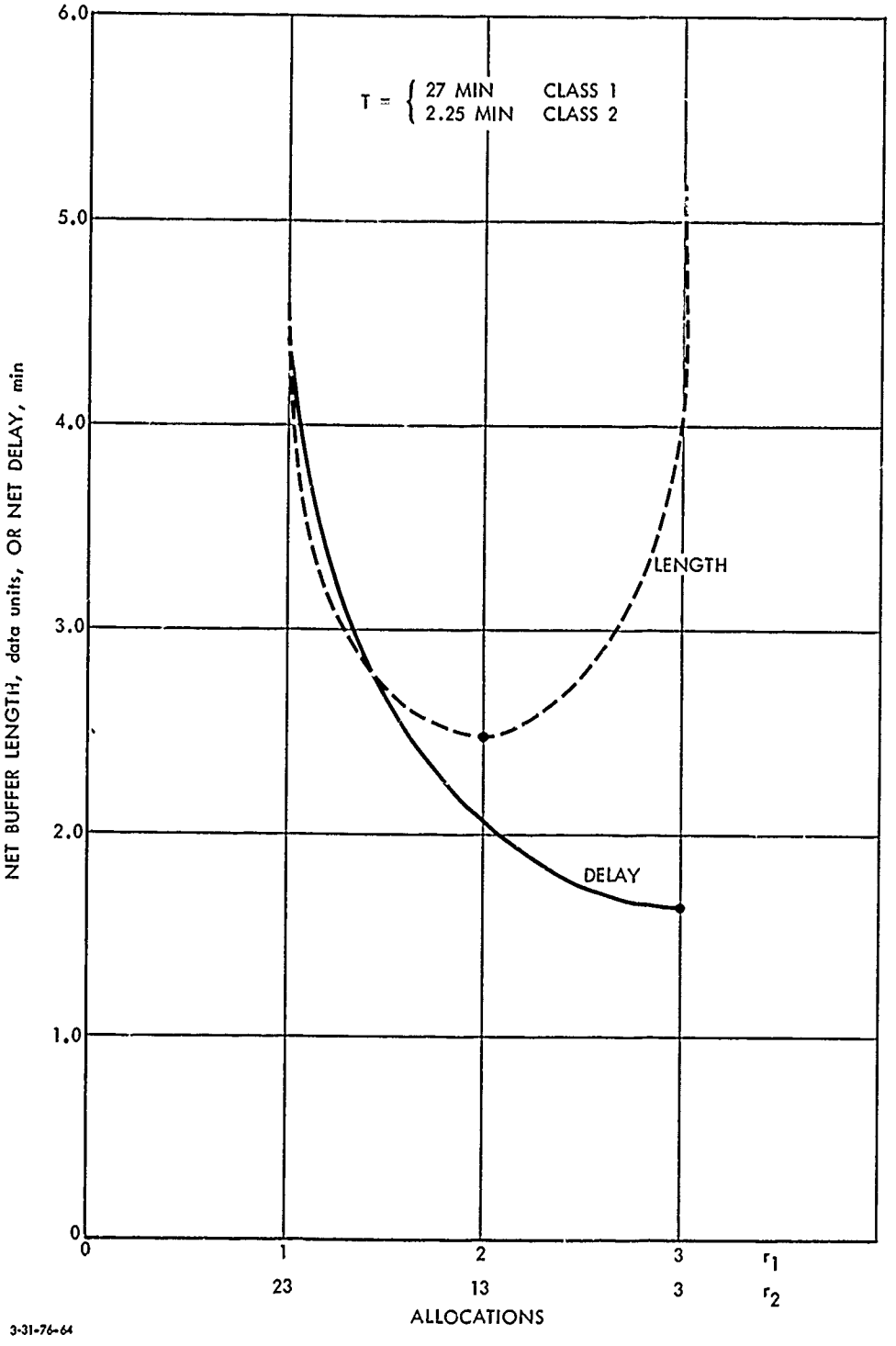


FIGURE IV-9. Optimization for Fixed Cycle Length (M = 38) for Overhead Case II

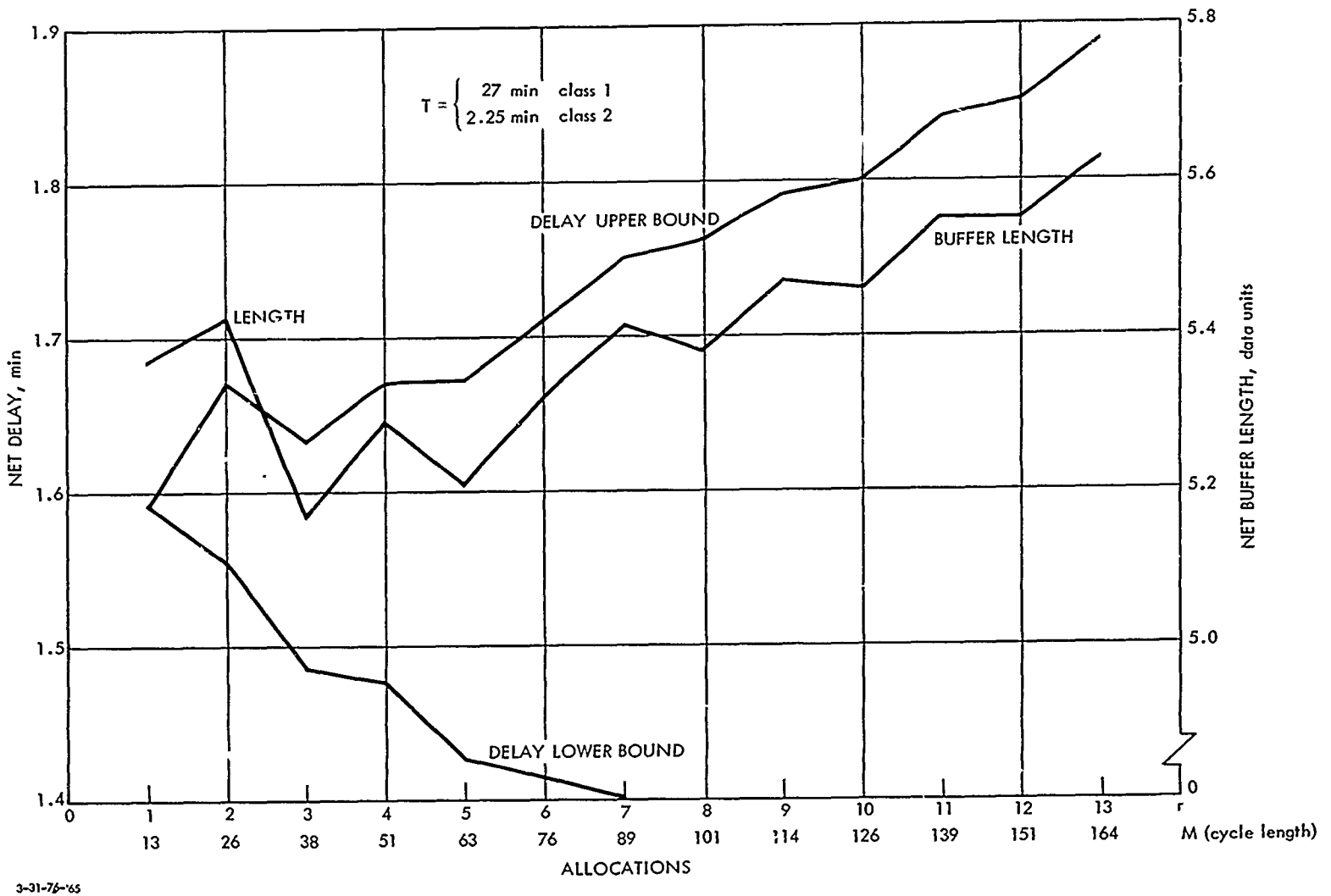


FIGURE IV-10. Equal Allocations for Overhead Case II

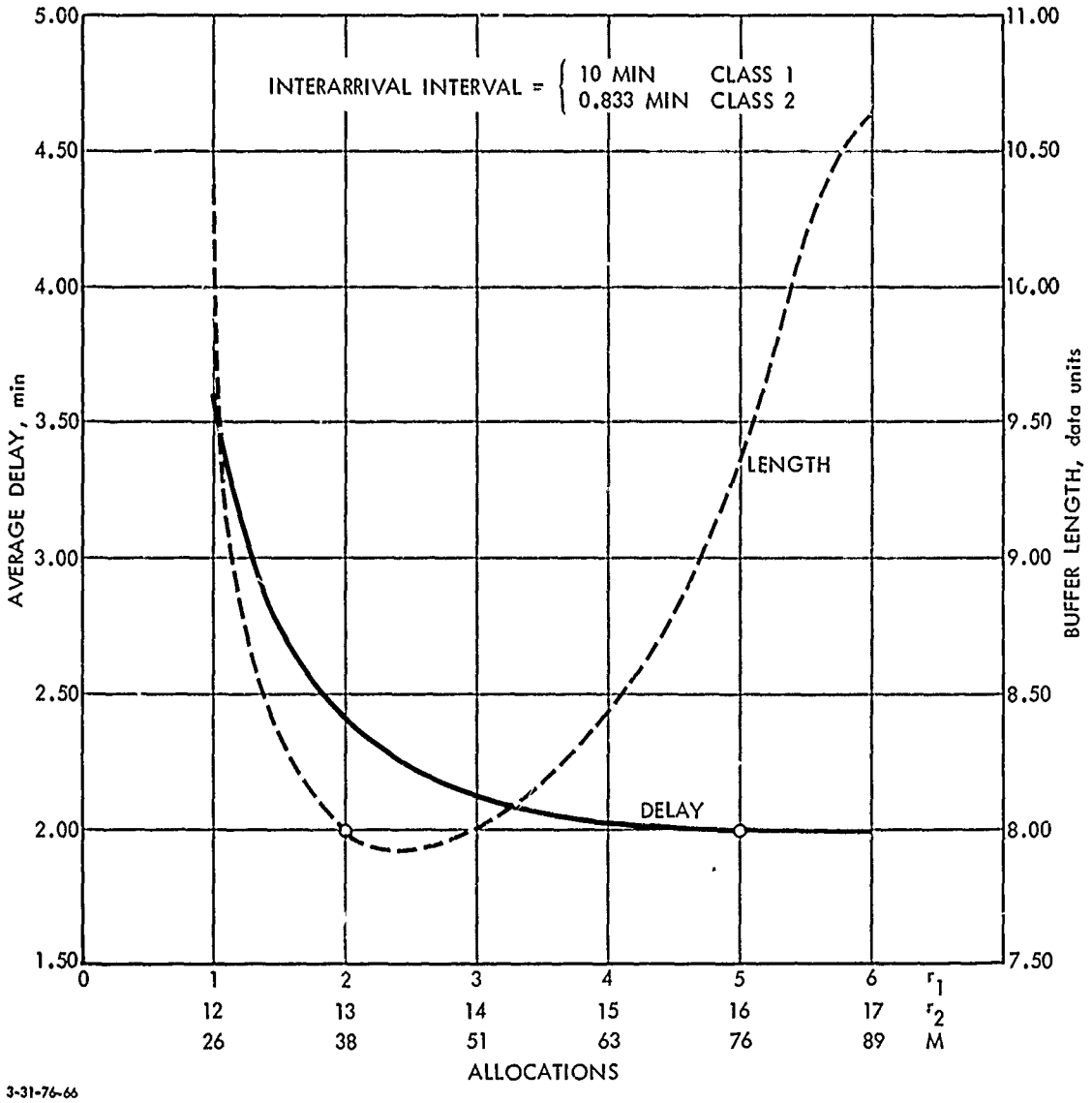


FIGURE IV-11. Optimization with Equal Allocations to a Terminal Class for Overhead Case II

TABLE IV-9. EFFECTS OF OVERHEAD

	Delay, min		Allocation, slots		Buffer Length, data units		Allocation, slots	
Interarrival Times, min	I[a]	II	I	II	I	II	I	II
T_1 = 27 T_2 = 2.25	2.03	1.59[b]	r_1=12 r_2=23 M=183	r_1=1 r_2=1 M=11	3.15	2.42	r_1=6 r_2=28 M=120	3 14 51
T_1 = 10 T_2 = 0.833	2.5	1.99	r_1=11 r_2=33 M=183	5 16 76	10.8	7.9	r_1=7 r_2=40 M=146	2 12 37

[a]Refers to the overhead case.

[b]The minimum was not determined.

to transmit for different time durations (i.e., r_i contiguous time slots once per cycle) to accommodate various user requirements (the user message load). Essentially, the user is allowed to transmit at most r_i data units in a time period of duration M. There are numerous ways, other than removing the units consecutively in time, of accomplishing this. For example, if the i^{th} users were allocated 6 time slots, Fig. IV-12 demonstrates two ways of accomplishing this. One technique is an example of contiguous allocations, where the user is allowed to transmit 6 data units consecutively and then waits M - 6 slots to transmit again. The other technique is an example of distributed allocations, where the user is allowed to remove one data unit at the beginning of the cycle, waits n_1 slots, transmits 3 data units, waits n_2 slots, transmits 2 units, and finally waits n_3 slots for the cycle to begin again. Obviously, there are many ways of distributing the allocations, contiguous allocations being a special case of distributed allocations. Distributed allocations permits the terminals to access the satellite more than once in a net cycle. Figure IV-4 depicts the cyclic behavior of the net for the case where each access is of equal duration, r slots, and user 1 is allocated four accesses, user 2 three accesses, and the N^{th} user one access.

In Fig IV-12, an implicit assumption was made that distributing the allocations did not change the cycle length. This is obviously false if each terminal transmission must be prefaced by synchronization preambles and if transmissions from different terminals must be separated by guard times. As a matter of fact, with n_i access allocated to the i^{th} user, the transmission overhead for the i^{th} terminal has been increased by a factor of n_i. Therefore, the net cycle time has been increased, resulting in a decrease in the effective data transmission rate. On the other hand, if the net is operating under

overhead case II (i.e., no transmission overhead), the cycle times and effective data transmission rates for both techniques are identical.

For the case where the cycle times are the same, it is not obvious why distributed allocations are being investigated, since under either technique r_i data units can be transmitted over a period of M slots. Figure IV-13 presents the behavior of the buffer contents under the contiguous and distributed allocations outlined in Fig. IV-12. The cycle is initiated at the point when the user has just used his r_i slots in the contiguous case. The buffer contents are assumed to be identical at this point because the arrivals and the number of units transmitted are the same. In the example, three messages arrive with lengths 1, 2, and 3 data units, resulting in a total of 6 data units, which is equal to the number removed. The buffer contents, as expected, are identical at the end of the cycle, but the dynamic behavior is different. The cycle time for a user is subdivided into a "waiting" or inactive time, when other terminals are transmitting, and an active or transmitting time. During the waiting time, the buffer contents can be increased by message arrivals but cannot be reduced, and therefore the buffer queue length is monotonically increasing. The contents can be reduced only during the transmission time. (The contents could also be increased by a message arrival.) In contiguous allocations, the subdivision results in two intervals, each with consecutive slots (e.g., the first M-r slots form the waiting period, and the active period occurs during the last r slots of the cycle). In distributed allocations, the active or inactive period consists of disjoint intervals. If the buffer contents are examined at any time within the cycle, the contents with distributed allocations will be less than or at most equal to the contents with contiguous allocations, because more data units could

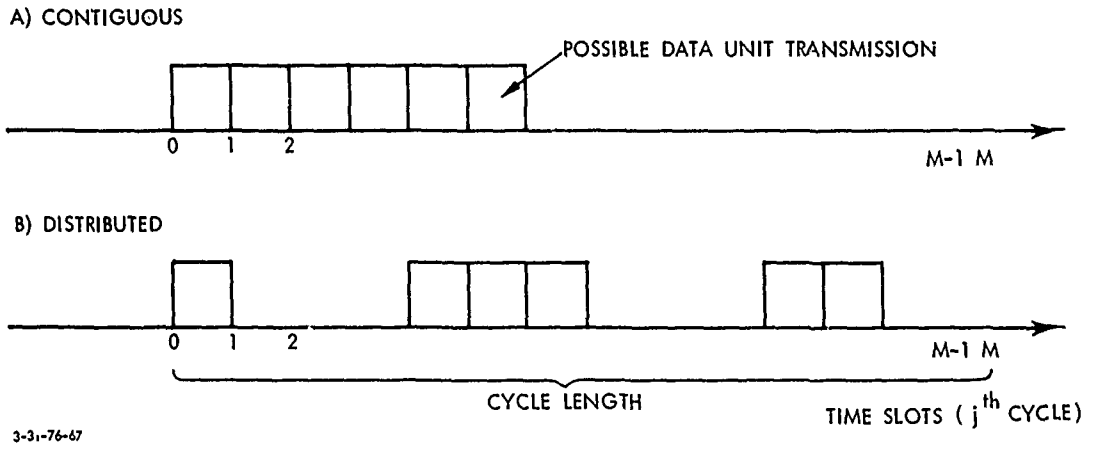


FIGURE IV-12. Time Sequence of When Data Units Can be Transmitted by the i^{th} User

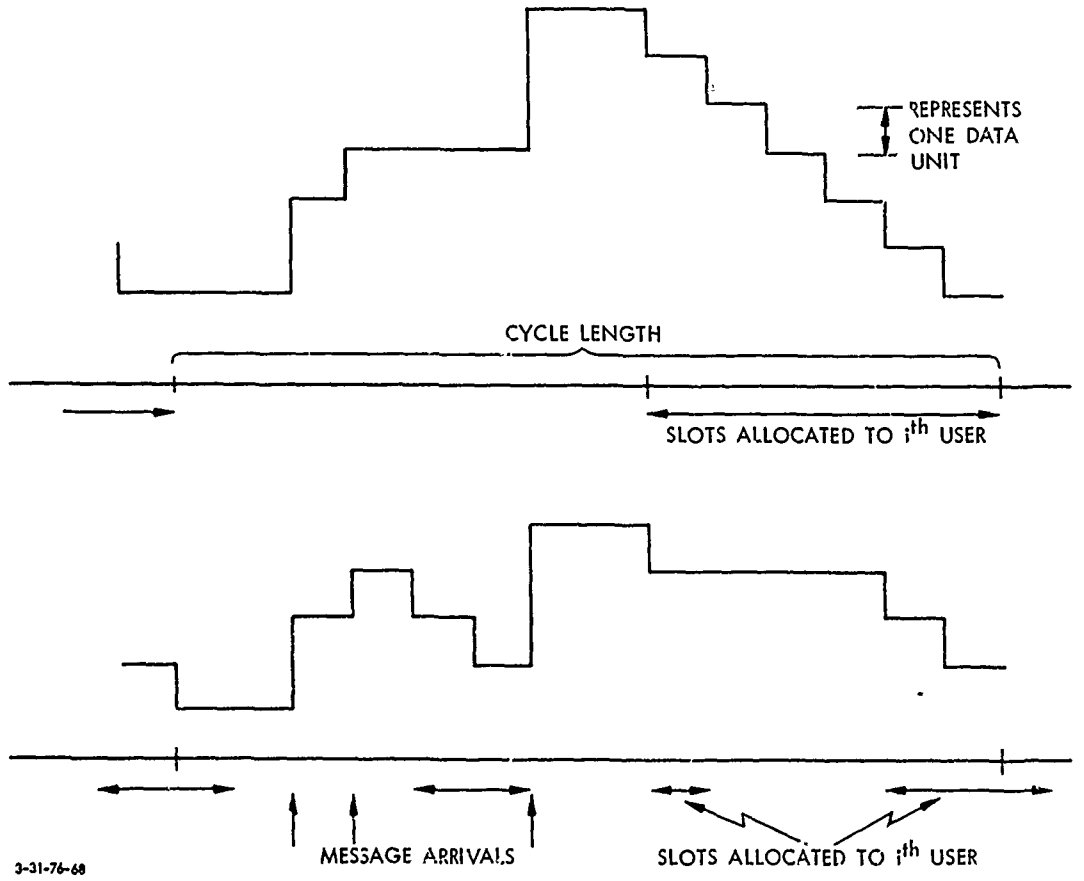


FIGURE IV-13. An Example of the Buffer Contents with Contiguous and Distributed Allocations

have been transmitted in the distributed case. Only at the end of the cycle will the number be the same, resulting in the same buffer contents at this point. Therefore, the average of the buffer contents over a cycle will be smaller in the distributed case. Further, since the waiting time is dependent upon the buffer contents at the time of the arrival, it is expected that the waiting time will also be smaller in the distributed case. In the example, the buffer contents at any point in distributed allocations was strictly smaller than the contents in contiguous allocations, except in the last three time slots, where they are equal. The cycle average of the buffer contents is 4.7 data units for contiguous allocations and 2.6 data units for distributed allocations.

The investigation of distributed allocations is primarily limited to the case where the terminals are allocated one data slot at each access. In the development of the equations, the allocations are assumed to be $n_1,\ldots n_i,\ldots n_N$, where n_i is the number of accesses allocated to the i^{th} user, each access lasting one data-unit time slot. The overhead in bits associated with each transmission is given by $t_w R + P$, where P is the number of bits in the synchronization preamble, t_w is the guard time between transmissions, and R is the data transmission rate. The overhead slots are again separated from the data slot allocations and are lumped together into the net overhead. For noncontiguous terminal accesses, the cycle or net overhead is given by

$$\phi_d = \left[\frac{(t_w R+P)}{B} \sum_{i=1}^{N} n_i \right]_+ + n_{new} ,$$

where n_{new} is the number of slots allocated for new users to request service. The net cycle length is given by

$$M_d = \sum_{i=1}^{N} n_i + \phi_d .$$

If the number of accesses n_i is equal to the number of slots r_i allocated in the contiguous allocations, the cycle length for distributed allocations is larger than that for contiguous allocations because of overhead. Therefore, the effective transmission rate for the i^{th} user, $R_i = n_i R/M_d$, is reduced when the transmission overhead is non-zero.

If the terminal accesses are uniformly distributed in the net cycle, the average time between transmissions for that terminal is equal to M_d/n_i. If M_d/n_i is an integer M_i, the i^{th} user transmits one data unit, waits $M_i - 1$ slots, transmits another data unit, waits $M_i - 1$ slots again, and so on (Fig. IV-4). This cycle behavior is identical to that of a net operating under fixed contiguous assignment with the terminal allocated one data unit and a cycle length of M_i slots (this cycle length will be referred to as the terminal cycle length to avoid confusion with the net cycle length). Therefore, the equations developed for the terminal buffer queue length (Eq. IV-3) and virtual delay (Eq. IV-8) are applicable for the i^{th} user, with $r_i = 1$ and $M = M_i$. For the case where M_d/n_i is not an integer, the terminal accesses are separated by either $\left[M_d/n_i\right]$* slots or $\left[M_d/n_i\right]_+$ slots, the relative occurrence being chosen so that the average separation between accesses is M_d/n_i. (The methodology used in developing the buffer length requires integer values for the allocation and the cycle length.) Unfortunately, the previous results are no longer applicable,

*[X] is the largest integer contained in X. In this case, where X is not an integer, $[X]_+ = [X] + 1$.

because the terminal cycle length is not constant but oscillates between two values, but the buffer contents and virtual delay at the i^{th} user can be bounded. The upper bound is obtained by assuming that the user is allocated one time slot in a cycle length of $\left[M_d/n_i\right]_+$ slots. The effective transmission rate in this net is smaller than the effective transmission rate of our net, and therefore the buffer contents and delay will be larger than in the distributed net. Similarly, the lower bound is obtained by assuming that the user is part of a net with a cycle length of $\left[M_d/n_i\right]$ slots and is still allocated one time slot. The upper and lower bounds on the average terminal buffer queue length and delay are given by Eqs. IV-3 and IV-8 evaluated at $r_i = 1$, and $M = \left[M_d/n_i\right]_+$ and $\left[M_d/n_i\right]$, respectively.

The results for the buffer contents can easily be extended to the case where the terminal is allowed to transmit r (>1) data units at each access as long as the allocation does not change from access to access of that terminal. In this distributed allocation case, the i^{th} terminal accesses the satellite n_i times in a net cycle and at each access can transmit r data units. Allowing the transmission of more than one data unit per access will reduce the net cycle overhead. The upper and lower bounds on the buffer length are obtained by evaluating Eq. IV-3 at $r_i = r$, and $M = \left[M'd/n_i\right]_+$ and $\left[M'd/n_i\right]$, respectively, where $M'd$ is the cycle length obtained using n_i accesses with r data units at each access. The bounds for the virtual delay are complicated by the fact that the virtual delay with r slots could not be found exactly except for r = 1. The upper bound for the virtual delay with n_i accesses and r slots at each access is given by the upper bound on the delay

in Eq. IV-7 with $M = \left[M' d/n_i\right]_+$ slots, and the lower bound is given by Eq. IV-8 with $M = \left[M' d/n_i\right]$.

Unfortunately, the upper and lower bounds for the buffer contents with distributed allocations included the buffer contents with contiguous allocations in many examples for overhead case II. Therefore, another model for distributed allocations is investigated because of these inconclusive results and the desire to study the effects of the access distribution. In the development, the i^{th} user is allowed to transmit one data unit and then must wait a random number of slots before transmitting again, and the process repeats. The time allocation to the i^{th} user is cyclic, but the cycle length is random. The terminal cycle length is a discrete random process that is independent and identically distributed from cycle to cycle with a distribution such that the average cycle length is $M d/n_i$. For example, with 3 accesses allocated to the i^{th} user and a net cycle length of 12 slots ($M d/n_i = 4$), the terminal cycle length is 4 slots (the random variable has only one value and that occurs with probability one) if the accesses are uniformly distributed in the cycle. On the other hand, the terminal cycle length could have two values (3 or 5 slots), each occurring with equal probability. Which of these choices is better?

The cycle length of the j^{th} (terminal) cycle for the i^{th} user is $M_j^{(i)}$, and the buffer contents in data units at the beginning of this cycle are $L_j^{(i)}$. The cyclic behavior of the buffer contents with the cycle initiated by the i^{th} transmission slot is described by

$$L_{j+1} = (L_j - 1)^+ + \sum_{k=1}^{M_j} X_k \,, \qquad \text{(IV-11)}$$

where the superscript i has been suppressed and where X_k is the number of data unit arrivals during the k^{th} slot of the j^{th} cycle (M_j is random). The expected value of the steady-state buffer contents can be derived by the generating function methodology (Appendix D) or by the easier technique used in deriving the Khintchine-Polloczek equations (Refs. 17, 40, 41). The stability condition, which is identical to that obtained for contiguous allocations, for the existence of a steady-state solution is $\bar{M}\mu < 1$, where $\bar{M} = {}^{M_d}/_{n_i}$ is the average cycle length, and $\mu = E\left\{X_k\right\}$ is the average number of data-unit arrivals during a time slot. The average value of the buffer contents in steady state is given by

$$E\left\{L^*\right\} = \frac{\bar{M}\,\sigma_x^2}{2(1-\bar{M}\,\mu)} + \frac{1}{2}\,\bar{M}\,\mu + \frac{\mu^2\,\sigma_M^2}{2(1-\bar{M}\,\mu)}, \quad \text{(IV-12)}$$

where σ_x^2 and σ_M^2 are the variances of the data-unit arrivals and the cycle length, respectively. The previously defined virtual delay is again used as the waiting-time measure for a message and is given by

$$D_m = \frac{\bar{M}+1}{2} + \bar{M}\left\{E\left\{L^*\right\} - \frac{\bar{M}+1}{2}\,\mu\right\} + (m-1)\,(\bar{M}-1), \quad \text{(IV-13)}$$

where M is the length of the virtual message. The derivation of Eq. IV-13 is included in Appendix D.

The particular distribution chosen for the accesses within the net cycle determines the value of σ_M^2, which is a measure of the dispersion of the cycle length values around the mean. The last term of Eq. IV-12 portrays the dependence of the buffer contents and also the virtual delay on the distribution strategy. This term is directly proportional to the cycle length variance. An increase in the variance (increasing the dispersion of the values) results in an increase in both

the buffer contents and the virtual delay. Therefore, the distribution strategy that has the smallest dispersion is optimum for both buffer contents and delay. For example, when $\bar{M} = {}^{M}d/n_i$ is an integer, the optimum distribution is uniform with only one value for the terminal cycle length, $M_j = \bar{M}$ ($\sigma_M^2 = 0$). When ${}^{M}d/n_i$ is a non-integer, the optimum distribution has the terminal cycle length assume two values $\left[{}^{M}d/n_i\right]$ and $\left[{}^{M}d/n_i\right]_+$ with appropriate probabilities to ensure that the average cycle length is ${}^{M}d/n_i$.

The behavior of the average buffer contents under the various allocation schemes (contiguous and distributed) is examined in the following. The first term of Eq. IV-12 has been rewritten with the substitution $\bar{M} = {}^{M}/r$ with $n_i = r$. For the allocations indicated below, the average buffer contents are given by the following equations:

a) One data slot per net cycle

$$E\left\{L\right\} = \frac{1}{2}\ \frac{M\ \sigma^2}{1-M\mu} + \frac{1}{2}\ M\mu \qquad \text{(IV-14)}$$

b) r contiguous data slots per cycle

$$E\left\{L\right\} = \frac{1}{2}\ \frac{M\ \sigma^2}{r\ -\ M\mu} + \frac{1}{2}\ M\mu\ +\ \frac{1}{2}\sum_{t=2}^{r}\ \frac{1\ +\ \theta_t}{1\ -\ \theta_t} \qquad \text{(IV-15)}$$

c) r randomly distributed data slots in the net cycle

$$E\left\{L\right\} = \frac{1}{2}\ \frac{M\ \sigma^2}{r\ -\ M\mu}\ +\ \frac{1}{2}\ \bar{M}\mu\ +\ \frac{1}{2}\ \frac{\mu^2\ \sigma_M^2}{1-\bar{M}\mu}\ . \qquad \text{(IV-16)}$$

The following queueing interpretation can be imparted. In all the models, service to a customer is initiated only at discrete points (i.e., the beginning of each cycle). Equation IV-14 is the average queue length for a single server queue, the service time for each customer lasting M slots. The model for

r contiguous allocations is a queue being served by r servers, and service is initiated at the discrete points, the service time for each customer again lasting M slots. The equivalent service time if there had been one server would be M/r slots. By distributing the r slots randomly in a net cycle, the queue model again becomes a single-server queue with a random service time whose average is equal to M/r slots. Service to a new customer is initiated by termination of service to the customer being served, as if a customer were being served when the queue is empty. The second term in each equation is seen to be proportional to the average number of arrivals during a service time, while the first term is intuitively proportional to the residual queue left behind the previous customer. [The equation for the queue behind the $n+1^{st}$ customer is given by the queue left behind the n^{th} customer, minus the $n+1^{st}$ customer, plus the arrivals that occur during service to the $n+1^{st}$ customer, i.e., $q_{n+1} = (q_n - 1)^+ + v_{n+1}$.] The third term, which includes σ_M^2 of Eq. IV-16, as was previously indicated, is a result of the chosen distribution of allocations. Likewise, the summation term of Eq. IV-15, is a result of allocating r contiguous slots every M slot.

If synchronization preambles and guard times are required every time a terminal transmits, distributing the allocations with one data slot at each access can significantly reduce the effective transmission rate, resulting in a poor utilization of the channel. This is demonstrated by the following example. Consider an eleven-member net operating under overhead case I, that is, 0.39 seconds of overhead is associated with each transmission. If contiguous allocations are used, the number of accesses or transmissions is eleven for any set of allocations. If it is desirable to have an allocation pool of 66 data slots, with contiguous allocations, 17 slots are required for transmission overheads and 12 slots for new entry

request time, resulting in a cycle length of 95 slots. With a bit transmission rate of 2.4 kb/s, the cycle overhead reduces the effective data transmission rate to about 1.7 kb/s, a reduction of about 30 percent.

With distributed allocations, an allocation pool of 66 data units results in 66 separate transmissions. Therefore, 102 slots are required for transmission overheads and an additional 24 slots are needed for new entry request time, resulting in a net cycle of 192 slots (more than twice the contiguous cycle length). The effective data transmission rate is reduced to 825 b/s, only 34 percent of the transmission rate (a reduction of 66 percent). The effective data rate for distributed allocations is half the data rate for contiguous allocations. The result is that it takes twice as long to transmit a data unit. It should be clear from this example that, with this transmission overhead, distributing the allocations is not effective. On the other hand, there may be other implementations of distributed allocations (e.g., increasing the transmission duration to more than one slot at each access) that could produce smaller buffer contents and shorter delays.

For overhead case II, comparisons of the distributed and contiguous fixed assignments with equal slot allocations per net cycle ($n_i = r_i$) are presented in Fig. IV-14 and Table IV-10. Both demonstrate that distributed allocations reduce (unfortunately, a small reduction) the net buffer contents with contiguous allocations. Figure IV-14 also demonstrates that the virtual delay in distributed allocations is smaller than the delay upper bound but larger than the lower bound in contiguous allocations. Both of these results are consistent with analytical comparisons of the respective equations. With higher arrival rates, more dramatic improvements may become evident. Further, the comparison should have been performed

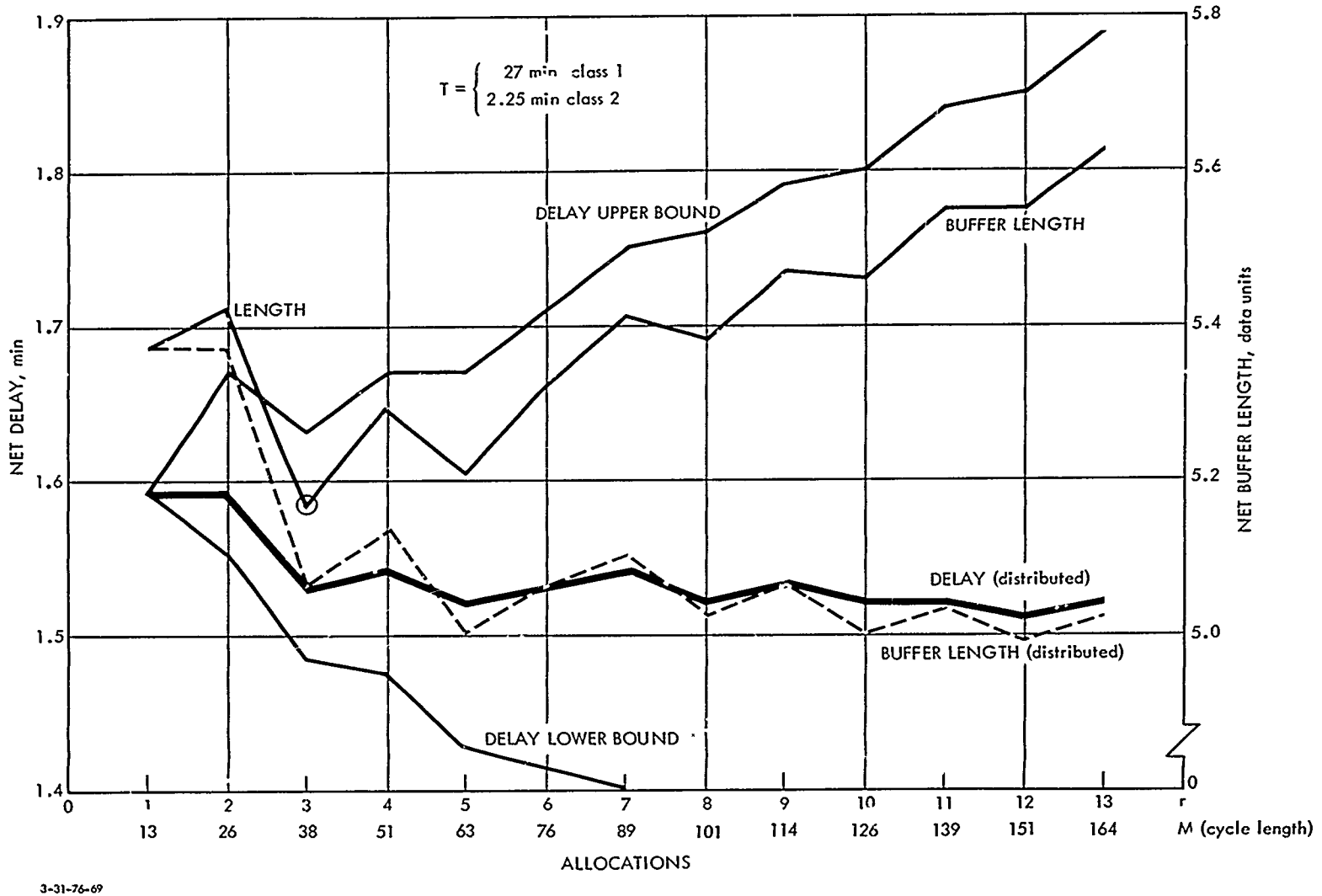


FIGURE IV-14. Equal Allocations for Overhead Case II

with the respective optimum allocations and not equal slot allocations for both techniques.

TABLE IV-10. A COMPARISON OF DISTRIBUTED AND CONTIGUOUS ALLOCATIONS

			$T = \begin{cases} 27 \text{ min--Class 1} \\ 2.25 \text{ min--Class 2} \end{cases}$				$T = \begin{cases} 10 \text{ min--Class 1} \\ 0.833 \text{ min--Class 2} \end{cases}$			
Allocations			Buffer, data units		Delay, min		Buffer, data units		Delay min	
r_1	r_2	M	D[a]	C[b,c]	D[a]	C[b]	D[a]	C[b,c]	D[a]	C[b,c]
2	13	38	2.37	2.48	1.94	2.06	7.67	(7.98)	2.29	2.41
3	14	51	2.26	(2.43)	1.73	1.86	7.56	8.02	2.01	2.15
5	16	76	--	--	--	--	8.62	9.35	1.80	(1.99)

[a]D: distributed allocations.

[b]C: contiguous allocations.

[c]Circled values are minimum for contiguous allocations.

In fixed assignments, the allocations are based on the average arrival rate, and therefore the terminal's use of its allocation may be inefficient. Messages arrive randomly and have random sizes, and therefore sometimes the allocation may be more than is required and sometimes it may be substantially less, hence causing the buffer contents to increase very rapidly to the steady-state value. Distributing the allocations tends to lessen the inefficient use of the allocated slots and therefore yields smaller buffer contents and possibly delays. On the other hand, in fixed assignments, the user always knows exactly when his next allocation occurs.

C. ACCESS WITH POLLING

In the previous section, the fixed-assignment technique of allocating transmission time to the users in the net was investigated. In that technique, a terminal's allocated transmission time or capacity is constant from cycle to cycle. Therefore, the amount allocated must be based on long-term statistics and not on instantaneous requirements. Because messages arrive randomly and have random sizes, the user's utilization of his allocated time may be inefficient. Access with polling is one technique that allocates transmission time on a requirements basis but still maintains the cyclic structure of Fig. IV-1. Figure IV-3 depicts one cycle of the net operation.

Polling can be accomplished with a central controller or a distributed controller. The controller sequentially polls the users according to their polling position. When a user is polled, he is allowed to transmit until his buffer is empty. Messages that arrive while the user is transmitting are transmitted during this period. The duration of the transmission and hence the cycle length are therefore random. When the buffer is emptied, the user transmits an "end of transmission" message. When this message is received, the process continues with a poll of the next user. With central control, polling is initiated by one member of the net. The controller transmits the polling message to the appropriate user and receives the "end of transmission" message. With distributed control, the polling sequence is established, the first user transmits until the buffer is empty, sends an "end of transmission" message to the second user, who beings transmitting when the end statement is received, and so on. The time between the end of a transmission to the time the next user begins transmitting data is the idle time. The idle time consists of the "walking" time, which is the time required to transfer control

from one user to the next and the time to transmit synchronization preambles. The walking time is initiated at the beginning of the "end of transmission" message to the time the next user is polled (informed that he can transmit). The walking time is dependent upon whether the control is distributed or centralized. The walking time for distributed control consists of the duration of the end message plus the propagation delay from the terminating user to the satellite and back to the next user. With centralized control, the walking time is composed of the durations of the "end" message and the "poll" message plus the single-hop delays from the terminating user to the controller and from the controller to the polled user. The idle time is assumed to last r slots.

The buffer contents, the virtual delay, and the average cycle length are extracted from Ref. 27. The results are only valid for identically distributed data-unit arrival processes at the users. The required results are presented with no derivations but some heuristic justification. The stationary expected length of the polling cycle is

$$E\{M\} = \frac{Nr}{1-N\mu}, \qquad \text{(IV-17)}$$

where r is the idle period in data slots, and N is the number of terminals in the net. The mean value of the terminal buffer queue length in steady state is identical for all terminals and is given by

$$E\{L^{(i)}\} = \frac{1}{2}\,\frac{\sigma^2}{1-N\mu} + \frac{1}{2}\,\frac{Nr\,\mu(1-\mu)}{1-N\mu}, \qquad \text{(IV-18)}$$

while the delay for a virtual message of m units is

$$D_m^{(i)} = \frac{1}{2}\,\frac{N\sigma^2}{1-N\mu} + \frac{1}{2}\,\frac{Nr\,(1-\mu)}{1-N\mu} + \frac{1}{2}\,(1-\mu). \qquad \text{(IV-19)}$$

The virtual delay is independent of virtual message size m. A terminal's buffer contents are emptied when it is polled, and therefore the delay consists of the time waiting for the channel to be made available and the transmission time for the already queued messages at the virtual message arrival. The message is completely transmitted at this time.

To explain the absence of N in the numerator of the first term of the queue length, since it was present with fixed assignments, consider a star communication system where the terminals transmit their messages on dedicated channels to a common buffer from which the data units are removed. The mean value of the stationary common queue length with identically distributed input processes is

$$E\left\{L_c\right\} = \frac{1}{2}\,\frac{N\sigma^2}{1-N\mu} + \frac{1}{2}N\mu .$$

If the common buffer contents are now uniformly distributed into the user buffers, the i^{th} buffer contents would be

$$E\left\{L^{(i)}\right\} = \frac{1}{2}\,\frac{\sigma^2}{1-N\mu} + \frac{1}{2}\mu .$$

The buffer contents are identical to average queue lengths in customers (messages) obtained by the Khintchine-Polloczek (K-P) formula for a single server queue with Poisson arrivals a. geometric service times. The K-P formulas are applicable to any queueing or dispatching discipline provided that the selection of the next item to be served does not depend on the service time. Therefore, the star system could have used a polling service discipline, but the walking time must be zero. The server can sequentially empty each queue in the star system with a zero walking time between queues, and the average queue length is still given by the above formula.

Note the behavior of the first term of the i^{th} buffer contents--it depends only on σ^2 and not on $N\sigma^2$. In the derivation of the stationary expected queue length (Eq. IV-18) for polling, the idle time r was assumed to be positive integer valued (nonzero). Otherwise, we would have had a contradiction because the buffer queue length with polling and $r = 0$ is smaller than the length obtained with the star system using the polling discipline. The first term of Eq. IV-18 accounts for the arrivals that occur while the user is transmitting. The average time the channel is being used by the other users is equal to $Nr(1-\mu)/(1-N\mu)$. Therefore, the second term accounts for the arrivals during this inactive period. The average number of arrivals is μ times the average inactive period.

The stability condition for a steady-state solution to exist is given by $N\mu < 1$; that is, the average number of arrivals in the net during a slot are less than the number of data units that can be transmitted during a slot. For $N\mu = 1$, the average cycle length, buffer contents, and delay become infinite. It is surprising that the idle time (net overhead) does not enter into the stability condition. The size of the idle time affects the average cycle length, the average contents, and delay in a linear manner.

Access with polling is interesting in that the polling order is the only quantity that has to be determined a priori. When the user is polled, he transmits until he has nothing more to say. This is in contrast to fixed assignments, where the relative positions and the number of slots allocated to each user have to be determined.

The virtual delays and average buffer contents are identical for all users because the arrival processes were the same. Therefore, the system performance measures in the net are simply equal to the terminal buffer contents and the terminal virtual delay, respectively.

The parameter values assumed for the evaluation of the performance of a net operating under access with polling are presented in Table IV-11. The channel and traffic parameters are identical to those assumed for the fixed-assignment net, with the major exception of having one class of users. The idle time per transfer is equal to the preamble transmission time, plus the single-hop propagation delay with distributed control, and also plus another propagation delay with a central controller. The total idle time in the net is equal to the number of users in the net N times the single-transfer idle time, which then is converted to data time slots and rounded up to the nearest integer. The channel utilizations ($\rho = N\mu$) are presented in Table IV-12 for the various interarrival times.

TABLE IV-11. PARAMETER VALUES FOR POLLING

Channel Parameter:
Transmission Rate = 2.4 kb/s
Traffic Parameter:
Number of Users (N) = 11 (one class)
Average Interarrival Times (T) = 27, 10, 2, 1.34 minutes
Average Message Length = 26 data units
Virtual Message Length (m) = 26 data units
Data Unit Length = 608 bits
Idle-Time (Overhead) Parameters:
Propagation Delay = 0.27 seconds
Synchronization Preamble = {936 bits (Normal) / 96 bits (Reduced)}

TABLE IV-12. CHANNEL UTILIZATIONS

Interarrival Time, min	Channel Utilization
27	0.045
10	0.121
2	0.604
1.34	0.9

The values for the average cycle length, buffer contents, and virtual delay are presented in Table IV-13 for normal overhead and both distributed and centralized control and in Table IV-14 for reduced overhead and distributed control. If centralized control were used in the reduced overhead case, the total idle time would be 26 slots, and the resultant performance measures would be slightly smaller than those tabulated in Table IV-13 with distributed control. Substantial reductions in total idle time (overhead) are achieved by changing the control from centralized to distributed or by reducing the preamble, but the resultant improvements in the system performance measures are smaller and decrease with increasing utilization. The average cycle length, buffer contents, and virtual delay all increase, as is expected, with increasing channel utilization. The average message throughput in a cycle is equal to the average message input for a stable (nonsaturated) system, and hence the cycle length increases and the cycle overhead percentage decreases with increasing channel utilization.

The results are surprisingly low, considering the length of the propagation delay. The buffer is emptied at each poll, and therefore at the next poll, a cycle length later, the buffer contains the arrival that occurred in the previous cycle but no residual from other cycles.

TABLE IV-13. ACCESS WITH POLLING FOR NORMAL OVERHEAD

Interarrival Time, min	Average Cycle Length, slots		Buffer Length, data units		Delay, mins	
	Centralized[a]	Distributed[b]	Centralized	Distributed	Centralized	Distributed
27	42.9	30.4	0.2	0.2	0.097	0.07
10	46.6	33	0.57	0.5	0.11	0.086
2	103.5	73.2	6.2	5.4	0.37	0.31
1.34	410.0	290.0	36.3	31.8	1.77	1.53

[a]Nr = 41 slots.

[b]Nr = 29 slots.

TABLE IV-14. ACCESS WITH POLLING FOR REDUCED OVERHEAD AND DISTRIBUTED CONTROL

Interarrival Time, min	Average Cycle Length,[a] slots	Buffer Contents, data units	Virtual Delay, min
27	14.7	0.14	0.04
10	15.9	0.40	0.05
2	35.4	4.45	0.24
1.34	140.0	26.12	1.23

[a]Nr = [illegible] slots.

D. ACCESS WITH RESERVATIONS

The allocation techniques investigated in the previous sections treat the users as entities to which capacity is allocated. The users are arranged in an orderly fashion, and, during the user's turn, the user is allocated a constant transmission time in fixed assignments or a random time sufficient to empty the buffer in polling. Under access with reservations, capacity (transmission time) is allocated to the users on a demand basis by treating messages as entities and allocating capacity to individual messages by the use of an orderwire and a central or distributed controller. With a central controller, the users request transmission time via the orderwire at the arrival of a message. The controller receives these requests and allocates time according to a queueing discipline. The user then transmits the message completely during the allocated time. With distributed control, users reserve "unclaimed" slots by transmitting a reservation message over the orderwire, and each user monitors the orderwire for the status of future slots.

The "request" or "reservation" message contains information describing the message length or transmission time, while the time these "messages" are received can be viewed as the origination time of the message. Therefore, the following interpretation is made of net behavior under access with reservations as an aid in understanding the technique and also developing system performance measures. By means of the [illegible] and the "request" messages, the controller forms a (local) common queue of messages requiring service and provides service (transmission time) to the queued messages using the queueing discipline, which is assumed to be first-in, first-out for consistency with the other allocation techniques. In the distributed control situation, each user forms this common queue, from which the user determines the

appropriate transmission time for his messages. Because messages are generated at the users, the time between the generation and the joining of the common queue must be accounted for. This time is the turnaround time, which can be interpreted as the walking time of a customer (message) from the user terminal to the common queue. The queueing model for the complete process can be thought of as consisting of a waiting room and the common queue. Customers arrive at the waiting room, are ordered, and wait a portion of time (the turnaround time) before joining the common queue. Service is provided only to those messages in the common queue. The turnaround time is not equivalent to another service time, because each message waits a time amount measured from its generation and not from a previous message departure to the common queue.

In the development, the orderwire is assumed to be a separate channel utilized on a contention (random-access) basis. Further, it is assumed that "messages" transmitted on the orderwire do not destructively interfere (i.e., there is a zero probability of blocking).* The characteristics of the orderwire and the control messages must be properly designed to ensure that the probability of blocking is small. The turnaround time for messages to join the common queue is therefore assumed to be constant and is determined by the duration of the "reservation" message and the single-hop propagation delay with distributed control, and by the durations of the "request" and "allocation" messages and the double-hop propagation delay with central control. The message arrival process at the common queue is assumed to be Poisson with an arrival rate equal to the sum of the terminal rates, and the message

*A more exact derivation is presented in Appendix E.

distribution is assumed to be identical for the users. These assumptions and the decomposition of the queueing model into separate parts allows one to determine the queueing parameters (queue length and queueing time) of the common queue and to separately determine the number of messages in the waiting room.

The contents of the common queue can be determined by using the Khintchine-Polloczek (K-P) equations for a single-server queue with geometrically distributed service time or by using the results presented for the star communication system (Section IV-C or Ref. 22) with compound Poisson arrivals and constant data-unit service time. The K-P equation yields the average number of messages in this queue, from which the average message waiting time is determined by Little's theorem, while the star communication result is in data units. The expected value of the stationary queue length, in data units, for the common queue is given by

$$E\left\{L_C\right\} = \frac{1}{2}\,\frac{\sigma_T^2}{1-\mu_T} + \frac{1}{2}\mu_T ,$$

where $\sigma_T^2 = \sum_{i=1}^{N} \sigma_i^2$ and $\mu_T = \sum_{i=1}^{N} \mu_i$, with μ_i and σ_i^2 the mean and variance of the compound Poisson distribution of the i^{th} terminal. The data units in the waiting room are the messages that have not been positioned in the common queue. These arrivals are those that occur during a walking period. The average arrivals during this period are given by the average data-unit arrivals during a slot, multiplied by the walking time in slots, which is assumed to be "a" slots. Therefore, the joint terminal buffer contents, equal to the common queue contents plus the waiting room contents, are given by

$$E\{L_T\} = \frac{1}{2}\,\frac{\sigma_T^2}{1-\mu_T} + \frac{1}{2}\,\mu_T + a\mu_T \qquad \text{(IV-20)}$$

From which the buffer contents of the i^{th} user is identified as

$$E\{L^{(i)}\} = \frac{1}{2}\,\frac{\sigma_i^2}{1-\mu_T} + \frac{1}{2}\,\mu_i + a\mu_i . \qquad \text{(IV-21)}$$

The average message waiting time at the i^{th} user, which is obtained from the joint message buffer queue length, is identical for all users and is given in data slots by

$$W_q^{(i)} = a + \frac{\sigma_T^2}{2(1-\mu_T)} . \qquad \text{(IV-22)}$$

The channel utilization ρ is equal to μ_T, and for stability $\mu_T < 1$, with a terminal utilization given by μ_i. The derivation of the above results assumes that the terminals transmit synchronously.

If synchronization preambles must be transmitted with each message, the common queue contents are derived from the K-P equation using a service distribution equal to a constant (preamble time) plus a geometrically distributed term. Synchronization preambles affect only the contents of the common queue. The resultant equations are presented in Table IV-15.

A common queue of messages is essentially formed in access with reservations. Because of this, (1) the queueing discipline can be applied throughout the net and not just at the terminals, as with fixed assignments or polling, and (2) the controller (centralized control appears to be required) can allocate capacity (time) to messages on more than one channel if they exist and if the user communication equipment

permits. Therefore, access with reservations can pool the available communication channels and thereby add more flexibility to the resource sharing, while potentially improving the grade of service to the users. Instead of having a separate queue for each available channel (in queueing theory, the C "1 server--1 queue" situation), pooling the available channels combines the separate queues into one queue, from which messages are transmitted on the next free channel (the "C servers--1 queue" situation). The waiting times are tabulated in Table IV-15 for a nonpreemptive priority queue discipline (one channel only) and for the pooling of 2 or 3 channels with a first-in, first-out queue discipline, with the assumption that the message lengths are exponentially distributed because these results do not exist for geometrically distributed message lengths. The average terminal buffer contents are given by $E\left\{L^{(i)}\right\} = \lambda_i \bar{m} (W_q + \bar{S})$, where $\bar{S} = \bar{m}$ time slots, the average message transmission time.

In the following, the improvements obtainable by pooling channels are examined. Queueing theory notation is utilized to simplify the comparison (e.g., M/M/C is used to refer to the queueing system, C servers--1 queue). It is assumed that the total message arrival rate for each queue (joint terminal buffer queue) is λ_T in the M/M/1 systems (1 server--1 queue). The total arrival rate in the M/M/C system is therefore $C\lambda_T$. The comparisons are made on the buffer contents and waiting times in the controller common queue (that is, neglecting the effect of the turnaround time, which is identical for the cases considered). For the queue length comparison, the controller common queue of the M/M/C system is compared to the sum of the common queues in the M/M/1 systems. On the other hand, the waiting time in the M/M/C system is compared to the waiting time in any M/M/1 system. If the arrival rates were different

TABLE IV-15. WAITING TIME EQUATIONS[a]

	Average Message Waiting Time W_q	Symbol Definitions
Synchronization Preambles	$X + a + \frac{\rho^2(1+C_b^2)}{2\lambda_T(1-\rho)}$	λ_T = total arrival rate $\rho = \lambda_T(X + \overline{S})$ X = preamble transmission time $\overline{S} = \overline{m}$ time slots $C_b^2 = q\,\overline{S}^2/(X + \overline{S})^2$
One Channel[b]	$a + \frac{\rho\,\overline{S}}{1 - \rho}$	$\rho = \lambda_T\,\overline{S}$
Pooling of Two Channels[a]	$a + \frac{\rho^2\,\overline{S}}{(1 - \rho^2)}$	$\rho = \lambda_{T_2}\,\overline{S}/2$
Pooling of Three Channels[b]	$a + \frac{3\,\rho^3\,\overline{S}}{(1-\rho)\,(2 + 4\rho + 3\rho^2)}$	$\rho = \lambda_{T_3}\,\overline{S}/3$
Nonpreemptive Priority Queue Discipline[b]	$W_q^{(i)} = a + \frac{\sum_{k=1}^{r} \rho_k\,\overline{S}_k}{(1-\sigma_{i-1})\,(1-\sigma_i)}$	$W_q^{(i)}$ = waiting time for message with priority i (i = 1 is highest priority) r priority classes $\rho_k = \lambda_T^{(k)}\,\overline{S}_k$ $\overline{S}_k$ = average message transmission time of priority k message $\sigma_k = \sum_{i=1}^{k} \rho_i$ $(\sigma_0 = 0)$

[a]Buffer contents $E\left\{L^{(i)}\right\} = \lambda_i\,\overline{m}\,(W_q + \overline{S})$.

[b]Reference 41 (exponentially distributed message length).

in the M/M/1 systems, the buffer contents comparison would still be the same, but the M/M/C waiting time would be compared to the average of the M/M/1 waiting times. The average contents and waiting time in the controller (hypothetical) queue in an M/M/C system are $L_{M/M/C}$ and $W_{q_{M/M/C}}$ and are related to the measures of the M/M/1 systems by

$$L_{M/N/C} = C\, L_{M/M/1} \cdot A_{c1}$$

$$W_{q_{M/M/C}} = W_{q_{M/M/1}} \cdot B_{c1} \cdot$$

Therefore, the reductions in the number of messages and waiting times at the common queue are simply A_{c1} and B_{c1}, respectively. These reductions are presented for several channel utilizations in Table IV-16. The results indicate that, with a per-channel utilization of 0.5, the waiting time in the common queue by pooling 3 channels is about 0.16 of the waiting time in the unpooled system. The total average message waiting time is the turnaround time plus the waiting time in the common queue.

TABLE IV-16. REDUCTIONS IN THE NUMBER OF MESSAGES AND WAITING TIMES BY POOLING CHANNELS

ρ	A_{21}[a]	B_{21}[a]	A_{31}[a]	B_{31}[a]
0.1	0.909	0.091	0.901	0.012
0.5	0.667	0.333	0.579	0.158
0.9	0.526	0.474	0.372	0.303

[a] A_{k1} and B_{k1} are the reductions in the common queue buffer contents and waiting times, respectively, by pooling K channels.

After having combined the queues, further improvements are possible if the available channels can be merged into one channel (e.g., C channels each with transmission rate R are merged to provide one channel with transmission rate CR). The transmission times of messages for these two cases are different, the user being able to transmit a message C times faster on the merged channel. Therefore, instead of waiting times, the queueing times (the total time a message spends in the system) are compared for the case of 3 channels. Using the equations developed for the M/M/1 queue with $\mu = 3\mu_1$ (the queue with the merged channel), and for the M/M/3 queue with $\mu = \mu_1$, smaller queues and queueing times of the common queue are obtained by merging the channels. The queue length for the M/M/1 queue is

$$L_{M/M/1} = L_{M/M/3}\ A_{13},$$

where $A_{13} = 1/(3\ A_{31})$ is the queue length reduction by merging the channels, and A_{31} is the reduction in combining the queues. The corresponding relationship for queueing times is identical to the queue length relationship, and therefore A_{13} is also the queueing time reduction.

TABLE IV-17. IMPROVEMENT IN THE COMMON QUEUE MEASURES BY MERGING THREE CHANNELS

ρ	A_{13}
0.1	0.37
0.5	0.58
0.9	0.9

Numerical Results

The values of the system performance measures for the net operating under reservations are presented in Table IV-18, using the parametric values of Table IV-1 (two user classes), and in Table IV-19, using the channel and traffic parameters of Table IV-11 (one user class). The results of an example of pooling channels are presented in Table IV-21. It is further assumed that the orderwire is a separate channel, the control messages required for reservations do not destructively interfere, the turnaround time is 3 data units (centralized control), and guard times are not required (the effect is discussed). For the case of two user classes and under synchronous operation (Table IV-19), the buffer contents in the class II terminal are twelve times larger than the class I contents, but the waiting times are identical.

TABLE IV-18. BUFFER CONTENTS AND WAITING TIMES FOR ACCESS WITH RESERVATIONS (Zero Synchronization Preambles and a = 3 Slots)

Class	Interarrival Time, min	Channel Utilization	Buffer Contents, data units	Waiting Time, min
1	27	0.089	0.13	0.023
2	2.25		1.54	0.023
Average			0.26	0.023
1	10	0.242	0.41	0.047
2	0.833		4.89	0.047
Average			0.81	0.047

TABLE IV-19. BUFFER CONTENTS AND WAITING TIMES FOR RESERVATIONS WITH IDENTICALLY DISTRIBUTED MESSAGE PROCESSES (a = 3 Slots)

Interarrival Time, min	Buffer Contents, data units		Waiting Time, min	
	A[a]	B[b]	A[a]	B[b]
27	0.1	0.1	0.02	0.02
10	0.4	0.4	0.04	0.03
2	4.2	3.7	0.21	0.18
1.34	45.2	21.2	2.22	0.98

[a]A: synchronization preamble = 936 bits.

[b]B: zero preamble.

For one user class, Table IV-19 contains the results for various average interarrival times and two values of synchronization preambles. As is expected, for each preamble value the buffer contents and waiting times increase with message arrival rates. The effect of the preamble on the system performance measures is small except for T = 1.34 minutes, where the performance-measure values for the 936-bit preamble have doubled over those for zero preamble. For an expanded definition of channel utilization as the average arrival rate times the average message transmission time including the preamble transmission time, Table IV-20 provides an explanation of the preamble effect demonstrated in Table IV-19. The preamble or any transmission overhead (e.g., guard times) increases the channel "utilization" value and thereby increases the buffer contents and waiting times and decreases the traffic load that results in saturation.

TABLE IV-20. EFFECTS OF PREAMBLE ON CHANNEL "UTILIZATION"

Interarrival Time, min	"Utilization"	
	A[a]	B[b]
27	0.047	0.045
10	0.128	0.121
2	0.64	0.604
1.34	0.95	0.9

[a]A: synchronization preamble = 936 bits.

[b]B: zero preamble.

The example for which the results are presented in Table IV-21 is chosen primarily to demonstrate the effect of the orderwire channel and secondarily to demonstrate the improvement of pooling channels (combining queues). The parameter values of Table IV-11 are used for each of the four nets. One of the communication channels is the orderwire, while the remaining three channels are shared by the combination of the four nets for data transmission. For the assumed average message length (26 data units), the capacity allocated to the orderwire is excessive because the average numbers of request messages per data slot are 0.007, 0.018 and 0.093 for the message interarrival times of Table IV-21. For this example, the assumption of no blocking on the orderwire is reasonable. On the other hand, for substantially smaller average message lengths, the orderwire capacity may be insufficient, resulting in saturation of the orderwire and eventually zero throughput of control messages. The removal of one channel for the orderwire results in an increase in the channel utilization for the pooled system as compared to the individual systems and a decrease in the message load that results in saturation (0.75 of the individual saturation load).

The results in Table IV-21 can be compared to those of Table IV-19 for identical interarrival times, although the actual improvements would be larger if the same number of data channels were provided in both cases (the pooled system utilizes 3 data channels, while the unpooled system has 4 data channels). Even with the removal of one channel, the terminal buffer contents and waiting times in the pooled system are smaller than those for the individual systems in the pooled stable region.

TABLE IV-21. BUFFER CONTENTS AND DELAYS WITH 44 USERS AND THREE COMMUNICATION CHANNELS (Zero Preambles and a = 3 Slots)

Interarrival Time, min	Utilization[a]	Buffer Contents, data units	Waiting Time, min
27	0.06	0.1	0.0127
10	0.16	0.3	0.0133
2	0.805	3.2	0.136

[a]For pooled system.

Under reservations, the orderwire reduces the available data capacity, and the saturation message load level, but reservations adds more flexibility by permitting the pooling of resources and by potentially improving the grade of service to the users.

E. COMPARATIVE EVALUATIONS

In the preceding sections, three allocation schemes were introduced, and the resultant terminal buffer queue lengths and virtual delays or waiting times were determined and evaluated for specific examples. In this section, comparisons of the techniques are presented.

The equations for the terminal buffer contents and delays are reproduced in Tables IV-22 and IV-23, respectively. From these tables, it is evident that for finite buffer contents and delay to exist the following stability conditions must be satisfied:

- For fixed assignments,
  $M\mu_i < r_i$ for all i
- For polling,
  $N\mu < 1$
- For reservations,
  $\lambda_T X + \mu_T < 1$.

Initially, it appears that, except in polling, transmission overhead reduces the saturation message load. What is surprising is that saturation in polling is unaffected by transmission overhead. For the case of identically distributed terminal message processes with equal allocations to the users and a net overhead that is independent of allocation size (this excludes the new entry request times), the stability condition for fixed contiguous assignments can be rewritten as $N\mu + \frac{\phi}{r} < 1$, which indicates that the saturation level can almost be made independent of the overhead for sufficiently large allocations. This also possibly explains why overhead values do not affect the saturation level in polling. The transmission overhead in all cases increases the buffer contents and the delays, but only in reservations will it reduce the saturation message load. It is evident from Table IV-23 that the virtual delay or waiting time is independent of the virtual or actual message lengths except in fixed assignments, where the delay is proportional to the virtual message length. In fixed assignments, the delay for small messages is less than the delay for large messages.

TABLE IV-22. TERMINAL AVERAGE BUFFER CONTENTS IN STEADY STATE FOR FIXED ASSIGNMENTS, POLLING, AND RESERVATIONS

	Buffer Contents $E\{L^{(i)}\}$	Definitions/Conditions
Fixed Assignments	$\frac{1}{2}\frac{M\sigma_i^2}{r_i - M\mu_i} + \frac{M\mu_i}{2} + \frac{1}{2}\sum_{k=2}^{r_i}\frac{1+\theta_k^{(i)}}{1-\theta_k^{(i)}}$	Allocation = r_i contiguous slots $M = \sum_{i=1}^{N} r_i + \phi$ (cycle length) ϕ = net overhead $\theta_k^{(i)}$: Solutions of $Z^{r_i} - \left(P^{(i)}(Z)\right)^M = 0$ $\mu_i = \lambda_i \bar{m}_i$ $\sigma_i^2 = \mu_i(2\bar{m}_i - 1)$
Polling	$\frac{1}{2}\frac{\sigma^2}{1-N\mu} + \frac{1}{2}\frac{Nr\mu(1-\mu)}{1-N\mu}$	Identical message processes r = single transfer idle time $E\{M\} = Nr/(1-N\mu)$
Reservations	$\mu_i\left\{X + \bar{S} + a + \frac{\rho^2(1+C_b^2)}{2\lambda_T(1-\rho)}\right\}$ for X = 0 $\frac{1}{2}\frac{\sigma_i^2}{1-\mu_T} + \left(a + \frac{1}{2}\right)\mu_i$	Zero blocking probability a = turnaround time X = preamble transmission time $\bar{S} = \bar{m}$ time slots $\rho = \lambda_T X + \mu_T$ $\lambda_T = \sum_{i=1}^{N}\lambda_i$, $\mu_T = \sum_{i=1}^{N}\mu_i$ $C_b^2 = q\bar{S}^2/(X + \bar{S})^2$

TABLE IV-23. VIRTUAL DELAY OR AVERAGE MESSAGE WAITING TIME

	Virtual Delay[a]
Fixed Assignments	$D_\ell^{(i)} \le D_m^{(i)} \le D_u^{(i)}$ $D_u^{(i)} = B^{(i)} + (\frac{M}{r_i} - 1)\, m$ $D_\ell^{(i)} = B^{(i)} + (\frac{M}{r_i} - 1)\, (m - r_i)^+$ with $B^{(i)} = \frac{M+1}{2} + \frac{M}{r_i}\left(E\left\{L^{(i)}\right\} - \frac{M+1}{2}\,\mu_i\right)$
Polling	$\frac{1}{2}\,\frac{N\sigma^2}{1-N\mu} + \frac{1}{2}\,\frac{Nr(1-\mu)}{1-N\mu} + \frac{1}{2}\,(1-\mu)$
Reservations[b]	$X + a + \frac{\rho^2(1+C_b^2)}{2\lambda_T(1-\rho)}$ for $X = 0$ $a + \frac{\sigma_T^2}{2(1-\mu_T)}$

[a]Virtual message length of m data units.
[b]Average message waiting time.

Any comparison with polling must be limited to the case of identically distributed terminal message processes. It is assumed that the net using fixed assignments requires no overhead (zero transmission overhead and no new entry request time) because the best performance (the smallest values of system performance measures) is obtained under this condition. With these assumptions and equal allocations to all users in the net under fixed contiguous assignments, the average buffer contents and the virtual delay upper bound both monotonically increase with allocation size r, and therefore the optimum allocation to each user is one slot ($r = 1$ and $M = N$). On the other hand, the delay lower bound monotonically increases for allocations larger than the virtual message length ($r > m$), but for smaller allocation sizes it monotonically decreases if the summation term is less than half the allocation size* and therefore the optimum allocation is equal to the virtual message length ($r = m$). With the above assumptions, fixed distributed assignments also allocate one slot to each user per cycle and are therefore identical to contiguous allocations with the optimum allocation for buffer contents and delay upper bound. In the comparisons, an allocation of one data slot to each user is assumed for fixed assignments. The system performance measures are simply equal to the average terminal buffer contents and delay, respectively.

The respective equations for the buffer contents and delay can now be compared in order to determine the regions in the parameter space where the individual techniques are preferred. The regions are determined only with respect to the buffer contents. The comparisons for reservations neglect the

*Always true in the cases considered (average message length of 26 data units).

orderwire capacity and blocking of the control messages because these were neglected in deriving the equations. The following inequalities are obtained, with the virtual length m set equal to $\bar{m}$, for comparisons between

- Polling and Fixed Assignments. The polling technique is preferred if

$$r < r_c = \frac{(1 - \frac{1}{N})(2\bar{m} - 1) + 1 - N\mu}{1 - \mu} \approx 2\bar{m} .$$

Surprisingly, the region is determined primarily by the average message length and the transmission rate which affects the idle time. Table IV-24 presents values for the crossover idle time and the corresponding transmission rate.

TABLE IV-24. CROSSOVER IDLE TIME AND CORRESPONDING TRANSMISSION RATE FOR DISTRIBUTED CONTROL

Average Message Length, data units	N	Utilization	r_c, slots	Transmission Rate,[a] kb/s
5	11	0.1	9.2	20
		0.9	9.0	20
26	11	0.1	47.7	107
		0.9	50.6	114
26	100	0.1	51.4	115
		0.9	51.0	115

[a]Data unit = 608 bits; preamble length = 96 bits for polling.

- Reservations and Fixed Assignments. The reservations technique is preferred if

$$a < \frac{N-1}{2} \quad \frac{2\bar{m} - N\mu}{1 - N\mu} \approx \frac{N\bar{m}}{1-N\mu} .$$

The region boundaries are strongly dependent upon all the system and traffic parameters (Table IV-25).

TABLE IV-25. CRITICAL TURNAROUND TIMES AND TRANSMISSION RATES FOR CENTRALIZED CONTROL

Average Message Length, data units	N	Utilization	a, slots	Transmission Rate, b/s
5	11	0.1	55	61×10^3
		0.9	455	511×10^3
26	11	0.1	288	323×10^3
		0.9	2555	2.9×10^6
26	100	0.1	2855	3.2×10^6
		0.9	25,295	28.5×10^6

- Reservations and Polling. The reservations technique is preferred if

$$a < \frac{Nr}{2} \quad \frac{1-\mu}{1-N\mu} - \frac{1}{2} .$$

The inequality relates the reservation turnaround time to the average nontransmission duration for a user in polling, and hence, for most cases, the reservations technique is preferred.

With the nets operating under reduced or zero transmission overhead, either reservations or polling are preferred over fixed assignments for a wide range of parameter values (R, $\bar{m}$, and N), fixed assignments being the preferred technique only for small average message lengths and/or high transmission rates where the propagation delay in slots becomes excessive. For most cases, the buffer contents in reservations is smaller than that in polling when the effects of the orderwire capacity and blocking are neglected. The orderwire capacity reduces the available data transmission capacity, and blocking of the control messages increases the turnaround time, and both thereby increase the buffer contents and waiting times. With the same average message traffic, blocking is increased by reduction in the average message length.

Table IV-26 presents a comparison of techniques for the example used throughout the chapter. The results support the conclusions drawn from the analytical comparison, but the vast improvement (an order of magnitude) obtained by using a technique other than fixed assignments is surprising. The reservations technique is slightly better than polling, and the improvement may be expected to increase with the number of users in the net. In the net using polling, 14 slots in every cycle are used for overhead, while in the net using fixed assignments, all the slots are possible data transmission slots. A possible explanation of why polling is more efficient is that the terminal buffer is emptied at each poll, and the average number of arrivals during the overhead slots is small. In fixed assignments, because the buffer is not emptied at each service, the buffer contents accumulate over many cycles, only one data unit being transmitted per cycle. Although the average number of data-unit arrivals is small during a cycle, when a message is generated its average length is 26 data units, which takes an average of 282 slots to completely transmit.

TABLE IV-26. COMPARISON OF TECHNIQUES FOR THE CASE OF IDENTICAL TERMINAL MESSAGE PROCESSES WITH N = 11, $\bar{m}$ = 26, AND R = 2.4 kb/s

Interarrival Time, min	Buffer Contents, data units			Delay, min		
	Fixed	Polling[a]	Reservations[b]	Fixed	Polling	Reservations
27	1.2	0.1	0.1	1.14 (0.66)[c]	0.04	0.02
10	3.6	0.4	0.4	1.25 (0.77)	0.05	0.03
2	39.2	4.5	3.7	2.89 (2.41)	0.24	0.18
1.34	230.0	26.1	21.2	11.74 (11.26)[d]	1.23	0.98

[a]Distributed control with Nr = 14 slots and 96-bit preamble.

[b]Centralized control with a = 3 slots and zero preamble.

[c]Minimum value of delay lower bound, neglecting summation term.

[d]Summation term increases value to 11.35 minutes.

During this time the average number of arrivals is 16 units for a net utilization of 0.6, which implies that 60 percent of the time another message arrives before the previous message has been cleared. Increasing the allocation may reduce the delay somewhat, but the variation between the upper and lower bounds (Table IV-26) is small for the higher arrival rates. If the arrival process were smoother (i.e., more message arrivals but smaller lengths), this accumulative effect would be lessened.

The delay results from Table IV-26 for polling and reservations are plotted as a function of channel utilization in Fig. IV-15. The behavior with utilization is similar to that of a typical queueing system and demonstrates the sensitivity of a performance measure to variations in utilization.

Table IV-21 is repeated here to demonstrate the improvement obtainable by pooling the nets using reservations and to demonstrate the effect of the orderwire capacity. Four nets with their respective channels are pooled together by using reservations. One channel is designated as the orderwire, while the remaining three are used for data communication. The controller allocates transmission time on the three channels on a demand basis. The capacity allocated to the orderwire is excessive for this example, but may be insufficient for shorter average message lengths. By comparing the results of Tables IV-26 and IV-21, it is seen that pooling with reservations reduces the performance measures even though one channel is removed. The removal of one channel for the orderwire increases the utilization of the remaining channels and causes the system to saturate at a smaller channel utilization, 0.75, per old net.

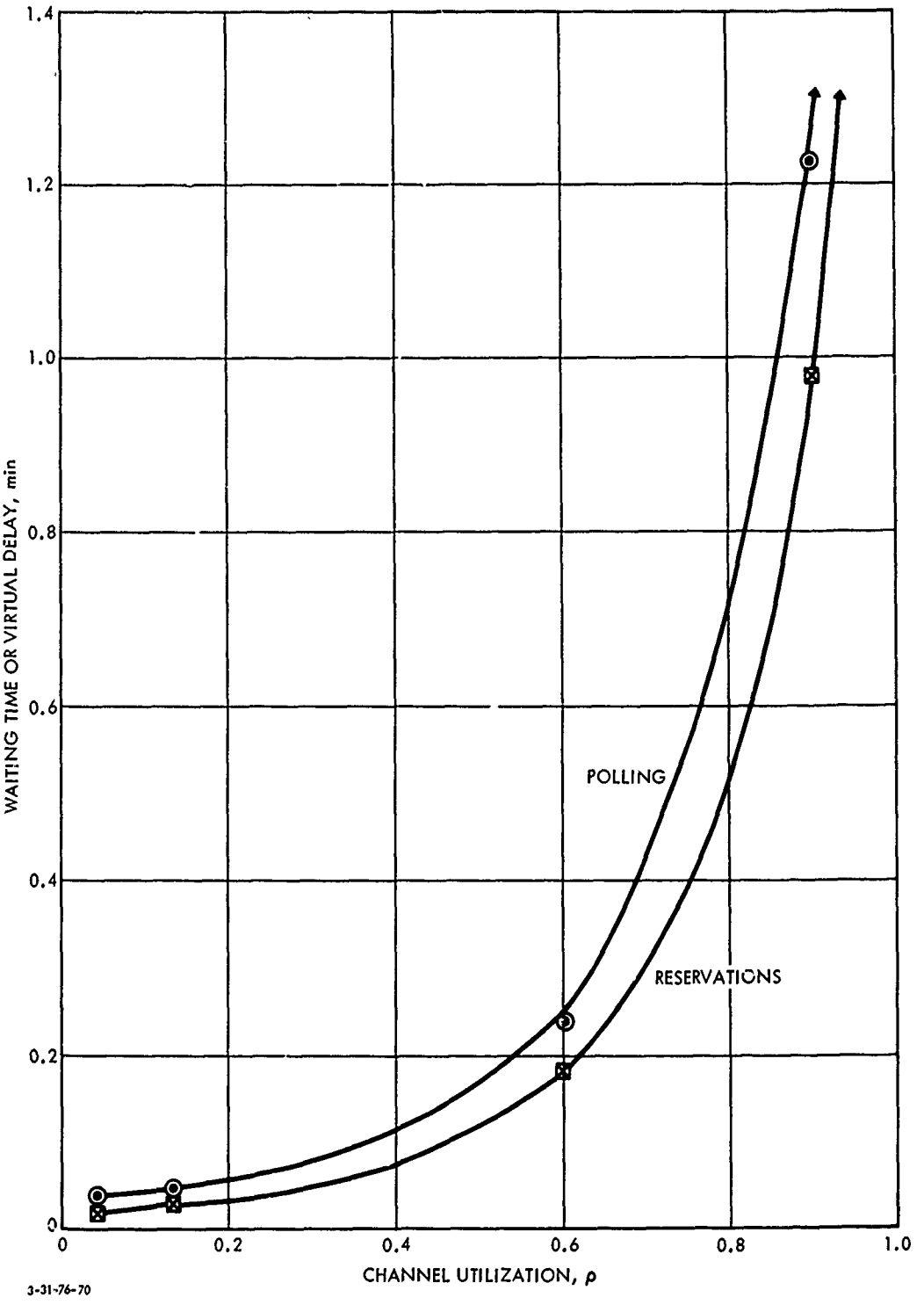


FIGURE IV-15. Effect of Channel Utilization on Waiting Time

TABLE IV-21 (REPEATED). BUFFER CONTENTS AND DELAYS WITH 44 USERS AND THREE COMMUNICATION CHANNELS

(Zero Preambles and a = 3 Slots)

Interarrival Time, min	Utilization[a]	Buffer Contents, data units	Waiting Time, min
27	0.06	0.1	0.0127
10	0.16	0.3	0.0133
2	0.805	3.2	0.136

[a]For pooled system.

Table IV-27 presents a comparison of polling and reservations for the case where synchronization preambles are transmitted. The performance measures for reservations are smaller than those for polling, except for the average interarrival time of 1.34 minutes. Although the results for fixed assignments were not calculated for the case of synchronization preambles and identical message processes, the system performance measures will be larger than those presented in Table IV-26, and therefore a polling or reservations technique is still preferable.

TABLE IV-27. COMPARISON OF POLLING AND RESERVATIONS TECHNIQUES HAVING IDENTICAL MESSAGE PROCESSES, N=11, $\bar{m}$ = 26, R = 2.4 kb/s, AND 936-bit PREAMBLE

Interarrival Time, min	Buffer Contents, data units		Delay, min	
	Polling[a]	Reservations[b]	Polling	Reservations
27	0.2	0.1	0.07	0.02
10	0.5	0.4	0.09	0.04
2	5.4	4.2	0.31	0.21
1.34	31.8	45.2	1.53	2.22

[a]Distributed control, Nr = 29 slots.

[b]Centralized control, a = 3 slots.

Table IV-28 presents a comparison of fixed assignments and reservations for the two user classes of Table IV-1 with synchronous operation. The reservations technique yields smaller buffer contents and delays, and only if the transmission rate were greater than 280 kb/s would fixed assignments yield a smaller buffer content for the higher utilization case.

F. NET CAPACITY ALLOCATION

The main interests of the previous sections were to evaluate system performance measures in the sharing of a communication channel by users organized in a net and to compare various schemes for allocating time to the users of the channel. In doing so, the number of users in the net and the transmission rate were specified, while the message arrival rates were varied (1) to determine the behavior of the schemes with respect to traffic loading and (2) to ensure that the comparisons would be valid for a wide range of situations. The system designer still must determine what net capacity (bit rates) will be assigned to the nets if the nets are predetermined, or he must determine the size and composition of a net if the corresponding bit rates have been predetermined. This problem is somewhat alleviated by having the pooled users time-share the available communication channels by access with reservations. It was also demonstrated in this case that the buffer contents and queueing times were reduced by time-sharing on one channel with capacity CR, instead of time-sharing on C channels, each with capacity R. On the other hand, there may be reasons other than queueing for having the users operate in nets, and therefore the net capacity allocation must be determined.

The allocated transmission rate affects net performance through channel utilization. For given message statistics,

TABLE IV-28. COMPARISON OF FIXED ASSIGNMENTS AND RESERVATIONS

Interarrival Time, min	Channel Utilization	Terminal Allocations for Fixed Assignments	Buffer Contents, data units		Delay, min	
			Fixed Assignments	Reservations	Fixed Assignments	Reservations
27 (10 terminals)	0.089	[a] 1	1.33	0.13	1.26	0.023
2.25 (1 terminal)		[a] 2	10.85	1.54	0.84	0.023
		[a] (average)	2.2	0.26	1.22	0.023
		[b] 1	1.57		1.50	
		[b] 4	5.59		0.38	
		[b] (average)	1.93		1.39	
10 (10 terminals)	0.242	[a] 1	4.32	0.41	1.53	0.047
0.833 (1 terminal)		[a] 3	34.77	4.89	1.01	0.047
		[a] (average)	7.09	0.81	1.48	0.047
		[b] 1	5.11		1.83	
		[b] 5	17.65		0.46	
		[b] (average)	6.25		1.71	

[a] Allocations for minimum average delay upper bound.

[b] Allocations for minimum average buffer length.

channel utilization is inversely proportional to transmission rate. With respect to the system performance measures (buffer contents and delay), it is desirable to have a low utilization because the performance measures increase with increased utilization. On the other hand, a low utilization implies inefficient use of the channel and affects the cost of the transmission facility. For example, a channel utilization ρ of 0.1 indicates that, on the average, one data unit is transmitted every 10 slots, while for $\rho = 0.6$, an average of 6 data units are transmitted every 10 slots. The performance measures for $\rho = 0.1$ are less than those for $\rho = 0.6$, but the $\rho = 0.1$ system requires six times the capacity of the $\rho = 0.6$ system. The system designer must make a tradeoff between the channel utilization, which affects the transmission facility cost, and the effects on the user (buffer contents and message waiting time).

The tradeoffs are pointed out in the following example, using the parameter values of Table IV-29. Two possible capacity allocations are (1) to provide one channel to all the users ($\rho = 0.6$) or (2) to provide 10 channels and to subdivide the users into 10 nets ($\rho = 0.06$). The respective buffer contents and delays are presented in Table IV-30 for the various time-sharing techniques and the capacity allocations. The users are allocated one data slot per net cycle in fixed assignments. Distributed control is used in polling, while a central controller with a separate orderwire is used in reservations. The increase in the average message waiting time due to increasing the channel utilization (reducing the number of channels) in this case is at least a factor of 20, except in reservations, where the increase is only a factor of 3. If the chosen time allocation scheme is either fixed assignments or polling, the system designer has to make a tradeoff between the buffer contents and delays on the one hand and the number of channels dedicated to serving this group of users on the

TABLE IV-29. ASSUMED PARAMETER VALUES

Transmission rate (R):	24 kb/s
Number of Users (N):	110
Average Interarrival Time (T):	2 minutes
Average Message Length ($\bar{m}$):	26 data units
Virtual Message Length (m):	26 data units
Data Unit Length:	608 bits
Preamble Length:	96 bits for polling 0 for fixed assignments and reservations

TABLE IV-30. SYSTEM PERFORMANCE MEASURES EVALUATED FOR N = 11 AND 110

	N	Fixed Assignments	Polling	Reservations
Buffer Length, data units	11	1.7	0.49	0.27
	110	39.2	8.6	0.48
Delay or Waiting Time, min	11	0.12	0.03	0.01
	110	2.98	0.65	0.03

other. The designer can reduce the cost of the system by decreasing the number of channels required to serve this group, but the users will have substantially larger buffer contents and delays. On the other hand, if the resultant delays are not acceptable, the designer must allocate more capacity to this group. It is extremely interesting that in reservations the increase in the performance measures were minimal, with a substantial reduction in the allocated capacity.

The results of Table IV-30 can also be used for determining the effect of the number of users on the performance measures and for a further comparison of the techniques. Reservations with N = 110 yielded delays that are smaller by a factor of 20 than the delays in polling and smaller by a factor of 100 than the delays in fixed assignments. For N = 11 and the same arrival rate, the improvements were only by factors of 3 and 12, respectively.

The results of Table IV-30, in conjunction with the results presented in Table IV-26 for the 2-minute interarrival time, can be used to compare two allocation strategies that have the same total capacity. Under the first strategy, the users are organized into 10 nets, each with 11 members and a 2.4-kb/s channel. The total capacity assigned to the group of users is 24 kb/s. Assuming that each user can transmit at 24 kb/s, the other strategy forms only one net consisting of 110 users and assigns a 24-kb/s channel to the net. The second strategy is that of merging the individual channels and pooling the users. The results, buffer contents and queueing times, are presented for both strategies in Table IV-31. The queueing time is used as a performance measure because the transmission time of a message at the 24-kb/s rate is an order of magnitude smaller than the time at the 2.4-kb/s rate. Under both strategies, the channel utilizations are identical. In a typical queueing system, combining the

TABLE IV-31. EFFECTS OF MERGING CHANNELS AND POLLING USERS (10 NETS, EACH WITH A TRANSMISSION RATE OF 2.4 kb/s)

	Buffer Contents, data units		Queueing Time, minutes	
	N = 11 R = 2.4 kb/s	N = 110 R = 24 kb/s	N = 11 R = 2.4 kb/s	N = 110 R = 24 kb/s
Fixed Assignments	39.2	39.2	3.00	2.99
Polling	4.5	8.6	0.35	0.66
Reservations	3.7	0.48	0.29	0.04

queues and increasing the service rate (analogous to merging the channels and pooling users) reduces the performance measures. If the nets are operating under fixed assignments, merging the channels has no effect on the performance measures, while for reservations, as expected, merging produces an order-of-magnitude improvement. Under access with polling, surprisingly, combining the channels and polling the users degrades the performance by almost a factor of 2. This is caused by the fact that the time for wnich the channel is not available to a user increases by an order of magnitude.

In designing the communication satellite system, the designer must decide what values of the buffer contents and delays are acceptable, which depends on the particular situation, and he must then allocate sufficient capacity to satisfy these design criteria. The number of channels that must be allocated to the user group with the parameter values of Table IV-29 to satisfy a delay criterion of either 0.03 or 1.0 minute is presented in Table IV-32. The transmission time of

TABLE IV-32. NUMBER OF CHANNELS REQUIRED TO ACHIEVE DELAY CRITERIA WITH N = 110 AND T = 2 MINUTES

Technique	Number of 24 kb/s Channels		
	A[a]	B[b]	Queueing Time, min
Fixed Assignments	30	2	0.85
Polling	10	1	0.66
Reservations	2[c]	1[d]	0.07
Random Access[e] (unslotted)	5	5	

[a]A: Delay criterion of 0.03 minute.

[b]B: Delay criterion of 1 minute.

[c]One data channel plus an orderwire channel.

[d]Data channel capacity = 20 kb/s plus a 4 kb/s orderwire.

[e]Lower-bound estimate using saturation value of 0.136.

TABLE IV-33. CAPACITY TO ACHIEVE 0.3-MINUTE QUEUEING TIME WITH N = 11 AND T = 2 MINUTES

Technique	Transmission Rate, kb/s
Fixed Assignments	11
Polling	2.6
Reservations	3.0[a]
Random Access (unslotted)	11[b]

[a]A 2.4-kb/s data channel plus a 600-b/s orderwire.

[b]Lower-bound estimate using saturation value of 0.136.

an average message is 0.01 minute at the transmission rate of 24 kb/s. Although the delay in a contention system was not determined, the minimum number of channels required to provide stable service can be estimated by ensuring that the channel utilizations are less than the saturation value (see footnote, p. 122, Section IV-A). The number of channels required may be reduced by using a slotted contention system. For the delay criterion of 0.03 minute, which is three times the average message transmission time, the reservations technique permits a saving of at least eight 24-kb/s channels over fixed assignments and polling. For the delay criterion of 1 minute, which is half the average message interarrival time, both polling and reservations require one channel, but the queueing time (waiting time plus transmission time) is an order of magnitude smaller than in polling. The queueing time is used here instead of the waiting time because the message transmission time is larger in the reservations system. The queueing value for reservations is obtained by using the expression in Appendix E that takes into account blocking on the orderwire.

If the users with the parameter values in Table IV-29 are partitioned into 10 nets, each with 11 users, the capacity allocations of Table IV-33 are required to achieve a queueing time criterion of 0.3 minute per net. Either polling or reservations using distributed control requires the least capacity to achieve this criterion and reduces the required capacity by about 8 kb/s per net below fixed assignments, polling requiring slightly less capacity than reservations.

After the system has been functioning for a while and new users have joined the respective nets, the capacity allocations to the nets can be reexamined to improve performance under the new conditions. Assume that the total capacity of the system, R, can be linearly subdivided into capacity allocations to

each net, R_i, such that $R = \sum_{i=1}^{N} R_i$, where N now refers to the number of nets. Overall performance measures, such as the net buffer contents averaged over the nets and the queueing time averaged over the nets, can be used to evaluate the net capacity allocations. The selected net capacity allocations are those that minimize the overall performance measure under the total capacity constraint.* This is similar to the optimization of the net performance measure with respect to the time allocation r_i to the user, but the user buffer contents and delays are replaced by the net buffer contents or queueing time. Unfortunately, as in the time-allocation case, no analytical solution could be found, and therefore the optimization would have to be performed numerically. The optimization was not performed because resources were limited and because the optimization would have to be performed numerically for a specific situation.

G. CONCLUSIONS

In a net operating under fixed assignments, the user allocations must be predetermined and should be selected carefully because both system performance measures are sensitive to the allocations. In the examples considered where more than one user class exists, the optimum allocations are not only different from the allocations obtained by an intuitive approach but also reduce the resultant delay by about 50 percent below the levels obtained by using intuitive allocations. In either the polling or reservation system, user allocations are dynamic and intrinsic to the technique and therefore require no optimization. On the other hand, the (minimum) capacity allocation

*The problem formulation in Ref. 42 is similar to this, but the results are not applicable.

to the net in any technique must be determined to ensure stable operation. This requires an estimate of the expected average message traffic load in the net.

Overhead can be characterized as of two types. One type, represented by synchronization preambles and transmission guard times, is independent of the allocated transmission times. The other type, represented by the time reserved for new entry requests in the cyclic techniques and the orderwire capacity in reservations, is dependent upon allocated transmission times. Both types of overhead reduce the effective data transmission rate and hence reduce the grade of service provided; the second type also reduces the message traffic load, resulting in saturation. In the cyclic techniques, the inefficiencies induced by allocation-independent overhead can be reduced by allocating longer transmission times, and the effect on the saturation level can be made small (actually, there is no saturation effect in polling). In reservations, because the allocated transmission times are determined by message lengths, the effect of the preamble overhead is not reduced but increases the channel utilization and also reduces the saturation level. The reservation technique is more sensitive to this overhead than the cyclic techniques.

In the sample comparisons for a fixed capacity allocation to the net, polling and reservations both show significant improvements over fixed assignments in the grade of service provided. The performance measures are shown to increase the utilization, and at higher levels of utilization they are tremendously sensitive to small changes in utilization. This behavior is typical of a queueing system and should be taken into account in the choice of an operating point (in channel utilization or in capacity allocation to the net).

The reservation technique increases flexibility and further improves the grade of service provided because it permits the pooling of users and the sharing of available channels among pooled users when the users cannot transmit at a higher rate than the individual channel rate. Further, if the users can transmit at the total capacity of the channels (sharing the merged channels), the grade of service is improved significantly only in reservations, while it degrades in polling.

As would be expected from the above results, polling and reservations in the examples considered require significantly less capacity than fixed assignments to achieve a grade of service.

Finally, the general conditions under which the techniques are preferred with respect to buffer-contents performance measures (i.e., the conditions under which the techniques yield lower values) are obtained by analytical comparisons of the respective buffer-contents equations. In comparing polling and fixed assignments, which are both cyclic time-sharing techniques, smaller buffer contents are obtained by using polling, except when the time to transfer control from one user to another in polling is about twice the average message transmission time. The only parameters that determine the preferred regions are the transmission rate and the average message length. On the other hand, in comparisons of cyclic techniques with reservations, all of the parameters--number of users, channel utilization, average message length (except in polling), and transmission rate--determine the preferred regions. (The orderwire capacity required for reservations is not included in these comparisons.) Polling is preferred over reservations for those systems where the time to effect a capacity assignment in reservations is larger than half of the total average time the channel is not available to the user in polling. Fixed assignments is preferred over reservations for those systems

where the time to effect a capacity assignment is larger than the product of the number of users and the average message transmission time, divided by one minus the channel utilization. The cyclic techniques are at a disadvantage with respect to reservations for large numbers of users. The cyclic techniques are definitely preferred over reservations when the number of users in the net and the average message transmission time are small.

REFERENCES

1. Institute for Defense Analyses, *A Review of Analytical Techniques and Data for Assessing the Effects of Striations on Satellite Communication Links in a Post-Nuclear-Event Environment*, IDA Paper P-1154, W. Wasylkiwsky, in publication.

2. Military Satellite Communication Systems Office, Defense Communications Agency, *Draft Military SatCom System Architecture*, November 1975.

3. J.W. Schwartz, J.M. Aein, and J. Kaiser, "Modulation Techniques for Multiple Access to a Hard-Limiting Satellite Repeater," *Proc. IEEE*, Vol. *54*, No. 5, May 1966, pp. 763-777.

4. Joint Tactical Communications Office, Fort Monmouth, N.J., *System Specification for a Demand Assigned, UHF TDMA Satellite System*, TT-A2-2206-0029, 15 January 1975.

5. Computer Sciences Corporation, Falls Church, Va., *Systems Level Analysis for Demand-Assigned UHF Time-Division-Multiple-Access Satellite System*, Vol. *I*, CSC/TR-75/3012, December 1974.

6. Lincoln Laboratory, Lexington, Mass., *A Preliminary Design of a TDMA System for FLEETSAT*, Technical Note 1975-5, J.D. Bridwell and I. Richer, 12 March 1975.

7. Raytheon Company, Sudbury, Mass., *TDMA*, Document 48461-31, Final Report, Contract DAAB07-72-C-0176, CDRL Item F002.

8. Comsat Laboratories, *Interim Report on Study of Functional Requirements for Demand Assigned SHF TDMA Modems*, J. Lucas, May 1975.

9. W.G. Schmidt, "The Application of TDMA to the INTELSAT IV Satellite Series," *Comsat Technical Review*, Vol. *3*, No. 2, Fall 1972, pp. 257-275.

10. Institute for Defense Analyses, *Multiple Access to a Communication Satellite with Hard-Limiting Repeater*, IDA Report R-108, Vol. I: *Modulation Techniques and Their Application*, J. Kaiser, January 1965, Vol. II: *Proceedings of the IDA Multiple Access Summer Study*, J.M. Aein and J. Schwartz, April 1965.

11. Institute for Defense Analyses, *Processing Communication Satellites: Proceedings of the IDA/RESD Summer Study 1966*, IDA Study S-268, J.M. Aein and J. Schwartz, July 1967.

12. Comsat Laboratories, *Interim Report on Study of Functional Requirements for Demand Assigned SHF TDMA Modems*, DAAB-07-74-C-0204, May 1975.

13. R.B. Cooper, *Introduction to Queueing Theory*, MacMillian Company, New York, 1972.

14. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I, Second Edition, John Wiley & Sons, New York, 1950.

15. J.W. Cohen, "The Generalized Engset Formulae," *Phillips Telecommunication Review*, November 1957, pp. 158-170.

16. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Vol. I, Part 1, Oliver and Boyd, Edinburgh and London, 1960.

17. L. Kleinrock, *Queueing Systems*, Vol. I: *Theory*, John Wiley & Sons, New York, 1975.

18. O.A. Pederson, "The Design of Gradings with Small Interconnection Numbers for Random Hunt Selectors," *IEEE Trans. on Comm.*, Vol. *COM-23*, July 1975, pp. 714-721.

19. L.A. Gimpelson, "Analysis of Mixtures of Wide- and Narrow-Band Traffic," *IEEE Trans. on Comm.*, Vol. *13*, No. 3, September 1965, pp. 258-266.

20. Eckart Wollner, "Queueing Problem in Data Transmission," *Proceedings of the Seventh International Teletraffic Congress*, Swedish Telecommunications Administration (Televerket), Stockholm, Sweden, 13-20 June 1973.

21. Gunnar F.W. Frederickson, "Analysis of Channel Utilization in Traffic Concentrators," *IEEE Trans. on Comm.*, Vol. *COM-22*, No. 8, August 1974.

22. W.W. Chu and A.G. Konheim, "On the Analysis and Modeling of a Class of Computer Communication Systems," *IEEE Trans. on Comm.*, Vol. *COM-20*, No. 3, June 1972.

23. J.F. Hayes and D.H. Sherman, "Traffic Analysis of a Ring Switched Data Transmission System," *The Bell System Technical Journal*, Vol. *50*, No. 9, November 1971.

24. E. Fuchs and P.E. Jackson, "Estimates of Distribution of Random Variables for Certain Computer Communications Traffic Models," *Communications of the Association for Computing Machinery*, Vol. *13*, No. 12, December 1970. (Also in Ref. 43).

25. A.G. Konheim, "Service Epochs in a Loop System," *Proceedings of Symposium on Computer-Communications Networks and Teletraffic,* Polytechnic Institute of Brooklyn, April 4-6, 1972.

26. A.G. Konheim and B. Meister, "Service in a Loop System," *Journal of the Association of Computing Machinery,* Vol. *19,* No. 1, January 1972.

27. A.G. Konheim and B. Meister, "Waiting Lines and Times in a System with Polling," *Journal of the Association for Computing Machinery,* Vol. *21,* No. 3, July 1974.

28. A.G. Konheim, "Chaining in a Loop System," *IEEE Trans. on Comm.,* Vol. *COM-24,* No. 2, February 1976.

29. N. Abramson, "Packet Switching with Satellites," *Proceedings of National Computer Conference,* 1973. (Also in Ref. 43)

30. J.D. Spragins, "Loop Transmission Systems--Mean Value Analysis," *IEEE Trans. on Comm.,* Vol. *COM-20,* No. 3, June 1972.

31. University of California, Los Angeles, *Packet Switching in a Multi-Access Broadcast Channel with Application to Satellite Communication in a Computer Network,* Technical Report UCLA-ENG-7429, S.S. Lam, April 1974.

32. L. Kleinrock and S.S. Lam, "Packet Switching in a Multi-Access Broadcast Channel: Performance Evaluation," *IEEE Trans. on Comm.,* Vol. *COM-23,* No. 4, April 1975.

33. A.B. Carleial and M.E. Hellman, "Bistable Behavior of ALOHA-Type Systems," *IEEE Trans. on Comm.,* Vol. *COM-23,* No. 4, April 1975.

34. S.S. Lam and L. Kleinrock, "Packet Switching in a Multi-Access Broadcast Channel: Dynamic Control Procedures," *IEEE Trans. on Comm.,* Vol. *COM-23,* No. 9, September 1975.

35. M.J. Ferguson, "On the Control, Stability and Waiting Time in a Slotted ALOHA Random-Access System," *IEEE Trans. on Comm.,* Vol. *COM-23,* No. 11, November 1975.

36. R.A. Kaye, "Analysis of a Distributed Control Loop for Data Transmission," *Proceedings of Symposium on Computer-Communications Networks and Teletraffic,* Polytechnic Institute of Brooklyn, April 1972.

37. C.D. Pack and B.A. Whitaker, "Approximate Access Delay Computations in Multipoint Private Line Networks with Polling," *Proceedings of National Telecommunications Conference,* Atlanta, Georgia, November 1973.

38. L.G. Roberts, "Dynamic Allocation of Satellite Capacity through Packet Reservations," AFIPS Conference Proceedings, Vol. *42,* 1943. (Also in Ref. 43)

39. Computer Sciences Corporation, *System Level Analysis for Demand-Assigned UHF Time-Division-Multiple-Access Satellite System*, CSC/TR-75/3012, Vol. I, December 1974.

40. D.R. Cox and W.L. Smith, *Queues*, John Wiley & Sons, New York, 1961.

41. D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*, John Wiley & Sons, New York, 1974.

42. L. Kleinrock, *Communication Nets*, McGraw-Hill Book Company, New York, 1964.

43. W.W. Chu, *Advances in Computer Communications*, Artech House, Incorporated, Dedham, Massachusetts, 1974.

APPENDIX A

TASK ORDER

OFFICE OF THE SECRETARY OF DEFENSE
DIRECTOR, TELECOMMUNICATIONS AND
COMMAND AND CONTROL SYSTEMS
WASHINGTON, D.C. 20301

ASSIGNMENT FOR WORK TO BE PERFORMED
BY
INSTITUTE FOR DEFENSE ANALYSES

Date: 29 AUG 1974

You are hereby requested to undertake the following task; D-10

1. TITLE: Telecommunications Studies

2. TECHNICAL SCOPE:

The objective of this task is to support the development and evaluation of military command and control concepts and techniques that use satellite relays. This task will emphasize the development of satellite terminal network control concepts and identification and evaluation of key technical problem areas. The most important part of this task will be to perform analyses, evaluations and to draw conclusions and make technical recommendations in the area of military satellite communication systems important to command and control of U.S. forces. The studies will include, but are not limited to, the following:

a. A study of the relative impact of nuclear event-induced propagation degradations to SatCom systems supporting command and control of General Nuclear War Forces. Estimate the benefits and effort required to convert 225-400 MHz SatCom WWMCCS terminals to 1-2 GHz with appropriate modulation in order to reduce the possibility of nuclear event-induced earth/space communication outages.

b. A study of the technical means for achieving, allocating, monitoring and controlling efficient, flexible, and timely netting of mobile SatCom radio terminals in the general purpose forces.

3. SCHEDULE:

This task will continue through FY 1975 with the items to be completed and reports submitted at such times as sufficient study has been completed to render a meaningful report. Periodic meetings should be scheduled with offices having technical cognizance to discuss progress and matters of mutual interest during the study periods.

4. TECHNICAL COGNIZANCE:

Assistant Director (General Purpose Systems), Office of Director, Telecommunications and Command and Control Systems (DTACCS).

5. SCALE OF EFFORT:

Approximately two and one-half man-years of effort, including consultants as required by IDA, is authorized for this task, provided that total expenditures shall not exceed $160,000 (one hundred sixty thousand dollars). Changes in the scale of effort will not be made without the consent of DTACCS.

6. REPORT DISTRIBUTION AND CONTROL:

The DTACCS will determine the number of copies of reports and their distribution. A "need-to-know" is hereby established in connection with this task and access to classified documents and publications, security clearances and the like, necessary to complete the task, will be obtained through the DTACCS.

Thomas C. Reed
Director, Telecommunications and
Command and Control Systems

ACCEPTED:

Alexander H. Flax, President
Institute for Defense Analyses

Date: 9 September 1974

APPENDIX B

CURRENT DOD SATCOM PROGRAMS

CONTENTS

APPENDIX B

CURRENT DOD SATCOM PROGRAMS

A. OVERVIEW

The current DoD satcom system programs* are the Defense Satellite Communications System (DSCS), Air Force Satellite Communications (AFSATCOM), and Fleet Satellite Communications (FLEETSATCOM).

The DSCS is an operational super-high-frequency (SHF) system under direction of the Defense Communications Agency (DCA) and is characterized by very wide-band (multi-megahertz) satellite transponders and earth terminals with relatively large directive antennas. It is currently used for wide-band digital circuits in point-to-point trunk telephony. Although the DSCS is in the development phase, general-purpose forces have not been equipped with terminals to use this resource.

The AFSATCOM program is under development by the Air Force to provide service between the following:

- The National Command Authorities (NCA)/Airborne Command Post (ABNCP) and Single Integrated Operational Plan (SIOP) force elements [Emergency Action Message (EAM) force direction and report-back]
- ABNCPs (in CINCNET)
- Presidential communications

and, on a not-to-interfere basis,

*Reference 1 contains additional summary information on these systems.

- Other long-range aircraft communications (e.g., Military Airlift Command and global weather).

AFSATCOM is an ultra-high-frequency (UHF) system consisting of narrow-band (5-kHz) transponders placed on host satellites, including the FLEETSAT spacecraft. The system is designed principally for aircraft terminals using low-gain nonsteered antennas. It basically provides narrow-band teletype service. However, to support CINCNET and Presidential communications, an additional wide-band channel is provided on the FLEETSAT spacecraft.

The FLEETSATCOM program is under direction of the Navy. Its spacecraft development is managed by the Air Force. Like AFSATCOM, FLEETSATCOM is also a narrow-band UHF system utilizing ten 25-kHz-wide transponders per FLEETSAT spacecraft. Terminals for ship, aircraft, submarine, and shore installations with a variety of antenna configurations are also in development in the pre-production phase. The channel bandwidth per transponder is wide enough to accommodate a variety of communication services, ranging from teletype through kilobit digital circuits capable of supporting secure voice. One transponder per satellite is dedicated to supporting the Fleet Broadcast function. The other nine transponders represent orbital assets that can be assigned to support various networks of Naval terminals with a variety of communication services. The current Navy planning for use of FLEETSATCOM is presented in Refs. 2 and 3.

Of the three major current DoD satcom programs, the one that most immediately determines the general framework for this study is FLEETSATCOM. Its example serves as a study model for multi-user class capacity allocation. As the DSCS program develops, and especially if transpondable light and mobile truck-mounted terminals are deployed to the Field Army, more advanced traffic planning needs for system use will be manifested.

The mathematical methodology for this study, although initially developed from a consideration of the FLEETSATCOM system, should eventually be of equal interest to DSCS. The AFSATCOM system, with its primary role of supporting command and control of U.S. nuclear-capable forces and with the special features attending such communications, played little part in shaping the study effort reported here. However, as the AFSAT program evolves, a need could develop to support the nuclear force elements with fewer transponders per satellite. Should this occur, study of store-and-forward data exchange techniques applicable to AFSATCOM missions would become necessary.

B. AN EXAMPLE: TECHNICAL FACTORS AFFECTING CAPACITY ALLOCATION FOR FLEETSATCOM

The planned FLEETSATCOM usage developed to date provides an excellent case study with which to relate resource-utilization study concepts to real hardware issues and engineering design. This example is provided here only for the purpose of demonstrating these real engineering relationships. No evaluation of this program was undertaken and none is intended here. At this time the program focus is still on acquisition of spacecraft assets and basic terminal RF components (transceivers, modems, and antennas). The current plans summarized here present a baseline concept of operation developed by the Navy. Considerable flexibility still remains in organizing the refinement and use of the assets. In point of fact, the Navy has recognized this and has the Lincoln Laboratory (Refs. 4-6) advising it on developing more advanced usage plans as well as second-generation components [specifically, spacecraft modifications and an advanced time-division multiple-access (TDMA) modem]. The Navy has also tasked the Computer Sciences Corporation (Ref. 7) to study demand-assignment TDMA for FLEETSAT.

1. System Organization

To review briefly, the Navy portion of the FLEETSATCOM system consists of a set of spacecraft (nominally, two to four in orbit) on each of which there are ten independent 25-kHz-wide, hard-limited UHF transponders on separated frequencies. Of the ten, one transponder is dedicated to providing the Fleet Broadcast service,* leaving nine transponders for other uses. For the Fleet Broadcast service, standard (AN/SSR-1) receive-only terminals are being purchased for deployment to all Naval elements afloat.

Separate from and in addition to the Fleet Broadcast receivers, procurement of terminals for deployment to ships, shore-based long-range patrol aircraft (typically P-3), submarines, and shore stations is under way. Variations in antenna gain and local ambient RF noise background at the terminals will require variations in the transmitter power and receiver sensitivity from terminal to terminal, and, in addition, the transmission system must meet the various data speed or bandwidth requirements of the planned communication services.

The communication services other than Fleet Broadcast are grouped (Ref. 2) into Digital Information Exchange Systems (IXS) and Secure Voice Communications. The store-and-forward type of data service is characterized by the Common User Digital Information Exchange Subsystem (CUDIXS)** and the Small Ship

---

*This service takes primary precedence. Should the assigned transponder fail, there are alternates amongst the remaining nine which can be seized for the Broadcast service.

The Broadcast signal is injected from a designated shore station, nominally one plus a backup per ocean area. For the purposes of this study, the Broadcast portion of the system is of no further direct interest, other than to note that it serves in some cases to partially reduce the shore-to-ship traffic demand on other transponders.

**The CUDIXS data transmission concept of Ref. 3 strongly focused the effort reported in Chapter III.

Teletype Information Exchange System (SSTIXS). These two systems are very similar, and their data exchange processes can be jointly modeled as computer polled terminals. They have at various times been planned and implemented as separate and/or integrated services; in the latter case, SSTIXS is referred to as "CUDIXS Primary" and CUDIXS is referred to as "CUDIXS Special Subscriber."

The Tactical Data Information Exchange System (TADIXS) is a shore/ship/air high-speed digital circuit for bulk data transfer, as is the Tactical Support Center Information Exchange Subsystem (TSCIXS) for shore/air ASW data exchange. TACINTEL is a special digital circuit capability. There are also two technically similar secure voice circuits, one, the High Command Network (HICOM), for interfacing Automatic Secure Voice Communications (AUTOSEVOCOM), and the other, the Fleet Command Voice Network (FLTCOM). Finally, there is a Submarine Satellite Information Exchange Subsystem (SSIXS).

These communication services aggregate as (1) store and forward, (2) digital circuit, and (3) special or denied. The role of the third category is to reduce overall available capacity and could include events such as transponder failure (permanent) or preemption by the NCA (temporary failure). The current plan has nine communication services and *physically assigns one satellite transponder to each service.* As shown in the next subsection, any terminal (e.g., major flagship or capital ship) needing to participate in more than one service with the currently developed equipment must acquire multiple radios (sharing the same antenna), one for each communication service in which it participates. Perhaps even more significant is the possible temporary or permanent loss of satellite transponders. Under the present plan, one or more of the services could not be provided if a transponder or transponders were lost.

Clearly, there is motivation to seek means to pool and share transponder capacities among the communication services to maintain service continuity in the event of transponder loss*. Against future growth in terminals and traffic, capacity concentration and load balancing may also achieve enhanced efficiency. For FLEETSATCOM, the TDMA technique can provide the means for sharing transponders. A TDMA modem also can provide RF multiplex hardware capability to allow multichannel operation by a simple terminal transceiver.

2. System Terminal Configuration

The FLEETSAT terminals have been organized into three basic subsystems: an antenna assembly, a WSC-3 radio transceiver, and a complex of digital base-band equipments. The antenna system types fall into the categories of (1) omnidirectional antenna ($\sim$0 dB) suitable for aircraft, (2) coarsely steered low- to medium-gain ($\sim$12 dB) antenna suitable for ships, and (3) a pointed medium-gain ($\sim$18 dB) antenna suitable for shore stations. The antenna system contains RF components in addition to a low-noise wide-band receiver preamplifier (e.g., diplexer and multicoupler) that will permit full duplex operation of the terminal and the simultaneous connection and operation of up to four separate radio transeivers ("stacked" WSC-3). An explanation of the need for this is provided in what follows).

The WSC-3 radio transceiver is standard to all the terminals and consists of a 100-watt transmitter, a receiving chain (receiver, downconverter, IF), a frequency synthesizer, and a set of internal simple modems (AM, FM, FSK, and PSK). The transmitter and receiver provide 30-kHz narrow-band channels with low spurious radiation and high-selectivity performance. The transceiver can be tuned by the frequency synthesizer to any of 7000 channels in the 225-400 MHz band. The WSC-3 is

*In this regard, when capacity is preempted it is desirable to yield only that amount which is needed, rather than a whole transponder (or, more generally, an integral number of transponders).

currently configured to operate only in a half-duplex mode, i.e., to transmit and receive but not simultaneously. A set of simple modems is provided so that the WSC-3 can also be employed from conventional UHF line-of-sight use (e.g., ship/air) with voice (AM and FM) and data link (Links 4 and 11).

The internally supplied simple binary-phase-shift-keyed (BPSK) modem is intended for satellite use and can operate at any one operator-selectable bit rate of 75,300, 1200, 2400, 4800, or 9600 bits/sec. The BPSK modem is a relatively simple unit with no data multiplexing capability on the base band and no multiple-access capability to share a satellite transponder simultaneously with another transmitting terminal. In addition, a standard 70-MHz IF interface is provided for attachment of any external narrow-band (25-kHz) modem desired.

The base-band output from the WSC-3 internal modem or an external modem then attaches through crypto units to a variety of base-band digital equipments, including voice compression encoders, teletype, and automatic data handling devices (including a minicomputer). The constellation of such devices depends on the type of terminal and the networks in which the terminal participates. It is important to note the wide diversity possible in functions and arrangements of base-band devices (see for example Ref. 2). The procurement, installation, and software development costs for larger terminals can easily match or even exceed the satellite terminal RF/modem costs.

Without care, information handling equipment architectures can result which tend to increase the number of satellite accesses needed. In the present state of system development, without any further upgrading of modem and (WSC-3) radio transceiver, any terminal wishing to "simultaneously" transmit and receive on more than one RF carrier/satellite transponder (i.e., to

participate in two or more functional communication services or networks) must acquire for each carrier/transponder a separate WSC-3 modem and base-band chain. This explains the provision in the antenna subsystem for connecting multiple or "stacked" WSC-3 transceivers.

The Navy is currently studying several TDMA modem designs (Refs. 5, 7, and 8). Because of the narrow-band hard-limiting characteristic of the FLEETSAT transponders, TDMA is the only multiple-access modulation option available for the sharing of transponders among different communication nets or services. An important factor in regard to TDMA design is the fast returning capability of the synthesizer in the WSC-3. Automatic circuitry to command and execute frequency returning does not currently exist in the WSC-3 but can be added. A fast-retune or frequency-hopping capability in the WSC-3 would allow a TDMA modem the flexibility of accessing on a burst-by-burst basis any one of the (up to nine) available FLEETSAT transponders.* TDMA without frequency-hopping capability allows the sharing of a single transponder. TDMA with frequency hopping allows sharing across several transponders. This provides functional capacity pooling by electrically gluing transponders together.

*Full-duplex capability would facilitate hardware execution of this important capability, alleviate software needs, and reduce blocked availability of time slots. Discussions with Electronic Communications, Inc., of St. Petersburg, Florida, the current WSC-3 manufacturer, suggest no technical impediment to upgrading at modest cost to full-duplex capability. This can even be done in the field to deployed units. Furthermore, Collins Radio of Cedar Rapids, Iowa, has under development a Navy Growth Radio [AN/ARC-178(V) and AN/URC-93(V)] with satcom-compatible features, including full-duplex and rapid-retune features.

In addition to serving as a multiple-access means of physically gluing together and then of flexibly subdividing transponder capacity, TDMA can also serve to multiplex the RF and baseband services. This would reduce the need for stacked WSC-3 operation. (The cost savings in a reduced WSC-3 buy should be evaluated to offset the added costs of a more complex modem.) Because of the single instantaneous RF carrier generated at any one time, TDMA can mitigate any potential problems of intermodulation noise generated by simultaneous RF frequencies interacting with nonlinearities in or near the transmitting terminal antenna.

Thus, TDMA represents an approach for evolving enhanced transmission capacity with potential for flexible capacity allocation. The conceptual advantages outlined are not without certain challenges in modem design. These are outlined in the next subsection. Even so, there will still remain the problems of network traffic organization and capacity allocation methodology. To those problems this report addresses itself. It is restated for emphasis that the two extremes of fixed, static, preassigned capability (transponders, channels, capacity) and full demand access are too limiting. The "electrical" advantages of TDMA (or, more generally, flexible multiple access) must be matched with a suitable traffic engineering methodology.

3. TDMA Design Issues

Some of the design issues associated with developing a TDMA modem capability are briefly reviewed here to show the relationship of transmission functional areas to hardware subsystems. For further detail, the interested reader is referred to Refs. 5, 7, and 9.

The modem can be thought of as being organized into three subsystems:

1. IF interface, signal acquisition, and signal modulation/demodulation
2. User base-band interface; modem I/O buffering
3. Transmission network and modem control.

The first subsystem is "classical" and has been given the lion's share of attention. The second and third subsystems have only recently come to be recognized as major design areas because they contribute as much to development costs as the first subsystem, if for no other reason. What is even more important, however, is that subsystems 2 and 3 determine the limits of network flexibility and the degree of transparency to the base-band user, while subsystems 1 and 3 provide the physical mechanisms ("actuators") of transmission control.

The typical problem areas of subsystem 1 are as follows:

- Selection of signal waveform and timing formats: time slots, frames, waveform selection, bit rate, coding, preambles, and unique words
- RF/IF filters: channel equalization, adjacent channel interference, and intersymbol interference
- Signal burst acquisition: carrier recovery, slot/frame recovery, and ranging
- Bit detector: sample timing, sampler or integrate and dump.

The separate problem areas become interactive as the ratio of RF bandwidth to bit rate is reduced to 1/2.* There is no adequate analytic theory of behavior in this regime. Design becomes empirical, relying on computer simulation and laboratory experiment. At ratios of one or more (for FLEETSAT transponders, bit rates less than 25 kb/s) the problems can be addressed independently, and adequate theory does exist for each problem area of design.

*This assumes a quaternary-phase-shift-keyed (QPSK) signal. For binary-phase-shift keying (BPSK), interactive problems develop at ratio values of one, but BPSK sensitivity to the interaction is less than that of QPSK.

It is the objective of the equipment of subsystem 2 to "manage" time and logical transfer between the bits flowing into and out of the "pure modem" of subsystem 1 and the user base-band equipment. Subsystem 2 must smooth out in time and organize in sequence (multiplex) the bits arriving from and departing to base-band equipments and operating modes. Included here would be interfaces with communication security devices.

The smoothing and organizing functions are performed in a memory system (I/O buffers) very similar in concept to those used for computers. The complexity of this memory system depends on the flexibility desired in the use of transmission capacity and the variety and adaptability to user base-band equipment terminal organization (e.g., see Fig. B-1, which would be representative of a TDMA modem single WSC-3).

The simplest memory system would be a single independent buffer for each base-band digital voice line which would compress the continuously flowing voice bits for burst-mode insertion into the TDMA time slots. As the number of base-band lines or "ports" are added, more buffers and different buffer actions are required. Thus, one must use buffer sharing and provide flexible buffer characteristics (e.g., length and I/O speed). These are problems analogous to a computer I/O with peripherals.

There are many options for physical realization and organization of the memory system, all transferable from computer-associated technologies.* These include hardware features of solid-state memories and microprocessor chips, together with software concepts of memory segmenting or paging and data base management algorithms for efficient use of shared memory space and speed of memory I/O. Selection of a "memory system" will be as much creative as deductive, and perhaps more creative.

Overlaying the core of the memory system must be the logical "process" that not only manages data movement into and

*In fact, all of the design areas in subsystems 2 and 3 are highly analogous to computer architecture and design technique.

out of memory but must interface with logical processes of the base-band devices (e.g., crypto, teletype, and printer), which require a common dialog as to device control, clocking, character coding formats, and procedures (e.g., start, stop, and carriage return).

Interacting with the device control process of subsystem 2 is the transmission system control of subsystem 3. The technology and software logic techniques of subsystem 3 are similar to those of subsystem 2 but are oriented toward the transmission system rather than the user base-band system. Subsystem 3 can be separated into (a) the terminal access control and (b) management of the TDMA time slots. Network access control includes all-channel signaling* and addressing functions, as well as arrangements and procedures for granting (or effecting) the allocation of system transmission capacity. Management and reassignment of the TDMA slots provides for enhanced flexibility of slot-to-terminal assignment according to terminal loading. For example, it is desirable to avoid slot fragmentation by contiguously grouping slots assigned to any one terminal. Thus, as terminal traffic varies, slots should be periodically reassembled to eliminate fragmentation** inefficiencies.

In point of fact, a strong similarity exists between a TDMA transmission system and a computer memory. The number of simultaneous operating frequencies (i.e., transponder channels) times the number of basic (i.e., shortest) time slots in a frame serves to define the number of memory locations. "Data" is periodically put in and retrieved sequentially by independent users (terminals) at the frame rate of the TDMA. Some terminals

*Call setup, takedown, busy, restoration, accounting.

**It is interesting to note that a very analogous problem occurs in managing computer memories. The process of memory aggregation or reaggregation to eliminate fragmentation is referred to as "garbage collecting."

can have more than one entry and/or more than one memory location per frame. Thus, the organization and access by the TDMA terminals is amenable to methodology developed in computer science for computer memory architecture.

The three modem subsystems relate in the following manner to the modeling used here. Subsystem 1 (the conventional modem) serves to achieve system capacity and defines the allocable capacity units per terminal. Subsystem 2 serves to manage and multiplex the users sharing a terminal, as opposed to the satellite sharing. This problem area is not addressed here,* it being assumed that a design can be achieved that provides relative transparency. Subsystem 3 serves to effect control of capacity allocation. By analogy to servomechanism controls, subsystem 1 is the plant and controllable variables, subsystem 2 is the inputs/outputs, and subsystem 3 is the actuators/servos. Missing is the method for designing feedback compensation and loop filters and evaluating performance.

*Subsystem 2 deserves further study. It can be viewed as resource management of the terminal "capacity" or "baseband demand assignment." The principal interest in this study was in resource management of satellite capacity.

REFERENCES, APPENDIX B

1. Military Satellite Communication Systems Office, Defense Communications Agency, *Draft Military SatCom System Architecture*, November 1975.

2. Naval Electronics Laboratory Center, San Diego, California *FLTSATCOM System Description (U)*, Technical Document 280 (Revision 1), 10 July 1974 (CONFIDENTIAL).

3. Navy Space Project Office (PME 106), Naval Electronic Systems Command, *The Fleet Satellite Communications Program (FLTSATCOM) Information Exchange Systems CUDIXS and SSICS System Description (U)*, (CONFIDENTIAL).

4. Lincoln Laboratory, Lexington, Massachusetts, *Options for FLEETSAT Growth (U)*, Technical Note 1975-2, S. L. Bernstein, 31 March 1975 (SECRET).

5. Lincoln Laboratory, Lexington, Massachusetts, *A Preliminary Design of a TDMA System for FLEETSAT*, Technical Note 1975-5, J. D. Bridwell and I. Richer, 12 March 1975.

6. Lincoln Laboratory, Lexington, Massachusetts, *A Jammer Nulling Concept for FLEETSAT (U)*, Technical Note 1975-9, A. F. Culmone, 13 May 1975 (SECRET).

7. Computer Sciences Corporation, Falls Church, Virginia, *Systems Level Analysis for Demand-Assigned UHF Time-Division-Multiple-Access Satellite System*, CSC/TR-75-3012, Vol. 1, December 1975.

8. Joint Tactical Communications Office, Fort Monmouth, New Jersey, *System Specification for a Demand Assigned, UHF TDMA Satellite System*, TT-A2-2206-0029, 15 January 1975.

9. Comsat Laboratories, *Interim Report on Study of Functional Requirements for Demand Assigned SHF TDMA Modems*, DAAB-07-74-C-0204, May 1975.

APPENDIX C

A MULTI-USER-CLASS, BLOCKED-CALLS-CLEARED, BIRTH-DEATH MODEL WITH EXPONENTIAL ARRIVAL AND HOLDING TIMES

CONTENTS

APPENDIX C

A MULTI-USER-CLASS, BLOCKED-CALLS-CLEARED, BIRTH-DEATH MODEL WITH EXPONENTIAL ARRIVAL AND HOLDING TIMES

This appendix provides the mathematical detail for the results presented in Section III, as well as the computational technique used. Prior to the mathematical development, a discussion of related work in traffic theory is presented in Section A. Section B provides the mathematical development of the theoretical results. Section C presents a description of the computational technique used. Section C can be read independently of Section B, which can be bypassed by those readers not interested in mathematical detail.

A. PREVIOUS RELATED WORK

1. General Telephony

To date, the principal concern of teletraffic and congestion theory and of its application in traffic engineering has been dealing with terrestrial telephony systems, where the capacity per active circuit is always the same (e.g., one voice-band, 3-KHz circuit). Principal theoretical concern has been addressed to geographical distribution of access (switching) nodes and transmission lines of a network. This has focused development of the theory into complexities dealing with topological factors. Considerable effort has also been devoted to considering the more general call arrival and holding probability distributions, rather than the exponential ones. This has led to much more complex state-equation techniques (Refs. 1-6). Reference 3 is one of the more comprehensive texts dealing with teletraffic theory as it has evolved for the conventional telephony plant. References 2-4 require of their readers

an advanced mathematical background in the theory of stochastic processes, while Refs. 1 and 5 provide extremely useful texts at a less sophisticated mathematical level.

Only very recently, with the advent of data communication and, in particular, data circuits of different bandwidths or bit rates, has theoretical interest developed in traffic sources having different circuit capacity requirements. This new wrinkle is of basic theoretical significance. The above-mentioned texts do not address this factor.

2. Multi-Class Digital Traffic Sources

The methods used here were motivated by two recent papers (Refs. 7, 8). The problem as stated in Ref. 7 arose in considering two access methods for synchronous time-division multiplexing of different bit-rate local loops onto a wide-band, very-high-speed digital trunk. Consequently, the model considered only integer multiple values assigned to the c_i (e.g., $c_2 = 2$, $c_3 = 4$, $c_i = 2^i$). A further restriction was placed on the traffic source model in that the call initiation and call holding parameters, λ and μ, were assumed to be the same for all sources. This latter restriction seems very limiting and, as will be shown in the following sections, is unnecessary. However, Ref. 7 does provide the basic elements for the theoretical direction taken here.

Motivated by satellite communication problems, Ref. 8 models the problem as used here. The per-unit capacity c_i need not be integer multiples, nor need the traffic activity parameters all be the same, as in Ref. 7. Moreover, in Ref. 8 an important theoretical concept is provided but is not fully exploited. This is the concept of representing the capacity allocation access strategy as an allowable set of states A,

which imposes boundary conditions on the state equations.* The solution of these equations then determines the probability of occurrence for each allowable state. In Ref. 8, the set representation was used as a notational convenience for evaluating the strategy in which all traffic classes fully share satellite capacity. In point of fact, such a set representation has a fundamental significance beyond notational convenience, and this is explored in the following sections.

Motivated by these two references, a careful reexamination was undertaken relating the equations of state to capacity allocation strategies specified as general sets of allowable states. Within the theoretical context, the properties required by the capacity allocation strategy were determined in order for the commonly used recursive solution technique to be valid.

Unfortunately, Ref. 9** became available to this study only at its conclusion. This reference addresses itself to the time-division multiplexer problem studied in Ref. 7. Although it does restrict itself to c_i that are integer multipliers, it does not require all traffic classes to have identical λ, μ parameters, which is by far the most important restriction in Ref. 7. Had Ref. 9 been available earlier, considerably less time would have been spent on carefully formulating the equations of state.

Reference 9, like Refs. 7 and 8, does not formulate the problem with emphasis on the relationship between the equation of state and the set A of allowable states. Only the fully shared capacity allocation strategy is of interest. Reference

*A point of some theoretical significance is that in Ref. 8 the fundamental equations of state transitions are presented in collapsed form, leading directly to the recursive relations. This obscures examination of conditions for the validity of the solutions when the allowable set of states is varied.

**This reference is published in the *Proceedings of the Seventh International Teletraffic Congress* and to date has not been generally available other than through the attendees and their parent organization. A summary of the papers given at the Congress is contained in Ref. 10.

9 provides a detailed study of the blocking performance as determined by multiplexer implementation constraints. The effects investigated are those produced by restrictions on availability of time slots to incoming calls, time slot organization, and time slot fragmentation. Analogous implementation considerations arise in the design of TDMA modems (especially in digital buffer organizations) for satellite applications. Consequently, Ref. 9 represents the kind of further effort required to study more complex models.

Other examples of previous work of interest are Refs. 11 and 12. These both limit themselves to a mix of only two traffic classes (K = 2), wide band and narrow band. In Refs. 11 and 12 interest centers on queueing analysis where for one or both classes of traffic blocked call requests are held in queue, as opposed to being cleared. Their work shows examples of the rapid complexification of the state equations in Blocked-Calls-Held operation. Reference 13 addresses a multi-class system with all $c_i \equiv 1$ and a very simple partitioned-queue protocol and access strategy. The problem is also computationally complex.

3. Computer Science

It is useful to point out the growing overlap between teletraffic theory and the theory of queueing networks as it is being currently advanced for modeling and evaluating advanced multiprocessor architectural configurations and teleprocessing network configurations. The purpose of this subsection is to exhibit the nature of the common elements of the theories and to advocate the utility of cross-fertilization between the computer-science and traffic-theory communities.

The components of an ADP complex, both real (e.g., input/output channels, memories, arithmethic units) and virtual (e.g., software processes such as a matrix inversion), can be modeled as nodes of an interconnected graph (and hence a network). Application programs enter the graph at a node representing the

input port or channel of the computer and proceed through the graph as the computational tasks are executed. On completion of their journey, they exit from a node representing the output port or channel of the computer. It is assumed that entering computational tasks required by user programs* can be broken down into fundamental processing units that move probabilistically ("flow stochastically") between nodes according to the nodal interconnections. The units queue at each node while waiting the computational service provided by that particular node. The computational capacity of each node i is modeled as a group of "m_i servers" where one server processes one computational unit. Each node, then, is an m_i-server queueing facility, and the overall system is a network of queues.

Such a model shows much promise of providing a theoretical basis for study of the sizing and interconnections of the overall processing system, i.e., system architecture in terms of job load and desired throughput. That there is overlap with teletraffic theory should not be surprising, for the model conceptually resembles a packetized store-and-forward message network such as ARPANET (Refs. 14, 15).

Now, to demonstrate the overlap between this computer science activity and teletraffic theory, consider grading theory (see, for example, Ref. 16, and if possible, Ref. 17). The objective is to match a limited trunk availability to local lines in such a way that blocking is minimized for a given level of trunk availability. A grading is a specification of selector interconnections of local lines to trunks that can be represented as an "incidence matrix" between local line and trunk circuits. Such incidence matrices have model graphs that are of a particular form (Ref. 17) and are a subclass of general

*The model can also include parts of the operating system software not associated with user-submitted programs. Operating system tasks are modeled to circulate continuously without leaving the graph.

graphs in the computer processing models. For example, in a grading there can be no "feedback" paths as occur in the more general network graph. Consequently, since a grading is a subclass of network graphs, any general results applicable to networks also apply to gradings.

One of the more important early results in the theory of networks of queues is known as Jackson's theorem (Ref. 18). It provides the solution for the steady-state probability of queue lengths at each of the network nodes. This result underlies the common mathematical theory for teletraffic and processor performance. A paper by Gordon and Newell (Ref. 19) extends Jackson's work and interrelates open (no permanently circulating computational units) and closed (only permanently circulating quanta) queueing networks. Apparently, neither of these papers was motivated by the computer architecture model.* One of the earliest papers to recognize the applicability of Jackson's formulation to computer architecture was by Kleinrock (Ref. 20). Further applications were evolved in Refs. 21 and 22.

In any application of significant scope, the computational complexity involved in evaluating the mathematical formulae can become prohibitive. Efforts are being made to reduce or avoid the computational burden. Examples representative of the approaches taken are given below:

1. Computational algorithms (Refs. 21-23)
2. Network equivalence** (Refs. 24, 25)
3. Combination of 1 and 2 (Ref. 26)
4. Limiting differential models or diffusion theory (Ref. 27).

*It is of historical interest to note that the stated motivation for this reference was from classical operations research problems relating to organizing efficient repair and maintenance of factory machines. The relationship to computer science and teletraffic theory was not perceived at that time.

**Analogous to Thevenin and Norton's equivalent-source impedance concept of a resistive electrical network.

With the exception of Ref. 27, all of the above references deal with a model in which one computational unit is served by one nodal server, the fundamental assumption being that computational tasks can be subdivided so as to make this match at all of the nodes in the graph.

Even though the single-node capacity allocation problem considered in this study is topologically trivial, it differs in one very important and fundamental aspect from the theory currently being developed in the computer-science context. In particular, a traffic source requesting a circuit of capacity cannot be subdivided into fundamental units (c individual circuits of capacity 1) each of which can be served in sequence. This currently limits the applicability of the computer-science-generated theory.

Practical computer constraints may not allow all user-application tasks, let alone operating-system activity, to be divisible into common fundamental units which match all possible nodal server units for nodal processing. Consequently, computer scientists should be interested* in extending their theory to more general cases in which one task takes more than one nodal server. The results from such an extension would have direct application to teletraffic theory.

B. STATEMENT OF MATHEMATICAL MODEL; DEFINITIONS

1. System Definition

Traffic Model:

There are K classes of system users who place and terminate calls randomly and independently with exponential interevent

*Following preparation of this report, Ref. 28 was discovered. This reference considerably generalizes the queueing network model and explores conditions for the "product form solution" to hold. Thus, the objective of this appendix parallels that of Ref. 28 in a specialized way. Most significantly, Ref. 28 suggests that Blocked Calls *Cleared* will not have product form in general.

times. If at some moment there is inadequate capacity available to support a new call request because of a larger number of calls in progress, the new call is not held in queue until capacity becomes available (Blocked Calls Cleared).

For each $i = 1, 2, \ldots K$, define user class i with parameters $(N_i, \lambda_i, \mu_i, c_i)$.

a. Population size = N_i, either finite or infinite

b. Differential call arrival probability

$= (N_i - j_i)\, \lambda_i dt$ when $0 \leq j_i \leq N_i < \infty$

$= \tilde{\lambda}_i dt$ when $N_i = \infty$, i.e., $\lambda_i \to 0$, $N_i \to \infty \ni$: $N_i \lambda_i \to \tilde{\lambda}_i$

c. Differential call holding probability = $\mu_i dt$

d. System state vector $j \equiv (j_1, j_2, \ldots j_K)$, where j_i is the number of calls in progress from each user class i

Capacity Model:

e. A call in progress uses capacity c_i (i.e., c_i "servers" needed to support call from class i)

f. Total capacity available is C_o

g. The capacity in use $C(j)$ is the linear sum of capacity used by each call in progress, i.e., $C(j) \equiv \sum_{i=1}^{k} c_i j_i \leq C_o$.

The problem is to determine the equations of state and solve for the stationary probability $P(j)$ that there are $j_1, j_2, \ldots j_K$ calls of each class in progress. From $P(j)$, the probability of finding j in any specified set of states B (such as blocking states) is given by $P(B) = \sum_{j \in B} P(j)$. In addition, the average of any function of j, $f(j)$, is given by $\sum_{\text{all } j} f(j)\, P(j)$.

From items a and g above, the admissible states j of calls in progress must be contained in the set Ω, composed of non-negative integer-valued points in the intersection of a rectangle resting on the positive coordinate axes with vertex at the

origin, sides of length N_i, and integer points lying below the plane defined by $(c,j) \triangleq \sum_{i=1}^{K} c_i j_i = C_o$:

$$\Omega = \left\{ j \mid (c,j) \leq C_o \right\} \cap \prod_{i=1}^{K} \left\{ 0 \leq j_i \leq N_i \right\}.$$

Define

h. Allowable set of system states $\equiv A \subseteq \Omega$.

The fundamental description of the "system" in the model used here is the specification of the user-class traffic parameters, the set A, and the probability distribution, $P(j)$, $j \in A$. Note that $P(A)$ always must equal one, and, since blocked calls are cleared, $P(j) = 0$ for j not in A. The sets A serve to define access "arrangements" afforded to the K different user classes. All calls are provided "demand access" but need not have access to all of the capacity C_o. The A-set is a means of apportioning capacity availability to the different user classes.

Examples of A-Sets (Fig. C-1):

1. $A = \Omega$

   Fully shared capacity. All calls are treated equally, with access to any user class of any available capacity in system.

2. $A = \prod_{i=1}^{K} \left\{ j_i \mid c_i j_i \leq C_i \right\}$, where $\Sigma C_i = C_o$

   Capacity dedicated to each user class. Capacity is apportioned to each user class, with demand access only within that capacity apportioned to each class.

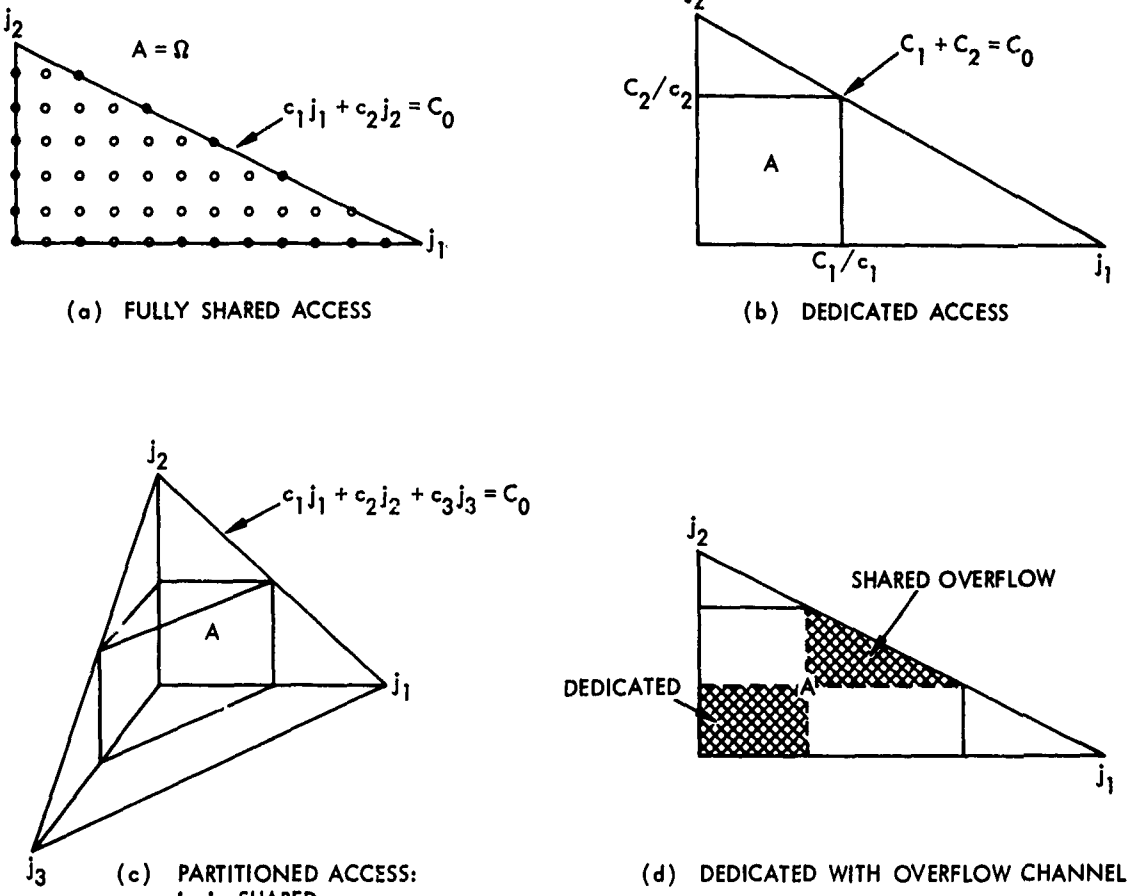


FIGURE C-1. Access Strategies and A-Sets

3. $A = \left\{ j \mid \sum_{i=1}^{K} c_i j_i \leq C_o - \sum_{i=M+1}^{K} C_i \right\} \times \prod_{i=M+1}^{K} \left\{ j_i \mid c_i j_i \leq C_i \right\}$

Partitioned access. The first classes have fully shared access per item 1 above; the last K - M classes have fully dedicated access per item 2 above.

4. $A = \Omega \times \prod_{i=M+1}^{K} \left\{ j_i \mid c_i j_i \leq C_i \right\}$, where $\sum_{i=M}^{K} C_i < C_o$

Capacity-limited access. Fully shared access to first M - 1 classes, with last K - M + 1 user classes capacity limited.

5. Imbedded capacity. Choose a joint j^* with coordinates j_1^*, j_2^*, etc., interior to Ω. Inscribe a right rectangle resting on the positive coordinate axes and with vertices at the origin and j^*. Then project forward each face of this rectangle not contained in some coordinate plane until it touches the plane $\Sigma c_i j_i = C_o$. Complete the rectangle that the projected faces determine. Let the set A be the intersection of Ω with the new rectangle. This A-set provides a level of dedicated capacity to each user class defined by the point j^* that uses up $\Sigma c_i j_i^*$ capacity units. The residual capacity $C_o - \Sigma c_i j_i^*$ is then fully shared for "overflow" traffic.

For the examples, it can be seen that the system architecture for capacity sharing is expressed by the specification of the admissible set of states A. It should be noted that there are limitations in specifying A-sets. Not only must they lie within Ω, but, because of the Birth-Death model, any j state in A must be reachable by passing through other j states within A (i.e., A must be connected). In every instance, an A-set is a geometrical solid in K dimensions resting on the positive coordinate axes.

A performance objective can be specified by determining blocking sets B that will be contained in A [e.g., in example 1, a blocking set could be $B = \{j \mid (c,j) = C_o\}$] and by computing P(B). The blocking set B is those states of A for which no new calls can be accepted until a call is terminated. With the user characteristics (N, λ, μ, c), the object is to study and compare alternative architectures A by evaluating P(B).

For K = 2, Fig. C-1 depicts the sets A. In Fig. C-1, example 3 is shown in three dimensions, since it becomes congruent to example 2 for K = 2. All the figures have an extremely important property: the orthogonal projections to the coordinate axis from any point in the set A are wholly contained in the set A. To coin a phrase, any figure possessing this last property can be said to have coordinate convexity. Note that in example 5, although A has coordinate convexity, it is not a convex* set. (The other four examples are convex.) A property equivalent to coordinate convexity is that the exterior surfaces of A be composed entirely of planar facets parallel to some coordinate plane. The lower exterior surfaces are contained in the coordinate planes (i.e., the planes formed by adjacent pairs of coordinate axes).

The property of coordinate convexity in A is important to the development of the geometrically simple** solutions to the equations of state that determine P(j). As will be shown below,

*A convex body is one in which a line drawn between any two points of the body is wholly contained in the body.

**The reader is warned that, although geometrically simple, the numerical computation can be very significant. The situation will be analogous to a finite system of linear constant coefficient differential equations wherein the geometrically simple solution is the diagonalization of the "characteristic matrix." Of course, all of filter theory and network synthesis follows from such a "simple" geometry.

these K-dimensional difference equations will separate variables and possess recursive solutions for coordinate-convex A. This means that $P(j) = \prod_{i=1}^{K} P_i(j_i)$ provided $j \in A$, and each $P_i(j_i)$ will solve its own difference equation independently of any other i provided the solutions are restricted to $j \in A$. Even further, the solutions will have the following property. Although the *coefficients* of the K-dimensional difference operator *change* when j is on the boundary of A and vanish when $j \notin A$, the solution is the same as if one were to ignore the restricted state space A and assume that the difference operator applied everywhere with no change in coefficients. Then, take the $P(j)$ so solved, arbitrarily set $P(j) = 0$ for $j \notin A$, and renormalize so that $P(A) = 1$. The above property is of theoretical significance in that it *completely separates* the solution technique for the difference equations from *the system architecture embodied in the set A.*

2. Equations of State

From the properties of the Birth-Death process (i.e., exponential interarrival time, exponential holding time, and state transitions only to adjacent neighbors), the steady-state probability $P(j)$ for occupying state $j \in A$ must satisfy a difference equation,* boundary conditions, and a normalization condition. (For a discussion of state equations for pseudo-random sources and infinite populations, see, for example, Refs. 1-6.)

For any point j in the strict interior of A, (A'), (i.e., adding or subtracting one to any j coordinate produces a new j' which is either in A or on its boundary), the homogeneous difference equation that $P(j)$ must satisfy is given by

$$+[(\mu,j)+(\lambda,N-j)]P(j)-(N+I-j,\ \lambda\Delta^{-1}P(j))-(j+I,\ \mu\Delta^{+1}P(j)) = 0 \quad (C\text{-}1)$$

for all $j \in A'$,

*These equations are obtained by equating probability flow into and out of state j.

where for brevity the following vector notation has been used:

$$j \triangleq (j_1, j_2, \cdots j_K)$$

$$\mu \triangleq (\mu_1, \mu_2, \cdots \mu_K)$$

$$\lambda \triangleq (\lambda_1, \lambda_2, \cdots \lambda_K)$$

$$N \triangleq (N_1, N_2, \cdots N_K)$$

$$I \triangleq (1, 1, \cdots 1)$$

$$(a,b) \triangleq \sum_{i=1}^{K} a_i b_i$$

$P(j) \triangleq$ scalar probability function

$j \pm \delta_1 \triangleq (j_1, j_2, \cdots j_i \pm 1, j_{i+1}, \cdots j_K)$

$\lambda\Delta^{-1} P(j) \triangleq$ backward difference vector whose i^{th} components are

$$\left\{\lambda\Delta^{-1} P(j)\right\}_i = \lambda_i P(j-\delta_i), \quad i = 1, 2, \cdots K$$

$\mu\Delta^{+1} P(j) \triangleq$ forward difference vector whose i^{th} components are

$$\left\{\mu\Delta^{+1} P(j)\right\}_i = \mu_i P(j+\delta_i)$$

The equations of state must also be specified on the boundaries of A, (A - A´). With coordinate convexity, the boundaries of A can be separated into K planar lower boundaries $\underline{A}(i)$, i = 1, 2, ... K, formed by the positive coordinate axes and an upper boundary $\overline{A}$. Each planar lower boundary $\underline{A}(i)$ is composed of j points with nonzero-valued coordinates only at j_i and j_i+1. In the case of $\underline{A}(K)$, the nonzero coordinates would be j_K and j_1. Thus, in what follows, when the notation reads i+1 and i is set equal to K, i + 1 should be read equal to 1 (e.g., j_{i+1} at i = K is j_1).

Equation C-1 must be modified to account for the boundary conditions. Since a "Birth" (new call) cannot occur from a disallowed state below the lower boundary $\underline{A}(i)$, there can be no backward difference* in the K - 2 directions normal to $\underline{A}(i)$ (i.e., the zero-valued j_i coordinates). In addition, the corresponding terms in the sum (μ, j) on the left-hand side of Eq. C-1 must also vanish, as they correspond to a Death to a state below $\underline{A}(i)$. This is accounted for automatically because the directions normal to $\underline{A}(i)$ are those coordinates with zero-valued j_ℓ $(\ell \neq i, i+1)$, so $\mu_\ell j_\ell = 0$.

Similarly, when j belongs to the upper boundary $\overline{A}$ of A, it is not possible to "die" from or be "born" into a disallowed state above the boundary. Thus, for $j \in \overline{A}$, Eq. C-1 must be modified by setting to zero the K - 2 vector coordinate components normal to $\overline{A}$ at j of the forward difference vector $\mu\Delta^{+1}P(j)$ on the right-hand side of Eq. C-1 and the corresponding λ vector coordinates on the left-hand side of Eq. C-1.

Thus, with the normalization condition that P(A) = 1, the homogeneous system of state difference equation in its briefest form is given by Eq. C-1 with boundary conditions determined by

$$P(A) = \sum_{j \in A} P(j) = 1 \quad .$$

If $j \in \underline{A}(i)$, that is to say, $j \in A$ but $j - \delta_i \in A$, then**

$$\left\{\lambda\Delta^{-1} P(j)\right\}_i = 0$$

---

*Backward differences are associated with "Birth" from a state below the given one, while forward differences are associated with "Death" from a state above the given one.

**This condition applies to those j states on a coordinate axis, $j \equiv (0, 0, j_i, 0, 0)$ and $j \equiv 0$. The coordinate axis i is the intersection of $\underline{A}(i-1)$ and $\underline{A}(i)$, while $j \equiv 0$ is the intersection of all K, lower boundary planes $\underline{A}(i)$. One must set the indicated terms to zero simultaneously for these states.

If $j \in \bar{A}$ (i.e., for some i, $j+\delta_i \notin A$), let $V(j)$ be the set of those indices of the j vector for which $j+\delta$ is outside A. Then, for $j \in \bar{A}$ and $i \in V(j)$,

$$\{\mu\Delta^{+1} P(j)\}_i = 0, \text{ if } i \in V(j)$$

$$(\lambda, N-j) = \sum_{i \in V(j)} \lambda_i (N_i - j_i) \quad .$$

For those classes which have infinite source population, the λ inner product terms must be modified in the following way. Each inner product in Eq. C-1 becomes two inner products--one for the infinite N_i population coordinates and one for the finite N_i population coordinates, thus ordering the finite population classes first, where it is assumed that there are $M \leq K$ finite population classes:

$$\begin{aligned}
(\lambda, N - j) &\to (\lambda, N - j) + (\tilde{\lambda}, I) \\
&= \sum_{i=1}^{M} \lambda_1 (N_i - j_i) + \sum_{i=M+1}^{K} \tilde{\lambda}_i \\
(N + I - j, \lambda\Delta^{-1} P(j)) &\to (N + I - j, \lambda\Delta^{-1} P(j)) \\
&\quad + (\tilde{\lambda}, \Delta^{-1} P(j)) \\
&= \sum_{i=1}^{M} (N_i + 1 - j_i) \lambda_i P(j - \delta_i) \\
&\quad + \sum_{i=M+1}^{K} \tilde{\lambda}_i P(j - \delta_i)
\end{aligned} \tag{C-2}$$

Equations C-1 and C-2 are second-order* K-dimensional homogenous difference equations over the set of allowable states A.

*There is a two-step difference between Δ^{-1} and Δ^{+1}, i.e., $P(j - \delta_i)$ and $P(j + \delta_i)$.

The fact that A is lower bounded by the planes generated by the coordinate axes and has the coordinate-convexity property will allow Eqs. C-1 and C-2 to be written as K "independent" one-dimensional second-order difference equations.

3. Separating Solution

It is assumed that all states j in A have a nonzero probability of occuring; $P(j) > 0$ for all $j \in A$. Furthermore, assume that $P(j)$ can be factored into the product of K functions $P_i(j_i)$:

$$P(j) = \prod_{i=1}^{K} P_i(j_i). \tag{C-3}$$

The object of using the assumed separated product form of Eq. C-3 in Eq. C-1 or C-2 is to produce a set of K second-order homogeneous difference equations, one for each j_i coordinate, in order to replace the K-dimensional Eq. C-1. Anticipating the result of substituting Eq. C-3 into Eq. C-1, define the second-order difference operator D_i on any function, F(s), on positive integers s = 1, 2, ...

$$\begin{aligned} D_i\,[F(s)] \triangleq\ & (\mu_i s + (N_i - s)\lambda_i)\, F(s) \\ & - (N_i + 1 - s)\lambda_i\, F(s-1) - (s+1)\mu_i\, F(s+1)\ , \end{aligned} \tag{C-4}$$

where μ_i, λ_i, N_i are the traffic source parameters for user class i (and hence the subscript on D).

Now, for any j vector on the strict interior of A, $A' \equiv \{j \in A \mid j \pm \delta_i \in A \text{ for all } i = 1, 2, \ldots K\}$ substituting Eq. C-3 into Eq. C-1, dividing* by $P(j) > 0$, and using the definition of D_i will produce

---

*This allows one to capitalize on the fact that $\prod_{\ell \neq i} P_\ell(j_\ell) / P(j) = P_i^{-1}(j_i)$.

$$\sum_{i=1}^{K} P_i^{-1}(j_i) \cdot D_i[P_i(j_i)] = 0; \text{ for } j \in A' \qquad \text{(C-5)}$$

Remembering that for $j \in \underline{A}(i)$, $j_\ell = 0$ for $\ell \neq i + 1$ (and $K + 1$ is equal to 1), and that the backward differences are zero at $\ell = i$ and $i + 1$, substitution of Eq. C-3 into Eq. C-1 and division by $P(j)$ yields for each $i = 1, 2, 3, \ldots K$:

$$\sum_{\ell=i}^{i+1} P_i^{-1}(j_i)\, D_i[P_i(j_i)] = \sum_{\substack{\ell=1 \\ \ell \neq i, i+1}}^{K} P_\ell^{-1}(0)(N_\ell \lambda_\ell P_\ell(0) - \mu_\ell P_\ell(1)) \qquad \text{(C-6)}$$

$$\text{for } j \in \underline{A}(i),\ j_i \geq 1,\ j_{i+1} \geq 1.$$

Equation C-6 holds when $j \in \underline{A}(i)$ and both j_i and j_{i+1} are at least as large as one.

The boundary conditions on a coordinate axis are obtained next. The i^{th} coordinate axis, $I(i)$, is the intersection of $\underline{A}(i - 1)$ and $\underline{A}(i)$, where $i - 1$ evaluated at $i = 1$ must be interpreted as equal to K. Thus, when only $j_i \geq 0$ and all other $j_\ell = 0$, $\ell \neq i$, the $i+1^{st}$ backward difference in Eq. C-5 vanishes, and Eq. C-6 is modified to read:

$$P_i^{-1}(j_i)\, D_i[P_i(j_i)] = \sum_{\substack{\ell=1 \\ \ell \neq i}}^{K} P_\ell^{-1}(0)(N_\ell \lambda_\ell P_\ell(0) - \mu_\ell P_\ell(1)) \qquad \text{(C-7)}$$

$$\text{for } j \in I(i),\ j_i \geq 1$$

On the lower boundary this leaves only one point unspecified, $j \equiv 0$, or the intersection of all $I(i)$. Since at $j \equiv 0$ all backward differences vanish, the left-hand side of Eq. C-7 vanishes and leaves, for $j \equiv 0$

$$\sum_{i=1}^{K} P_i^{-1}(0)\left[N_i\lambda_i P_i(0) - \mu_i P_i(1)\right] = 0 \quad . \tag{C-8}$$

Equations C-6 through C-8 provide the lower boundary conditions on Eq. C-5. It should be noted that separation of the K-dimensional difference Eq. C-1 has been only partially accomplished, in that sums of separated difference operators equate to zero. After developing the upper boundary condition, it will be shown that the solution can be found by equating to zero each separated difference operator of any sum.

The upper boundary conditions on Eq. C-5 are not as easily stated as the lower ones due to the freedom in having more complex upper boundary surfaces to A. If $j \in \bar{A}$, there exists at least one i such that $j+\delta_i$ does not belong to A. Recall that $V(j)$ are those indices of $j \in \bar{A}$ for which $j+\delta_i \notin A$. Thus, when $j \in \bar{A}$, the cardinality of $V(j)$, i.e., the number of elements in the set $V(j)$, must lie between 1 and K. Remember, by coordinate convexity, for every $j \in A$ (including $j \in \bar{A}$) and for any i, $j-\delta_i \in A$ unless $j_i = 0$.

Now, for $j \in \bar{A}$, both the forward difference as well as the $(\lambda, N-j)$ components at those coordinate indices contained in $V(j)$ go to zero. Then, substituting Eq. C-3 into Eq. C-1 for $j \in \bar{A}$ and dividing by $P(j)$ yields for $j \in \bar{A}$

$$\sum_{i \notin V(j)} P_i^{-1}(j_i) D_i[P_i(j_i)] =$$

$$\sum_{i \in V(j)} P_i^{-1}(j_i)[\mu_i j_i P_i(j_i) - (N_i+1-j_i)\lambda_i P_i(j_i-1)] \tag{C-9}$$

Note that the right-hand side of Eq. C-9 is a first-order difference operator on indices $i \in V(j)$, while the usual second-order operator D_i is in effect on those $i \notin V(j)$.

Thus, Eqs. C-5 through C-9 (with D_i defined by Eq. C-4), together with the normalization

$$\sum_{j \in A} \prod_{i=1}^{K} P_i(j_i) = 1 , \qquad \text{(C-10)}$$

provide an equivalent set of difference equations for a separable solution to Eq. C-1. (For those classes with N_i infinite, replace all terms containing $N_i\lambda_i$ with $\tilde{\lambda}_i$ and set to zero $\lambda_i j_i$.) Because of coordinate convexity in A, Eqs. C-5 through C-9 can be solved recursively. One starts at $j \equiv 0$ and works upward to A until $\overline{A}$ is reached.

Starting with $j \equiv 0$ and Eq. C-8, one would like to set each bracketed component of the sum in the left-hand side of the equation to zero. One might reason that the $P_i(0)$ are like arbitrary constants of integration in the difference equations and depend only on the probability normalization, which is not explicit to Eq. C-8. If the $P_i^{-1}(0)$ could be arbitrarily varied in Eq. C-8, each i bracketed component of the summand would have to be individually zero. Another rationale for setting each i bracketed component to zero is that otherwise one would have a nonlinear difference equation deriving from the original linear Eq. C-1. A fundamental property of the Birth-Death Markov process, stated in Ref. 29, p. 408, is that a solution to the state equations having the property P(A) = 1 (i.e., Eq. C-10) is *unique**. Consequently, if, by whatever method, a $P(j)$ is found to solve Eqs. C-5 through C-10 such that P(A) = 1, it solves Eq. C-2 and is unique.

Setting each i summand of Eq. C-8 to zero and defining $a_i \triangleq \lambda_i/\mu_i$ produces

---

*In fact, Eq. C-1 is the steady-state ($t \to \infty$) version of the Kolmogorov-Chapman equations. It is the solution of these more general time-dependent equations that is unique. Unique solutions relate to ergodic properties of Markov chains and determine the existence of the steady state. If the steady state exists and A has only a finite number of states, as is the case for finite-capacity systems, clearly P(A) can be normalized to one.

$$P_i(1) = N_i a_i \; P_i(0) \qquad \text{(C-11)}$$

This imediately implies that the right-hand sides of Eqs. C-5 through C-7 are zero. Consequently, from Eq. C-7, for each i and $j \in A - \bar{A}$, $j_i \geq 1$.

$$D_i\Big[P_i(j_i)\Big] = 0 \; . \qquad \text{(C-12)}$$

Note that Eq. C-7 equated to zero also forces each i^{th} bracketed component of the left-hand side of Eqs. C-5 and C-6 to zero. This, in turn, produces Eq. C-12. Thus, Eq. C-12 solves Eqs. C-5 through C-7. Dividing Eq. C-12 through by μ_i yields

$$[j_i + (N_i - j_i)a_i] \; P_i(j_i) = (N_i = 1 - j_i)a_i \; P_i(j_i - 1) + (j_i + 1) \; P_i(j_i + 1) \qquad \text{(C-12')}$$

It can be verified by direct substitution that for each i = 1, 2, ... K

$$P_i(j_i) = \binom{N_i}{j_i} a_i^{j_i} \; P_i(0), \qquad \text{(C-13)}$$

where

$$\binom{N_i}{j_i} = \frac{N_i!}{(N_i - j_i)! j_i!}$$

satisfies Eq. C-12' and hence Eq. C-12. Thus, Eq. C-13 in conjunction with Eq. C-3 is a solution to the state equations at all of the lower boundaries and interior j states of A. If it

can be demonstrated that Eq. C-13 satisfies the state equation (Eq. C-9) on the upper boundary of A ($\bar{A}$), then the solution will have been found.

Let j be any state of the boundary of $\bar{A}$. Then surely Eq. C-13 guarantees that for those $i \in V(j)$, $D_i\,[P_i(j_i)] = 0$, and hence the left-hand side of Eq. C-9 is zero. Consequently, if for each $i \in V(j)$ the corresponding summand of the right-hand side of Eq. C-9 is zero, Eq. C-13 will also satisfy Eq. C-9. To show this, substitute Eq. C-13 into the right-hand summand of Eq. C-9 and divide by μ_i, to obtain

$$j_i P_i(j_i) - (N_i+1-j_i)\, a_i P_i(j_i-1) =$$

$$\left[j_i \binom{N_i}{j_i} a_i^{j_i} - (N_i+1-j_i) \binom{N_i}{j_i-1} a_i a_i^{j_i-1} \right] P_i(0) \equiv 0.$$

Thus, whether or not $i \in V(j)$, Eq. C-13 forces the right-hand side of Eq. C-9 to zero, and the solution for $P(j)$ is given by combining Eqs. C-3 and C-13:

$$P(j) = P_A(0) \cdot \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i}$$

$$P_A(0) \equiv \left(\prod_{i=1}^{K} P_i(0) \right) \qquad \text{(C-14)}$$

$$j = (j_1, j_2, \ldots j_K)$$

$$a_i = \lambda_i/\mu_i \,,$$

where the as yet undetermined product of $P_i(0)$ is denoted as $P_A(0)$. The subscript A indicates that this constant depends on the set A. Note that the rest of Eq. C-14 *does not* depend on A. The *remaining* K-fold product on the j_i coordinates of the

state j depends only on the traffic characteristics of the user classes and is wholly independent of A. The impact of various access strategies (choices of A-sets and available capacity) are mathematically manifested only in $P_A(0)$, a crucial number. $P_A(0)$ is determined by the normalizing condition $P(A) = 1$. Summing* Eq. C-14 over all $j \in A$ and setting the result equal to one produces the separated solution:

$$P(j) = P_A(0) \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i}$$

$$P_A(0) = \left(\sum_{j \in A} \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i} \right)^{-1} \qquad \text{(C-15)}$$

The coordinate-convexity property in the A-set allowed one to start with Eq. C-8, $j \equiv 0$, and then recursively proceed back to Eq. C-7 and, in turn, to Eqs. C-6 and C-5, obtaining the solution form of Eq. C-13. Consequently, coordinate convexity is a sufficient condition for Eq. C-15 to solve Eq. C-11 and, as will be seen in the next subsection, Eq. C-2. Moreover, coordinate convexity considerably simplifies the topology of the upper boundary $\overline{A}$.

Whether coordinate convexity is a necessary condition is an open question. That is to say, is there at least one A-set not coordinate convex for which Eq. C-15 fails, or does Eq. C-15 solve Eq. C-1 for all A-sets? As will be shown in the next subsection, this question is physically uninteresting.

4. Coordinate Convexity and Infinite Population Classes

Because A was postulated to have coordinate convexity, Eqs. C-5 through C-10 could be solved recursively, producing Eq. C-15. Recall that for an A-set not coordinate convex there

*For even modest values of K (> 3) and N_i, this summation is computationally nontrivial.

exists at least one j state, j' in A with at least one coordinate i' such that $j_{i'} \geq 1$ and the state $j' - \delta_{i'}$ is no longer in A. This implies that the termination of a call in progress from an i' user class cannot take place until some other event (call arrival or departure from other than an i' user) transpires. Physically, the capacity relinquished by a departing call would blocked for reassignment to a new call arrival until the system transitions to a suitable j state. Coordinate convex A-sets guarantee that call completion from all j states returns the newly freed capacity for use by new call arrivals.

Consider the simple example in Fig. C-2 for an A-set in two dimensions that does not have the coordinate-convexity property. There exists a portion of the lower boundary of A that lies above the j_1 axis. For any state j on this lower boundary of A, a blocked-service release (as opposed to blocked-service request) will occur when a class 2 call terminates. Note that although the departing user capacity is blocked from returning to the capacity pool, additional users can be serviced. If a user of class 1 departs, the system moves away from j, in which case both capacities are returned for further use.

A-sets lacking coordinate convexity are certainly pathological in communications systems. For application to the sharing of communications capacity, it is reasonable to allow only A-set strategies that possess coordinate convexity. In all cases studied here, this restriction is presumed.

With the separation of variables in Eqs. C-5 through C-10, those user classes with infinite population can be treated by appropriately modifying the difference operators D_i and boundary conditions at those i indices with $N_i = \infty$.

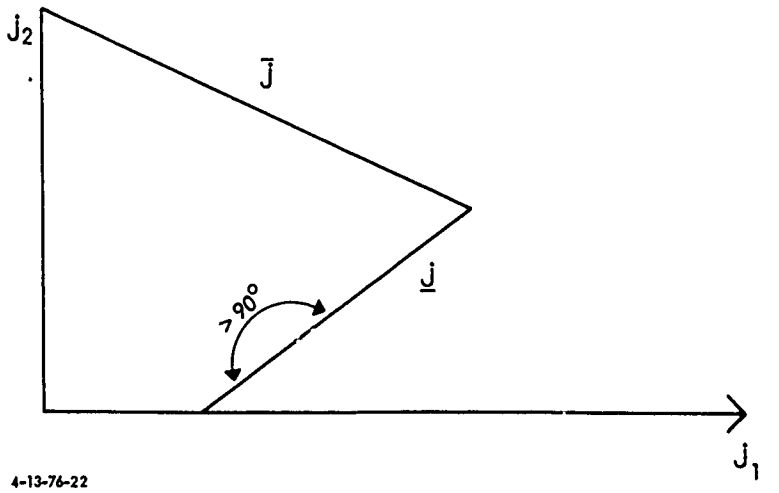


FIGURE C-2. An Example of Blocked Call Release

By direct substitution, it can be verified that for those user classes with $N_i = \infty$

$$P_i(j_i) = P_i(0) \frac{1}{(j_i)!} (\tilde{a}_i)^{j_i} \qquad \text{(C-16)}$$

$$\tilde{a}_i \triangleq \tilde{\lambda}_i/\mu_i = \lim_{\substack{N_i \to \infty \\ \lambda_i \to 0}} N_i\lambda_i/\mu_i \ .$$

Equation C-16 can also be derived from Eq. C-15 by taking the limit $N\lambda \to \tilde{\lambda}$ as $N \to \infty$ (or equivalently $Na \to \tilde{a}$). Thus, for all $j \in A$, Eq. C-15 generalizes to

$$P(j) = P_A(0)\left[\prod_{i \in K-\tilde{I}} \binom{N_i}{j_i} a_i^{j_i}\right]\left[\prod_{i \in \tilde{I}} \frac{1}{(j_i)!} \tilde{a}_i^{j_i}\right]$$

$$P_A(0) = \left\{\sum_{j \in A}\left[\prod_{i \in K-\tilde{I}} \binom{N_i}{j_i} a_i^{j_i}\right]\left[\prod_{i \in \tilde{I}} \frac{1}{j!} \tilde{a}_i^{j_i}\right]\right\}^{-1} \qquad \text{(C-17)}$$

$$\tilde{I} = \{i | N_i = \infty\}$$

$$K-\tilde{I} = \{i | N_i < \infty\}$$

$$a_i = \lambda_i/\mu_i, \quad i \in K-\tilde{I}$$

$$\tilde{a}_i = \tilde{\lambda}_i/\mu_i, \quad i \in \tilde{I} .$$

To reemphasize, in the Birth-Death model used here for Blocked Calls Cleared, the only portion of $P(j)$ that exhibits dependency on the set of admissible states A is the normalization constant $P_A(0)$. The product portion is completely determined by the charactieristics of the user class, i.e., N_i, λ_i, μ_i, ($a_i = \lambda_i/\mu_i$). The parameters N_i, c_i, and C_o establish the constant set Ω within which A must be contained. Thus, in a formal sense, *choice of system architecture A is primarily manifested in the calculation of* $P_A(0)$.

5. Statistical Dependence

A point of considerable importance is that, apart from a constant factor $P_A(0)$, the form of Eqs. C-15 and C-17 is a K-fold product of terms in the coordinates of the state $j = (j_1, j_2, \ldots j_K)$. This product form might indicate that the user classes are independent. In a strict probabilistic sense, the user classes are statistically dependent, although the form of the state probability has the computational convenience as if the user classes were independent. To see this, consider only the finite population case, and note that under the assumption that all of the K user classes were mutually independent* it would have to follow that for all $j \in A$

$$P(j) = \prod_{i=1}^{K} Q_i(j_i), \qquad \text{(C-18)}$$

where $Q_i(k) \triangleq$ probability k users are active in class i.

*More generally, one could have some of the user classes independent, for which $P(j) = \left\{ \prod_{i=1}^{L} Q_i(j_i) \right\} \cdot Q(j_{L+1}, \ldots j_K)$, assuming the first L classes as independent.

For $k = 0, 1, \ldots \hat{j}_i = \max_{j \in A} j_i$, then from first principles of probability,

$$Q_i(k) = \text{Prob}\left\{j_i = k\right\} = P(\sigma_i(k))$$

$$= \sum_{j \in \sigma_i(k)} P(j) \qquad \text{(C-19)}$$

$$Q_i(k) = P_A(0)\binom{N_i}{k} a_i^k \, p_{\sigma_i}(k) \,,$$

where

$$p_{\sigma_i}(k) \triangleq \sum_{j \in \sigma_i(k)} \prod_{\ell \neq i} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell}$$

and the set $\sigma_i(k)$ is the collection of all j states in A whose i^{th} coordinate is equal to k. Note that for any $i = 1, 2, \ldots K$, as k varies from 0 to its largest value $\hat{j}$, the $\sigma_i(k)$ are disjoint and $A = \bigcup_{k=0}^{\hat{j}} \sigma_i(k)$.

Introducing Eq. C-19 into C-18, equating to Eq. C-15 and then dividing by $P(j)$ produces the following condition for statistical independence:

$$\prod_{i=1}^{K} \sum_{j \in \sigma_i(j_i)} \prod_{\ell \neq i} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell} = (1/P_A(0))^{K-1} \,. \qquad \text{(C-20)}$$

The left-hand side of Eq. C-20 formally depends on j, while the right-hand side is a constant independent of j. This, in turn, implies that the sets $\sigma_i(k)$ for each $i = 1, 2, \ldots K$ would have to be invariant to the k value. The only A-set that will satisfy this condition (Eq. C-20) is the A-set for dedicated

capacity. This physically separates the available capacity into user-dedicated pools, which are obviously independent.

To demonstrate mathematically, let $\hat{\jmath}_i$, i = 1, 2, ... K, be the coordinate values of a point $\hat{\jmath}$ on the capacity-bounding plane such that $\sum C_i = C_o$ and $\hat{\jmath}_i = [C_i/c_i]$. In the dedicated strategy, the sets A and $\sigma_i(k)$ are the inscribed rectangular solids

$$A = \prod_{i=1} \{j_i \mid 0 \geq j_i \leq \hat{\jmath}_i\} \tag{C-21}$$

$$\sigma_i(k) = \{k\} \times \prod_{\ell \neq i} \{0 \leq j_\ell \geq j_\ell\} \quad ,$$

from which the summations over A and $\sigma_i(k)$ are:

$$\sum_{j \in A} = \sum_{j_1=0}^{\hat{\jmath}_1} \cdots\cdots\cdots \sum_{j_K=0}^{\hat{\jmath}_K}$$

$$\sum_{j \in \sigma_i(k)} = \underbrace{\sum_{j_1=0}^{\hat{\jmath}_1} \cdots\cdots\cdots \sum_{j_\ell=0}^{\hat{\jmath}_\ell}}_{\substack{\text{(K-1) independent sums over } j_\ell \text{ with} \\ \ell=i \text{ deleted.}}}$$

Introduce the above sums back into Eq. C-20 using Eq. C-15 for $P_A(0)$ to obtain

$$\left\{ \sum_{j \in A} \prod_{i=1}^{K} \binom{N_i}{j_i} a_i^{j_i} \right\}^{K-1} = \left\{ \prod_{i=1}^{K} \sum_{j_i=0}^{\hat{\jmath}_i} \binom{N_i}{j_i} a_i^{j_i} \right\}^{K-1}$$

$$= \prod_{i=1}^{K} \left\{ \prod_{\ell \neq i} \sum_{j_\ell=0}^{\hat{\jmath}_\ell} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell} \right\}$$

$$= \prod_{i=1}^{K} \left\{ \sum \cdots \sum_{\ell \neq i} \prod_{\ell \neq i} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell} \right\}$$

$$= \prod_{i=1}^{K} \left\{ \sum_{j \in \sigma_i(k)} \prod_{\ell \neq i} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell} \right\},$$

which satisfies Eq. C-20.

The key condition for statistical independence of a user-class i is the invariance of $\sigma_i(k)$ over k values, which in turn allows separation of the sums $\sum_{j \in A}$ and $\sum_{j \in \sigma_i(k)}$ and the interchange of product and sum operations. Any other* A-set (such as fully shared access) will fail to satisfy Eq. C-20, and thus the user classes will not be statistically independent, even though Eqs. C-15 and C-21 are in product form.

6. Computational Factors

There are a number of computational challenges and limitations to evaluating even the simple model here addressed. The overflow problem can be mitigated by multiplying the $P_i(j_i)$ by an arbitrary constant such as 10^{-x} prior to normalization.** The normalization will cancel out in computing the constant P(0). The exponential value x must be chosen experimentally, as too large a value will cause computational underflow. This problem is accentuated for large (e.g., 10-to-1) differences between c_i values.

The much more serious problems reside in enlarging the number of user classes K and coping with increased levels of offered traffic per class and available capacity (i.e., increasing the upper limit on the number of active users before blocking). Both of these factors increase the number of state points j in the space $A \subseteq \Omega$, with a very rapid increase in computational steps. This issue is clearly appreciated by recent work in queueing networks as applied to computer science (Refs. 21-27).

A key factor which is exploited is the fact that all the users, no matter of what "class" (node of the network), use one

---

* The discussion addresses the case where all the user classes are mutually independent. This generalizes the cases where only some of the user classes are independent. Only those classes i for which $\sigma_i(k)$ is invariant on k are statistically independent. The others are dependent.

** This is most easily achieved when computing $P_i(j_i)$ recursively by initializing $P_i(0) = 10^{-x}$ rather than $P_i(0) = 1$

server, or, in the notation here, all $c_i \equiv 1$.* This makes a significant difference in that the constraint reads $\Sigma j_i \leq [C_o]$. This, in turn, allows summing (Ref. 21) the space $\Omega = \{j | j_1 \geq 0, \Sigma j_i \leq C_o\}$ recursively along disjoint parallel planes $P_k = \{j | j_i \geq 0, \Sigma j_i = k\}$, $k = 0, 1, 2, \ldots [C_o]$. Note that $\Omega = \sum_k P_k$. All the points j in Ω are "captured" this way because the c_i are the same (nominally equal to 1). This considerably simplifies (Ref. 25) the summation technique and reduces the number of computational steps. When the c_i are not equal, this approach fails, because not all the states are "captured" if one tries to sum the planes $\Sigma c_i j_i = [C_o/\max c_i]$. This is best seen in Fig. C-3, where in (a) $c_1 = c_2 = 1$ and in (b) $c_1 = 1$, $c_2 = 2$. In Figure C-3b the states off the plane are crossed. If the c_i are integer multiples of one,** the state space can be augmented with dummy states with probability weight of zero in order to obtain "equivalent" equal c_i. For example, in Fig. III-10b each j_2 state could be split in half (and the j_2 axis doubled in length), and every other "new j_2" state would have probability zero. This would allow using the planar sum technique. However, since the number of state points is increased, it is not clear that any reduction in computational steps will be achieved. The software advantages in writing the summation algorithms do pertain.

Another computing technique that might be of interest is contained in Refs. 24 and 25. Conventional (i.e., old-fashioned RLC) techniques of network and filter theory are reviewed for possible application to queueing networks. These approaches

---

*These efforts are addressing a computer science problem where the user tasks are assumed divisible into basic units of work. Thus, one user (unit of work) uses one server (unity) of capacity.

**This occurs in time-division multiplex where the c_i are multiples of 2 times basic rate. However, in satellite systems with different EIRP x (G/T), the c_i need not be integer multiples.

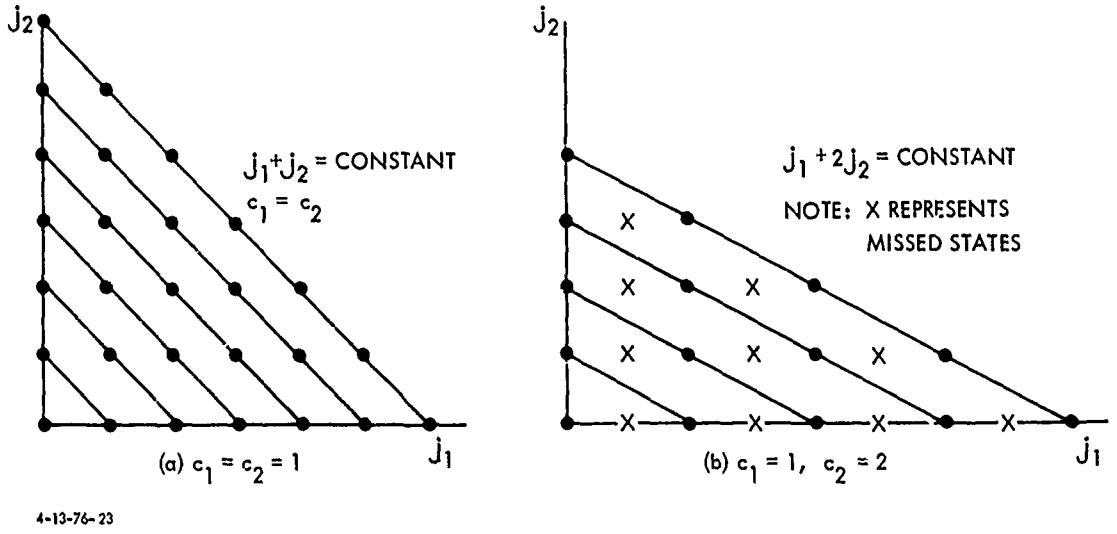


FIGURE C-3. Planar States Missed

may have more utility to the problem here, in that they do not depend inherently on the equality of the c_i.

7. Performance Measures

Equation C-17 gives a basic probabilistic description for experimentally distributed traffic of all the admissible states $j \in A$ for Blocked-Calls-Cleared systems. From $P(j)$, the blocking probabilities and utilization can be calculated. The average percentage of capacity used gives a measure of facility utilization, a quantity of interest to the *system provider*. From the opposite point of view, any specific user of a class wants to know the percentage of time (i.e., the probability) he will be blocked or denied service (e.g., a circuit). These two aspects are now discussed in turn. Facility utilization, being the easier, is treated first.

a. Capacity Utilization. Since in state $j \in A$ the fractional capacity used is $(c,j)/C_o$, the average utilization of the system is given by

$$U_A \triangleq \langle (c,j)/C_o \rangle \tag{C-22}$$

$$U_A = \sum_{j \in A} (c,j)\ P(j)/C_o \ ,$$

where U_A depends on traffic characteristics, a ailable capacity, and access strategy or A-set. But, by the linearity of the averaging process and the definition $Q_i(k)$ of Eq. C-19

$$U_A = \sum_{i=1}^{K} \frac{c_i}{C_o} \langle j_i \rangle \ . \qquad \text{(C-23)}$$

The general ν^{th} ($\nu > 0$) movement* of j_i is given by

$$\langle j_i^{\nu} \rangle = \sum_{k=0}^{\hat{j}_i} k^{\nu} \, Q_i(k) \ . \qquad \text{(C-24)}$$

b. Blocking States. There are any number of blocking conditions and grade-of-service probabilities of interest. They all depend on the method of partitioning the available capacity or, more specifically, the set A of admissible states. The blocked states of interest will be subsets of A. Their probability of occurrence will be given by summing $P(j)$ given in Eq. C-17 over those states in the blocking set B of interest. They will also depend on the user class. The probability that at least one user class is blocked is the probability that $j \in \overline{A}$ (the upper boundary of A). That is,

$$P(\text{blocking on at least one user class}) = P(\overline{A}), \qquad \text{(C-25)}$$

where

*In Eq. C-23, $\nu = 1$.

$$\{\overline{A} = j \in A \mid j + \delta_i \notin A \text{ for some } i = 1, 2, \ldots K\}$$

$$\{\delta_i\}_\ell = 0 \text{ if } \ell \neq i$$

$$= 1 \text{ if } \ell = i \ .$$

Consequently, the average unused capacity or "overhead reserve" is $1 - P(\overline{A})$.

Let the set $B_i \subseteq \overline{A}$ be given by

$$B_i \triangleq \{j \in \overline{A} \mid j_i + 1 \notin A\} \ . \tag{C-26}$$

B_i is the subset of $\overline{A}$ with j states whose i^{th} coordinate is on $\overline{A}$. Note that the B_i, $i = 1, 2, \ldots K$, need not be mutually disjoint, but that

$$\bigcup_{i=1}^{K} B_i = \overline{A} \ .$$

Thus, it follows that

$$P(i^{th} \text{ user class blocked}) = P(B_i) \tag{C-27}$$

$$P(B_i) \leq P(\overline{A}) \leq \sum_{i=1}^{K} P(B_i).$$

An illustrative example showing B_1 and B_2 in two dimensions is given in Fig. C-4 for partitioned and fully available access. Note that in fully shared access $B_1 \subset B_2 = \overline{A}$. Figure C-3 suggests the (obvious) demand-access ordering lemma:

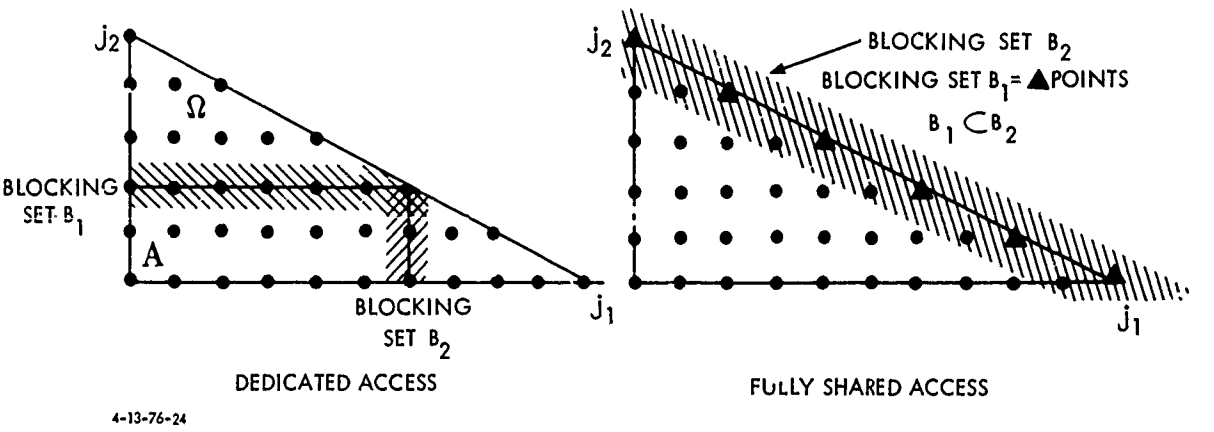


FIGURE C-4. Blocking Sets

Lemma:* Order the user classes $i = 1, 2, \ldots K$ so that $c_1 \leq c_2 \leq \ldots \leq c_K$. Then, for fully shared access, the blocking states are ordered $B_i \subseteq B_{i+1}$, and consequently $P(B_i) \leq P(B_{i+1})$.

Note that from the definition of B_i, $i = 1, 2, \ldots K$, for any $j \in B_i$,

$$C_o - c_i < \sum_{\ell=1}^{K} c_\ell j_\ell \leq C_o \quad .$$

But $c_{i+1} \geq c_i$, so that

$$C_o - c_{i+1} < \sum_{\ell=1}^{K} c_\ell j_\ell \leq C_o \quad .$$

*Physically, one would expect that those users requiring more capacity would be blocked more often. However, on reflection one would pause to contemplate the effect the parameters λ, μ might imply.

It then follows that $j \in B_{i+1}$. If the c_i are integer multipliers of each other (e.g., $c_i = 2^{i-1}c_1$), the B_i are properly included in B_{i+1}, and $P(B_i) < P(B_{i+1})$.

Remark: The ordering of the blocking probabilities does not depend on the values of the offered traffic (a_i, N_i).

c. Blocking Probabilities and Congestion. An extremely important distinction for finite population sources ($N_i < \infty$) is the concept of the probability of service being blocked as seen by an arriving customer (call congestion) in contrast to the probability of full capacity usage (time congestion). The calculations for $P(j)$ correspond to time congestion. For finite populations, an arriving customer implies knowledge that at least one traffic source (himself) is inactive. Thus, active population in class i is reduced by one to $N_i - 1$. For infinite populations the distinction is unnecessary.

To calculate for the finite population, the call congestion, a simple guess would be modified $Q_i(k)$ in Eq. C-29 by replacing N_i with $N_i - 1$ (an $N_i - 1$ source model). It is shown next that for the multidimensional Birth-Death model with coordinate-convex A-sets and Blocked Calls Cleared, this guess is correct. In Ref. 1 this property is shown to hold for a one-dimensional Birth-Death process with quasi-random arrivals and general holding times (not necessarily exponential) and queueing strategies (e.g., Blocked Calls Held). However, Ref. 1 shows by example that for multidimensional quasi-random Birth-Death processes the desired property need not hold. Paralleling Ref. 1, it is verified here that for the case at hand with exponentially distributed traffic and Blocked Calls Cleared, the call congestion probabilities (arriving calls) can be computed by replacing N_i with $N_i - 1$ in Eq. C-15.

The steady-state probability of there being k calls in progress from class i with a user population of N is given by Eq. C-19 with N_i replaced by N. For a more explicit notation,

replace $Q_i(k)$ with $Q_i(k;N)$ to show the dependency on N in user class i. Note that where N_i is reduced by one to $N_i - 1 = N - 1$, the set A may change, and thus $P_A(0)$ can be affected.

Following the heuristic approach of Ref. 3, for a long period of time T, the average number of call attempts by idle users (users without calls in progress) who also find k calls already in progress is given by the product of the average number of calls a single user places in time T, $(\lambda_i T)$, the number of idle users given k calls in progress, $(N - k)$, and the probability of k calls in progress, $Q_i(k,N)$. The probability that an arriving call finds k calls in progress, $\gamma_i(k,N)$, must be proportional to this product $(\lambda_i T)\ (N-k)\ Q_i(k,N)$ and must sum to one over all possible k values. Since there is always at least one arriving call, the largest k value, $\hat{k}_i$, is the smallest of $\hat{j}_i$ and $(N_i - 1)$. Thus $\gamma_i(k,N)$ is obtained by dividing $(\lambda_i T)\ (N - k)\ Q_i(k,N)$ by the sum over $0 \le k \le k_i$ of this product and letting $T \to \infty$, when this is done $\lambda_i T$ divides out, and the result is given by

$$\gamma_i(k,N) = \frac{(N-k)\ Q_i(k,N)}{\sum_{m=0}^{\hat{k}_i} (N-m)\ Q_i(m,N)} \qquad \text{(C-28)}$$

$$k_i = \min\ [\hat{j}_i,\ N_i - 1] \quad .$$

The objective is to show that $\gamma_i(k,N) = Q_i(k,N-1)$. Using Eq. C-19 and remembering $k \le \hat{k}_i$,

$$\gamma_i(k,N) = \frac{P_A(0)\ (N-k)\ \binom{N}{k} a_i^k\ p_{\sigma_i}(k)}{\sum_{m=0}^{\hat{k}_i} P_A(0)(N-m)\binom{N}{m}\ a_i^m\ p_{\sigma_i}(m)}$$

$$= \frac{NP_A(0)\ \binom{N-1}{k}\ a_i^k\ p_{\sigma_i}(k)}{NP_A(0)\ \sum_{m=0}^{\hat{k}_i}\ \binom{N-1}{m}\ a_i^m\ p_{\sigma_i}(m)}\ , \qquad \text{(C-29)}$$

where

$$p_{\sigma_i}(m) \triangleq \sum_{j \in \sigma_i(m)}\ \prod_{\ell \neq i}\ \binom{N}{j_\ell} a_\ell^{j_\ell}\ .$$

Cancelling the $NP_A(0)$ term in Eq. C-29 produces

$$\gamma_i(k,N) = G^{-1} \cdot \binom{N-1}{k}\ a_i^k\ p_{\sigma_i}(k) \qquad \text{(C-30)}$$

$$G \equiv \sum_{m=0}^{\hat{k}_i}\ \binom{N-1}{m}\ a_i^m\ p_{\sigma_i}(m)\ .$$

Equation C-30 will now be compared to $Q_i(k, N-1)$. Note that a change in N_i from N to N - 1 may cause a change in the normalization constant $P_A(0)$ in Eq. C-19.

In Eq. C-19 the particular i^{th} coordinate $\binom{N_i=N}{j_i}$ term for $Q_i(k, N-1)$ will be reduced $\binom{N-1}{j_i}$. Since the i^{th} coordinate is restricted not to exceed N - 1, the A-set must be modified. From the definition of $\sigma_i(k)$ in Eq. C-19, the original A-set (i.e., $N_i = N$) is given by the disjoint union of $\sigma_i(m)$ over m (for any i):

$$A = \bigcup_{m=0}^{\hat{J}_i} \sigma_i(m). \qquad \text{(C-31)}$$

Reducing N_i by one to N - 1 can only limit the largest m value in Eq. C-31 to the smaller of N - 1 or $\hat{J}_i$ (i.e., $\hat{k}_i$). Thus the modified A, $\hat{A}$, is the disjoint union of $\sigma_i(m)$ sets:

$$\hat{A} = \bigcup_{m=0}^{\hat{k}_i} \sigma_i(m) \quad .$$

Since the $\sigma_i(m)$ are disjoint, the inverse of the modified normalization $P_{\hat{A}}(0)$ is given by

$$P_{\hat{A}}^{-1}(0) = \sum_{j \in \hat{A}} \left\{ \binom{N-1}{j_i} a_i^{j_i} \prod_{\ell \neq i} \binom{N_\ell}{j_\ell} a_\ell^{j_\ell} \right\}$$

$$= \sum_{m=0}^{\hat{k}_i} \binom{N-1}{j_i} a_i^{j_i} \, p_{\sigma_i}(m)$$

$$\equiv G \quad .$$

Then $Q_i(k, N-1)$ is given by

$$Q_i(k, N-1) = G \binom{N-1}{k} a_i^k \, p_{\sigma_i}(k)$$

$$\equiv \gamma_i(k, N), \quad 0 \leq k \leq \hat{k}_i$$

C. COMPUTER PROGRAM

This section discusses the computer program that was used for the comparison of dedicated and fully shared capacity allocation examples. The program was written in Fortran IV

and run on a commercial time-sharing service. The program is limited to only three user classes (K = 3).

1. Grade-of-Service Bound

Before discussion of the program, the upper bound used on blocking probabilities in the fully shared allocation strategy to satisfy a grade-of-service (GOS) objective is reviewed. In the fully shared allocation the A-set is identically equal to Ω. Define the Ω^+ of j states as

$$\Omega^+ = \{j \in \Omega \mid C_o - c_3 < j_1 + c_2 j_2 + c_3 j_3 \leq C_o\} \quad . \qquad \text{(C-33)}$$

Thus, Ω^+ is the collection of those states in which the arrival of one more class 3 customer would overrun the available capacity C_o. The blocking states B_i must be contained in the upper boundary states of Ω, denoted $\overline{\Omega}$ (i.e., $B_i \subseteq \overline{\Omega}$). If $\overline{\Omega}$ were a subset of Ω^+, then $P(B_i) \leq P(\overline{\Omega}) \leq P(\Omega^+)$. Suppose the state j belongs to $\overline{\Omega}$. Then, for some i = 1, 2, 3,

$$C_o - c_i < j_1 + c_2 j_2 + c_3 j_3 \leq C_o \ ,$$

and since $c_2 \leq c_3$, the left-hand of the inequality is $\geq C_o - c_3$, and it follows that j belongs to Ω^+. Thus, $\overline{\Omega} \subseteq \Omega^+$, and $P(\Omega^+)$ is an upper bound to blocking on any user class. This use of the GOS is extremely easy to incorporate in the program, as the set Ω^+ is one of a nested sequence of sets used to recursively sum the whole space Ω.

Description

The concept employed here starts with user data as given and a specified blocking probability objective. It is assumed that there is no limit to available capacity. Capacity needed to meet the blocking objectives for each class is then calculated. The sum of these capacities, C_o, is what is needed to

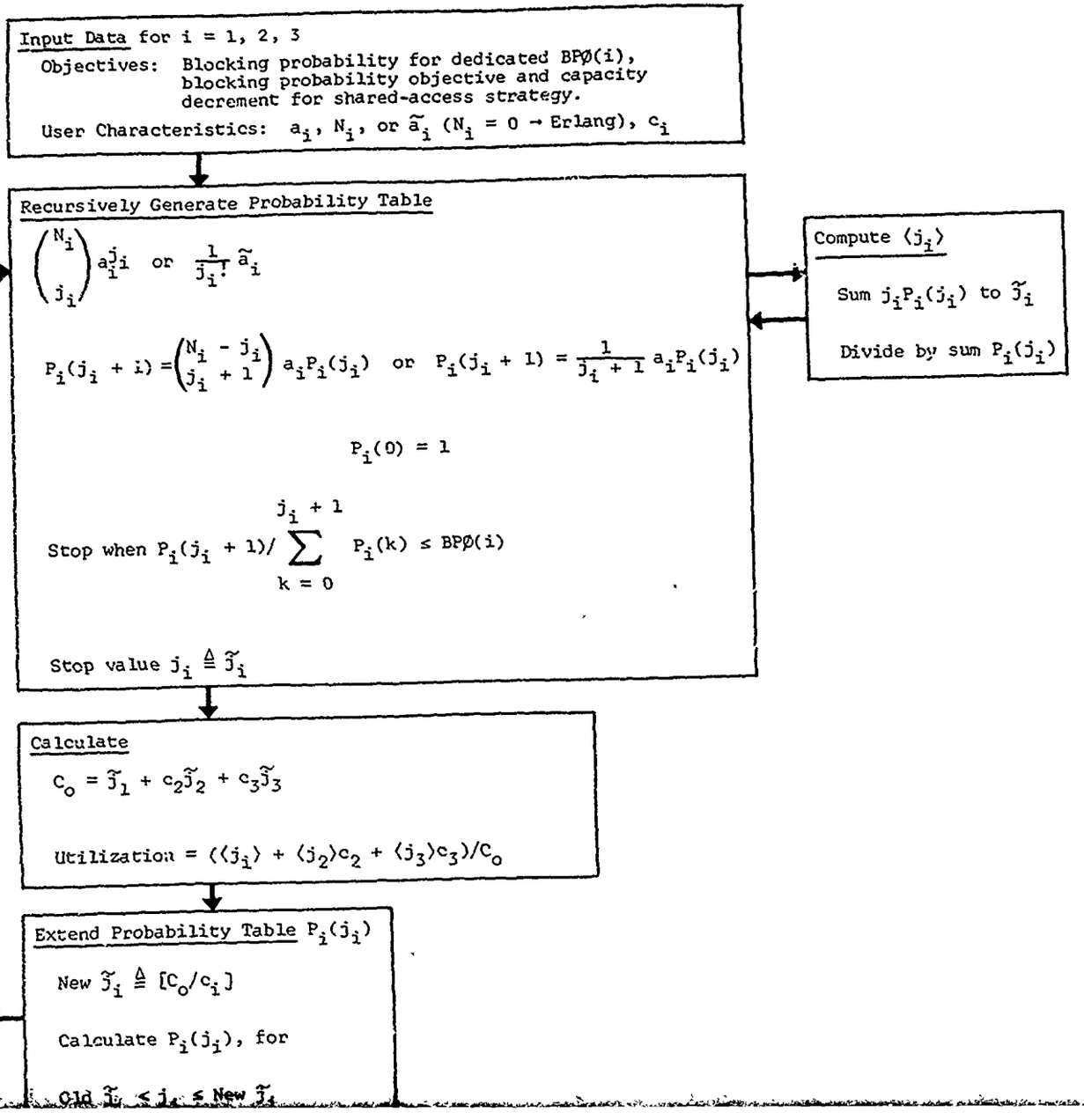
meet the GOS objective with dedicated allocation. This value of C_o is then used to initialize a calculation of $P(\Omega^+)$ for the fully shared allocation. C_o is then iteratively decremented a fixed amount until $P(\Omega^+) > GOS$. The C_o needed for the fully shared allocation is then taken to be that value on the immediately preceding iteration.

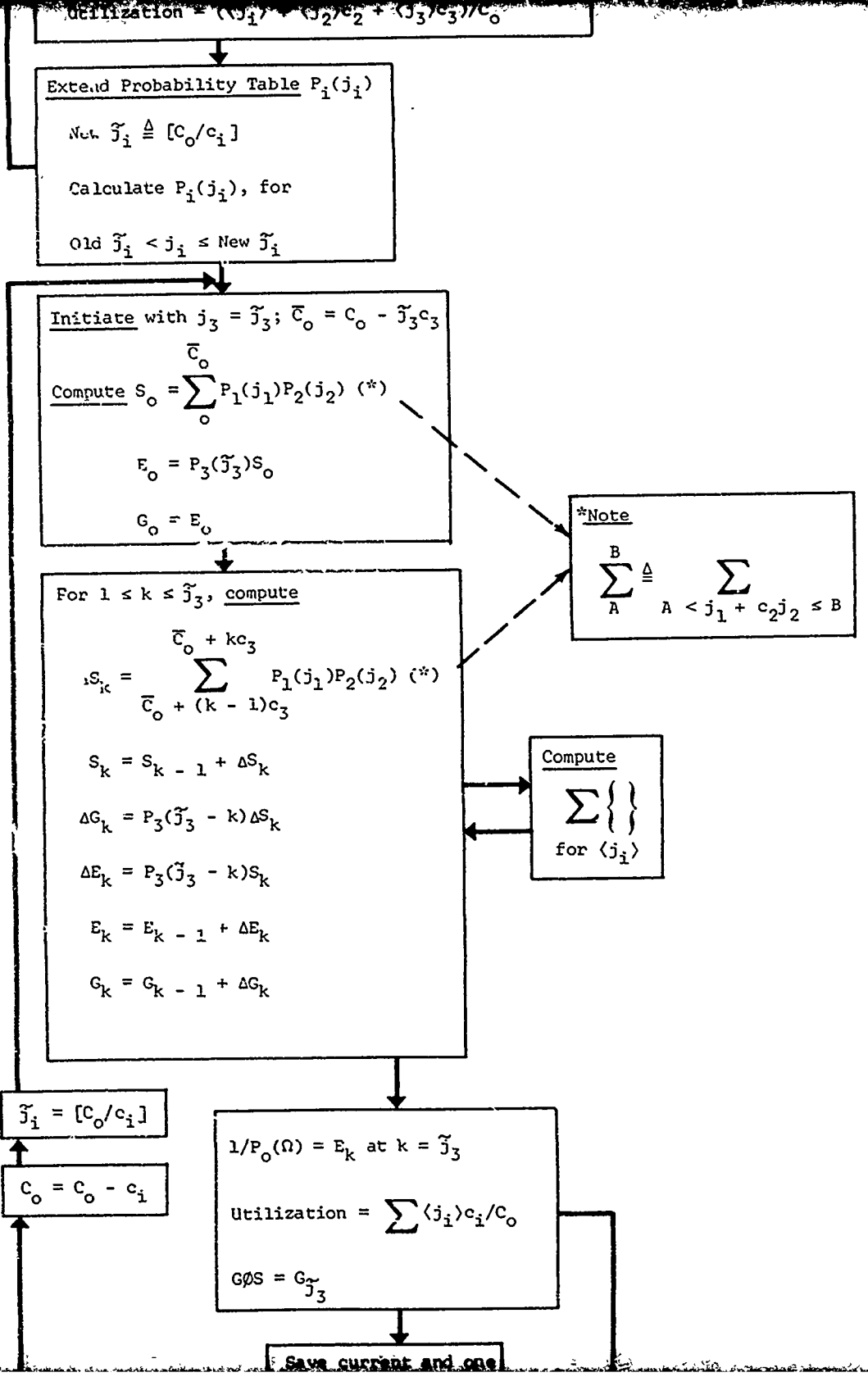
The computational method is flow-charted in Fig. C-5. The input data as to user characteristics and Engset or Erlang probability tables are recursively generated to obtain the necessary number of "circuits," to meet the GOS objectives. This, in turn, sets the total capacity C_o required for dedicated uses as $\Sigma\, c_i \tilde{j}_i$. With this amount of capacity available, the maximum number of available circuits to each user class is increased to $\tilde{j}_i = [C_o/c_i]$, and the probability tables are extended upward to these values of maximum circuits. Then the sum over Ω is performed recursively by starting with j_3 at its maximum $\tilde{j}_3$ and cyclically reducing by one unit of c_3 (dropping out a class 3 customer), until $j_3 = 0$. Let $P(j_i)$ denote the unnormalized Engset or Erlang probabilities (see Eq. III-17 of the main text). The calculation of $P_\Omega(0)$ by summing the j states in Ω is performed by using the following recursive representation:

$$\sum_{j\in\Omega} \begin{Bmatrix} j_1 \\ j_2 \\ j_3 \end{Bmatrix} \prod_{i=1}^{3} P_i(j_i) =$$

$$\begin{Bmatrix} 1 \\ 1 \\ \tilde{j}_3 \end{Bmatrix} P_3(\tilde{j}_3) \sum_{o}^{\bar{C}_o} \begin{Bmatrix} j_1 \\ j_2 \\ 1 \end{Bmatrix} P_1(j_1)P_2(j_2)$$

$$+ \begin{Bmatrix} 1 \\ 1 \\ \tilde{j}_3-1 \end{Bmatrix} P_3(\tilde{j}_3-1)\left[\sum_{o}^{\bar{C}_o} + \sum_{\bar{C}_o}^{\bar{C}_o+c_3} \begin{Bmatrix} j_1 \\ j_2 \\ 1 \end{Bmatrix} P_1(j_1)P_2(j_2)\right]$$


Input Data for i = 1, 2, 3
Objectives: Blocking probability for dedicated BPØ(i), blocking probability objective and capacity decrement for shared-access strategy.
User Characteristics: a_i, N_i, or $\tilde{a}_i$ ($N_i = 0 \rightarrow$ Erlang), c_i
Recursively Generate Probability Table
$\binom{N_i}{j_i} a_i^{j_i}$ or $\frac{1}{j_i!}\tilde{a}_i$
$P_i(j_i + 1) = \binom{N_i - j_i}{j_i + 1} a_i P_i(j_i)$ or $P_i(j_i + 1) = \frac{1}{j_i + 1} a_i P_i(j_i)$
$P_i(0) = 1$
Stop when $P_i(j_i + 1) / \sum_{k=0}^{j_i + 1} P_i(k) \leq$ BPØ(i)
Stop value $j_i \triangleq \tilde{j}_i$
Compute $\langle j_i \rangle$
Sum $j_i P_i(j_i)$ to $\tilde{j}_i$
Divide by sum $P_i(j_i)$
Calculate
$C_o = \tilde{j}_1 + c_2\tilde{j}_2 + c_3\tilde{j}_3$
Utilization = $(\langle j_1 \rangle + \langle j_2 \rangle c_2 + \langle j_3 \rangle c_3)/C_o$
Extend Probability Table $P_i(j_i)$
New $\tilde{j}_i \triangleq [C_o/c_i]$
Calculate $P_i(j_i)$, for
Old $\tilde{j}_i \leq j_i \leq$ New $\tilde{j}_i$


Utilization = $(\langle j_1\rangle c_1 + \langle j_2\rangle c_2 + \langle j_3\rangle c_3)/C_o$
Extend Probability Table $P_i(j_i)$
New $\tilde{j}_i \triangleq [C_o/c_i]$
Calculate $P_i(j_i)$, for
Old $\tilde{j}_i < j_i \leq$ New $\tilde{j}_i$
Initiate with $j_3 = \tilde{j}_3$; $\bar{C}_o = C_o - \tilde{j}_3 c_3$
Compute $S_o = \sum_o^{\bar{C}_o} P_1(j_1)P_2(j_2)$ (*)
$E_o = P_3(\tilde{j}_3)S_o$
$G_o = E_o$
*Note
$\sum_A^B \triangleq \sum_{A < j_1 + c_2 j_2 \leq B}$
For $1 \leq k \leq \tilde{j}_3$, compute
$\Delta S_k = \sum_{\bar{C}_o + (k-1)c_3}^{\bar{C}_o + kc_3} P_1(j_1)P_2(j_2)$ (*)
$S_k = S_{k-1} + \Delta S_k$
$\Delta G_k = P_3(\tilde{j}_3 - k)\Delta S_k$
$\Delta E_k = P_3(\tilde{j}_3 - k)S_k$
$E_k = E_{k-1} + \Delta E_k$
$G_k = G_{k-1} + \Delta G_k$
Compute
$\sum\{\ \}$
for $\langle j_i\rangle$
$\tilde{j}_i = [C_o/c_i]$
$C_o = C_o - c_i$
$1/P_o(\Omega) = E_k$ at $k = \tilde{j}_3$
Utilization = $\sum \langle j_i\rangle c_i/C_o$
GØS = $G_{\tilde{j}_3}$
Save current and one

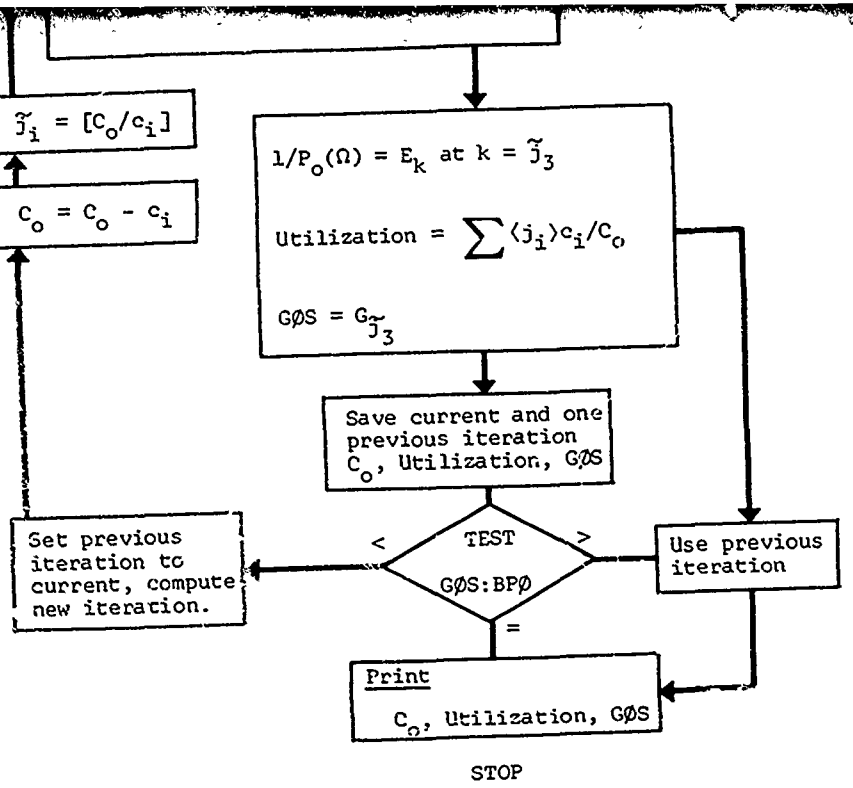


FIGURE C-5. Program Flow Chart

$$+ \begin{Bmatrix} 1 \\ 1 \\ \tilde{j}_3-2 \end{Bmatrix} P_3(\tilde{j}_3-2)\left[\sum_{o}^{\bar{C}_o} + \sum_{\bar{C}_o}^{\bar{C}_o+c_3} + \sum_{\bar{C}_o+c_3}^{\bar{C}_o+2c_3} \begin{Bmatrix} j_1 \\ j_2 \\ 1 \end{Bmatrix} P_1(j_1)P_2(j_2)\right]$$

$$\vdots$$

$$+ \begin{Bmatrix} 1 \\ 1 \\ 0 \end{Bmatrix} \cdot 1 \cdot \left[\sum_{o}^{\bar{C}_o} + \ldots + \sum_{C_o-c_3}^{C_o} \begin{Bmatrix} j_1 \\ j_2 \\ 1 \end{Bmatrix} P_1(j_1)P_2(j_2)\right] ,$$

where

$$\bar{C}_o \triangleq C_o - c_3\tilde{j}_3$$

$$\sum_{A}^{B} \triangleq \sum_{A<j_1+c_2j_2\leq B}$$

$$\begin{Bmatrix} j_1 \\ j_2 \\ 1 \end{Bmatrix} = \begin{matrix} 1 & \text{for calculating } P_o(\Omega) \text{ or } \langle j_3\rangle \\ j_1 & \text{for calculating } \langle j_1\rangle \\ j_2 & \text{for calculating } \langle j_2\rangle \end{matrix}$$

$$\begin{Bmatrix} 1 \\ 1 \\ \tilde{j}_3-k \end{Bmatrix} = \begin{matrix} 1 & \text{for } P_o(\Omega), \langle j_1\rangle \text{ , or } \langle j_2\rangle \\ \tilde{j}_3-k & \text{for } \langle j_3\rangle \text{ .} \end{matrix}$$

The expansion of the sum is taken with decreasing j_3 to get recursively increasing j_1, j_2 states. At each recursive step k, the only new sum states needed are those which lie between $(k-1)c_3 + C_o \leq j_i + c_2j_2 \leq kc_3 + C_o$. The other states lying between 0 and $(k-1)c_3 + \bar{C}_o$ were calculated in the k - 1 preceding stages. The curly bracket terms are shown to indicate how the average j_i are found. This is not shown in the flow chart so as to retain relative simplicity. Once the technique of summing over Ω space is adopted, average j_i calculations are trivial.

$P_\Omega(o)$ corresponds to the sum with $\begin{Bmatrix}1\\1\\1\end{Bmatrix}$. Moreover, the GOS estimate is also trivial, as it corresponds to summing only

$$\sum_{k=0}^{\tilde{j}_3} P_3(\tilde{j}_3-k)\left[\sum_{\bar{c}_o+(k-1)c_3}^{\bar{c}_o+kc_3} P_1(j_1)P_2(j_2)\right],$$

which has to be accumulated in any event.

3. Granularity

The effects of granularity in the above scheme must be noted. These will be most pronounced for light to moderately loaded low-capacity systems. The granularity develops basically because the system states (i.e., circuits) are integers. Consequently, for small system capacities the rounding up will cause noticeable percentage jumps. (The circuits needed will exhibit more noticeable jump phenomena with respect to offered traffic at low Erlang values.) In addition, in the approach used here, the capacity for the shared mode is decremented in discrete steps. Thus, although the capacity required versus offered traffic will be monotonically increasing, they will exhibit discontinuities and variations about a smooth curve. Even more noticeable are fluctuations in utilization. Although capacity is monotonically increasing, the effect of granularity on utilization can cause increased utilization efficiency at some immediately lower system sizing. This is purely an artifact of the computation. Variations in utilization efficiency of a few percent should not be overemphasized because, in addition to granularity, the Birth-Death model need not be accurate to that precision.

REFERENCES, APPENDIX C

1. R. B. Cooper, *Introduction to Queueing Theory*, MacMillan Company, New York, 1972.
2. J. W. Cohen, "The Generalized Engset Formulae," *Phillips Telecommunication Review*, November 1957, pp. 158-170.
3. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Vol. 1, Part 1, Oliver and Boyd, Edinburgh and London, 1960.
4. V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.
5. L. Kleinrock, *Queueing Systems*, Vol. I: *Theory*, John Wiley & Sons, New York, 1975.
6. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, John Wiley & Sons, New York, 1974.
7. Gunnar F. W. Fredrickson, "Analysis of Channel Utilization in Traffic Concentrators," *IEEE Trans. on Comm.*, Vol. COM-22, No. 8, August 1974.
8. G. Frenkel, "The Grade of Service in Multiple-Access Satellite Communications Systems with Demand Assignments," *IEEE Trans. on Comm.*, Vol. COM-22, No. 10, October 1974, pp. 1681-1685.
9. O. Enomoto and H. Miyamoto, "An Analysis of Mixtures of Multiple Bandwidth Traffic on Time Division Switching Networks," *Proc. Seventh International Teletraffic Congress*, Swedish Telecommunications Administration (Televerket), Stockholm, Sweden, 13-20 June 1973.
10. K. M. Ollsson, M. Anderberg, and G. Lind, "Report on the Seventh International Teletraffic Congress in Stockholm, June 13-20, 1973," *Ericsson Technics*, Telefonaktiebolaget L. M. Ericsson, Stockholm, Sweden, 1974.
11. L. A. Gimpelson, "Analysis of Mixtures of Wide- and Narrow-Band Traffic," *IEEE Trans. on Comm.*, Vol. 13, No. 3, September 1965, pp. 258-266.
12. Eckart Wollner, "A Queueing Problem in Data Transmission," *Proc. Seventh International Teletraffic Congress*, Swedish Telecommunications Administration (Televerket), Stockholm, Sweden, 13-20 June 1973.

13. B. A. Whitaker, "Analysis and Optimal Design of a Multi-server, Multiqueue System with Finite Waiting," *Bell System Technical Journal*, Vol. 54, No. 3, March 1975, pp. 595-623.

14. Network Analysis Corporation, *The Practical Impact of Recent Computer Advances on the Analysis and Design of Large Scale Networks*, First Semiannual Technical Report, H. Frank, May 1973.

15. L. G. Roberts and B. D. Wessler, "Computer Network Development in Achieving Resource Sharing," *Proc. Spring Joint Computer Conference*, 1970, pp. 543-549.

16. O. A. Pederson, "The Design of Gradings with Small Interconnection Numbers for Random Hunt Selectors," *IEEE Trans. on Comm.*, Vol. COM-23, July 1975, pp. 714-721.

17. A. Lotze, "History and Development of Grading Theory," *Proc. Seventh International Teletraffic Congress*, Swedish Telecommunications Administration (Televerket), Stockholm, Sweden, 13-20 June 1973.

18. J. R. Jackson, "Networks of Waiting Lines," *Operations Research*, Vol. 5, 1957, pp. 518-521.

19. W. J. Gordon and G. F. Newell, "Closed Queueing Systems with Exponential Servers," *Operations Research*, 1965, pp. 254-265.

20. L. Kleinrock, "Analysis of a Time-Shared Processor," *Naval Res. Logistics Quart.*, Vol. 11, 1964, pp. 59-73.

21. J. P. Buzen, "Computational Algorithms for Closed Queueing Networks with Exponential Servers," *Communications of the ACM*, Vol 16, No. 9, September 1973, pp. 527-531.

22. Department of Industrial Engineering, University of Michigan, Ann Arbor, *Network Models for Large-Scale Time-Sharing Systems*, Ph.D. Thesis, IR-71-1, C. G. Moore, III, April 1971.

23. IBM Research Center, *Horner's Rule for the Evaluation of General Closed Queueing Networks*, RC-5219, M. Reiser and H. Kobayashi, 16 January 1975.

24. K. M. Chandy, U. Herzog, and L. Woo, "Parametric Analysis of Queueing Networks," *IBM J. Res. Develop.*, January 1975, pp. 36-42.

25. K. M. Chandy, U. Herzog, and L. Woo, "Approximate Analysis of General Queueing Networks," *IBM J. Res. Develop.*, January 1975, pp. 43-49.

26. IBM Research Center, *Numerical Methods in Queueing Networks*, RC-5344, M. Reiser and H. Kobayashi, 27 March 1975.

27. H. Kobayashi, "Application of the Diffusion Approximation to Queueing Networks, I: Equilibrium Queue Distributions, II: Nonequilibrium Distributions and Applications to Computer Modeling," *J. Assoc. Comp. Mach.*, Vol. 21, No. 2, April 1974, pp. 316-328, and No. 3, July 1974, pp. 459-469.

28. Computer Sciences Department, University of Texas, Austin, *Local Balance, Robustness, Poisson Departures and Product Form in Queueing Networks*, Research Report TR-15, K. M. Chandy, J. Howard, T. W. Keller, and D. J. Towsley, 1973.

29. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley & Sons, New York, 1957.

APPENDIX D

DERIVATIONS FOR FIXED ASSIGNMENTS.

CONTENTS

APPENDIX D

DERIVATIONS FOR FIXED ASSIGNMENTS

A. BUFFER CONTENTS

The generating function for the steady-state buffer queue length at a user in a net operating under fixed assignments is developed in this appendix. The steady-state moments of the buffer contents are obtained from this generating function. The development parallels that of Ref. 1.

For the purpose of the derivation, assume that the cycle with respect to the i^{th} user is initiated at the beginning of his allocated transmission time. For example, with contiguous allocations, the user is allowed to transmit data units in the first r slots and then is inactive with respect to data transmission for M - r slots. The buffer queue length in data units at the i^{th} user at the beginning of the j^{th} cycle (where the superscript i is suppressed) is L_j. The buffer contents at beginning of the $j+1^{st}$ cycle are given by the following equation:

$$L_{j+1} = (L_j - r)^+ + Y_j, \qquad \text{(D-1)}$$

where Y_j is the total number of data-unit arrivals in the j^{th} cycle, and r is the maximum number of data units that can be transmitted in a cycle. A data unit can only be transmitted if at least one is stored, and the + indicates this. The buffer contents are examined only at the beginnings of the cycles, since a steady-state solution does not exist if the contents are examined at all of the time slots. Equation D-1 indicates that the contents at the beginning of the $j+1^{st}$ cycle are equal to the contents at the start of the j^{th} cycle,

minus the data units that are transmitted during the j^{th} cycle, plus the new arrivals that occur. The equation is written in this form instead of that of Eq. IV-2 so that the buffer contents can be determined for both contiguous allocations where the length of any cycle is M slots and distributed allocations where the length is random, M_j being the length of the j^{th} cycle. The total number Y_j of data units that arrive during the j^{th} cycle with length M_j is given by

$$Y_j = \sum_{k=1}^{M_j} X_k, \qquad \text{(D-2)}$$

where X_k is the number of data-unit arrivals during the k^{th} slot of the j^{th} cycle. The length of a cycle is a discrete random variable which assumes the values $m_1, m_2, \ldots m_N$ with $P\{M_j=m_i\} = p_i$. The cycle lengths are chosen from this distribution independently from cycle to cycle. The data-unit arrival process was assumed to be compound Poisson, arrival instances being Poisson distributed and the number of data units at an arrival being geometric. Since the arrival times are Poisson, the number of arrivals X_k during the k^{th} slot are independent of the arrivals during the i^{th} slot for $k \neq i$.

The generating function for the buffer contents at the j^{th} cycle is, by definition,

$$H_j(z) = E\left\{z^{L_j}\right\} = \sum_{k=0}^{\infty} z^k \, P\left\{L_j = k\right\} \qquad \text{(D-3)}$$

The generating function is analytic within the contour $|z|=1$. By using Eq. D-1, the generating function for the buffer contents at the $j+1^{st}$ cycle is given by

$$H_{j+1}(z) = E\left\{z^{(L_j - r)^+ + Y_j}\right\} = E\left\{z^{(L_j - r)^+}\right\} E\left\{z^{Y_j}\right\} \qquad (D\text{-}4)$$

since the arrivals during the j^{th} cycle are independent of the buffer contents at the beginning of the cycle. We see from Eq. D-4 that the generating function for the input process Y_j has to be determined:

$$S_j(z) = E\left\{z^{Y_j}\right\} = \sum_{n=0}^{\infty} z^n \, P\left\{Y_j = n\right\}.$$

The probability of exactly n data-unit arrivals can be obtained by using conditional probabilities:

$$P\left\{Y_i = n\right\} = \sum_{k=1}^{N} P\left\{Y_i = n / M_j = m_k\right\} p_k.$$

The arrivals in a cycle, given that the cycle length $M_j = m_k$, are $\sum_{i=1}^{m_k} X_i$, from Eq. D-2 where the X_i's are independent and identically distributed. The generating function of a sum of independent random variables is equal to the product of the respective generating functions. Therefore, the input process generating function is

$$S_j(z) = \sum_{k=1}^{N} p_k \, P^{m_k}(z), \qquad (D\text{-}5)$$

where $P(z)$ is the generating function of X_k and is equal to $\exp \lambda\left(\frac{z-1}{1-qz}\right)$. The input process generating function is independent of j, a. i therefore the subscript j can be dropped. Returning to Eq. D-4, we see that

$$E\left\{z^{(L_i - r)^+}\right\} = \sum_{k=0}^{r=1} P\left\{L_j = k\right\} + \sum_{k=r}^{\infty} z^{k-r} \, P\left\{L_j = k\right\}$$

$$= z^{-r}\left\{H_j(z) + \sum_{k=0}^{r-1} (z^r - z^k) \, P\left(L_j = k\right)\right\}.$$

The buffer-contents generating function satisfies the following difference equation, which is obtained by substituting the above equation into Eq. D-4:

$$H_{j+1}(z) = z^{-r}\, S(z) \left\{ H_j(z) + \sum_{k=0}^{r=1} (z^r - z^k)\, P(L_j = k) \right\} \quad \text{(D-6)}$$

If a stationary solution exists, that is, $\lim_{j\to\infty} H_j(z) = H(z)$, the solution will satisfy the following equation:

$$H(z) = \frac{S(z) \sum_{k=0}^{r=1} (z^r - z^k)\, P(L_j = k)}{z^r - S(z)} .$$

The probabilities $P\{L_j = k\}$ $k = 0, \ldots, r-1$ are unknown. The denominator $z^r - S(z)$ may have zeroes within the unit circle, and, because $H(z)$ is a generating function which must be analytic within the unit circle, the numerator must have zeroes which coincide with the zeroes of $z^r - S(z)$, and therefore some of the unknown will be determined. Unfortunately the number of zeroes $z^r - S(z)$ cannot be determined in general. Another generating function

$$H(z,w) = \sum_{j=0}^{\infty} H_j(z)\, w^j \quad \text{(D-7)}$$

is introduced to ensure that the number of zeroes can be determined exactly. If a steady-state generating function for the buffer contents exists, $\lim_{j\to\infty} H_j(z) = H(z)$, $H(z)$ is given by

$$H(z) = \lim_{w\to 1^-} (1 - w)\, H(z,w) . \quad \text{(D-8)}$$

Proof:

$$(1-w)\,H(z,w) = \sum_{j=0}^{\infty} H_j(z)\,w^j - \sum_{j=0}^{\infty} H_j(z)\,w^{j+1}$$

$$= H_o(z) + \sum_{j=1}^{\infty} \left(H_j(z) - H_{j-1}(z)\right) w^j .$$

Note that the sum $H_o(z) + \sum_{j=1}^{\infty} \left(H_j(z) - H_{j-1}(z)\right)$ has a partial sum $S_N = H_o(z) + \sum_{j=1}^{N} \left(H_j(z) - H_{j-1}(z)\right) = H_N(z)$, which converges to $H(z)$ as $N \to \infty$. Therefore, by taking the limit as $w \to 1^-$, Eq. D-8 is obtained, which is the generating function we want.

The equation for $H(z,w)$ is obtained by multiplying Eq. D-6 by w^{j+1} and summing over j:

$$H(z,w) = \frac{z^r\,H_o(z) + w\,S(z)\sum_{k=0}^{r-1} (z^r - z^k)\,A_k(w)}{z^r - w\,S(z)} , \qquad \text{(D-9)}$$

where $A_k(w) = \sum_{j=0}^{\infty} w^j\,P(L_j=k)$, $k = 0, \ldots, r-1$, are the unknowns and have to be determined. The denominator has exactly r zeroes within the contour $|z| = 1$ for any $|w| < 1$ by Rouché's theorem (Ref. 2):

> Rouché's Theorem: If $f(z)$ and $g(z)$ are analytic inside and on a closed contour C and $|g(z)| < |f(z)|$ on C, then $f(z)$ and $f(z) + g(z)$ have the same number of zeroes inside C.

In the previous case $z^r - S(z)$, the functions are equal on the contour $|z| = 1$. Further, on another contour $|z_o|$ where $|z_o| > 1$, the inequality need not be satisfied, but with

the introduction of w the conditions of the theorem are satisfied for any $|w| < 1$. Therefore, for any $|w| < 1$, the equation

$$z^r - w\,S(z) = 0$$

has exactly r complex zeroes in the interior of the contour $|z| = 1$. The zeroes are denoted as $\theta_1(w), \theta_2(w), \ldots, \theta_r(w)$. It is shown in Ref. 3 that, for some real $w_o (0 \le w_o < 1)$ and all real w $(w_o \le w < 1)$, the zeroes $\theta_i(w)$, $i = 1, \ldots, r$, are distinct functions of w with limits as $w \to 1^-$. The limits of the zeroes are given by

$$\lim_{w \to 1^-} \theta_i(w) = \theta_i$$

and one and only one zero tends to unity as $w \to 1^-$. Reorder the zeroes so that we obtain $\theta_1 = 1$ and $\theta_i \neq 1$, $i = 2, \ldots, r$. Since $H(z,w)$ is analytic within $|z|$ and $|w| < 1$, the numerator of Eq. D-9 must be zero at the zeroes of the denominator, i.e.,

$$H_o\left(\theta_j(w)\right) = \sum_{k=o}^{r-1} \left(\theta_j^k(w) - \theta_j^r(w)\right) A_k(w)$$

$$\text{for} \qquad j = 1, \ldots, r$$

since $\theta_j^r(w) = w\,S\left(\theta_j(w)\right)$. The solutions to this system of equations are presented in Ref. 1 with

$$\sum_{k=o}^{r-1} \left(z^r - z^k\right) A_k(w) = (z - 1) \sum_{j=1}^{r} \left[\prod_{\substack{m=1 \\ m \neq j}}^{r} \frac{z - \theta_m(w)}{\theta_j(w) - \theta_m(w)}\right] \frac{H_o\left(\theta_j(w)\right)}{1 - \theta_j(w)} .$$

Therefore, $H(z,w)$ is given by

$$H(z,w) = \frac{z^r H_o(z) + w(z-1) S(z) \sum_{j=1}^{r} \frac{H_o\left(\theta_j(w)\right)}{1 - \theta_j(w)} \prod_{\substack{k=1 \\ k\neq j}}^{r} \frac{z - \theta_k(w)}{\theta_j(w) - \theta_k(w)}}{z^r - w\, S(z)} \qquad \text{(D-10)}$$

Finally multiplying Eq. D-10 by (1 - w) and taking the limit on $w \to 1^-$, the generating function of the buffer contents in steady state is obtained and is equal to

$$H(z) = \frac{(r - \mu_Y)(z - 1)\, S(z)}{z^r - S(z)} \prod_{k=2}^{r} \frac{z - \theta_k}{1 - \theta_k}, \qquad \text{(D-11)}$$

where $\mu_Y = E\{Y\}$. The condition for the existence of the steady-state solution is $\mu_Y < r$; that is, the average arrivals in a cycle must be smaller than the maximum number of data units that can be transmitted. The expected value and variance of the buffer contents in steady state, obtained from the first and second derivatives of H(z) evaluated at z = 1, are given by the following equations:

$$E\{L\} = \frac{1}{2} \frac{\sigma_Y^2}{r - \mu_Y} + \frac{1}{2} \mu_Y + \frac{1}{2} \sum_{k=2}^{r} \frac{1 + \theta_k}{1 - \theta_k} \qquad \text{(D-12)}$$

$$\text{Var}\{L\} = \sigma_Y^2 + \frac{1}{12} + \frac{\mu_{3_Y} - r^3}{3(r - \mu_Y)} + \left[\frac{\sigma_Y^2 - (r^2 - \mu_Y^2)}{2(r - \mu_Y)}\right]^2$$

$$+ \sum_{k=2}^{r} \frac{-\theta_k}{(1 - \theta_k)^2}, \qquad \text{(D-13)}$$

where $\sigma_Y^2 = \text{Var}\{Y\}$ and $\mu_{3_Y} = E\{Y^3\}$.

With contiguous allocations, the length of every cycle is M, and therefore $M_j = M$ with probability one. With this value for the cycle length, the generating function of Y, the total arrivals in the cycle are equal to

$$S(z) = P^M(z),$$

resulting in $\mu_Y = M\mu_x$ and $\sigma_Y^2 = M\sigma_x^2$. The zeroes are now solution of

$$z^r - wP^M(z) = 0. \qquad \text{(D-14)}$$

Equations IV-3, IV-4, and IV-5 of Chapter IV are obtained by substituting these values of the average value and variance of Y into Eqs. D-12, D-13, and D-14, respectively.

For distributed allocations where the cycle lengths are random with possible values $m_1, m_2, \ldots, m_N$ with probabilities $P\{M_j = m_i\} = p_i$, the average value of the total arrivals in a cycle $\mu_Y = \bar{M}\mu$, where $\bar{M} = \sum_{i=1}^{N} m_i p_i$, is the average value of the length, and $\sigma_Y^2 = \bar{M}\sigma_x^2 + \mu^2\sigma_M^2$, where σ_M^2 is the variance of the cycle length. Equation IV-12 is obtained by substituting these values into Eq. D-12 and setting r = 1. The Eqs. D-11, D-12, and D-13 describe the behavior of distributed allocation when more than one data unit is allocated at an access.

We will now proceed to develop an analytic expression for the zeroes of Eq. D-14 with w real and $0 < w < 1$. Let us first define an r^{th} root of a function f(z) which is analytic for $|z| < 1$; g(z) is said to be an r^{th} root of f(z) if g(z) is analytic for $|z| < 1$ and $g^r(z) = f(z)$. One r^{th} root of $P^M(z)$ is $P^{M/r}(z)$. Consider the following set of equations:

$$z - w^{1/r} P^{M/r}(z) \exp\left\{\frac{i2\pi(j-1)}{r}\right\} = 0 \qquad (D\text{-}15)$$

$$j = 1, \ldots, r$$

Each of the Eqs. D-15 has exactly one zero by Rouché's theorem for $w < 1$. Any solution of one Eq. D-15, say $z_k(w)$, which is a solution for $j = k$, is also a solution of Eq. D-14. Further, it is easily seen that the solution of k equation is not a solution of the n^{th} equation for $k \neq n$. Therefore, the set of Eqs. D-15 yields r distinct zeroes, each of which is also a solution of D-14 which has exactly r zeroes. Hence the r zeroes of Eq. D-14 can be generated from the solutions of Eq. D-15. The solutions of the set of Eqs. D-15 are obtained by using the power series expansion (Ref. 4, page 133):

$$\theta_j(w) = \sum_{n=1}^{\infty} \frac{w^{n/r}}{n} e^{-n\phi_j} \sum_{k=1}^{n-1} \binom{n-2}{k-1} \frac{q^{n-1-k}}{k!} \left(\frac{n\lambda(1-q)}{r}\right)^k \qquad (D\text{-}16)$$

$$j=1,\ldots,r$$

where $\phi_j = \frac{M\lambda}{r} + \frac{2\pi i(j-1)}{r}$. The final set of equations is obtained by taking the limit as $w \to 1^-$. The zeroes are already arranged with $\theta_1 = 1$ and $\theta_j \neq 1$, $j = 2, \ldots, r$.

B. VIRTUAL DELAY

The cycle with respect to the i^{th} user is initiated at the beginning of his allocated transmission duration. The virtual message, consisting of m data units, arrives randomly in the j^{th} cycle and joins the queue at the i^{th} user. The virtual delay for this customer is the difference between the queueing time (total time in the system) and the message transmission time. Delay is due to two sources:

1. Data units already buffered at the terminal before the arrival of the virtual message

2. The periodic availability of the channel to the i^{th} user.

1. Contiguous Allocations

The virtual message (m data units) is assumed to arrive at the i^{th} user at the start of the α^{th} slot in the $(j+1)^{st}$ cycle, with α a random variable uniformly distributed in the interval [0,M). The queueing time consists of:

1. Waiting $M - \alpha$ slots for the start of the next transmission allocation to the i^{th} terminal.

2. Waiting the number of the slots needed to transmit the already buffered data units. The contents of the buffer at the virtual arrival time are

$$Q_j^{(i)} = (L_{jM+k}^{(i)} - r_i)^+ + \sum_{n=1}^{\alpha} X_{jM+k+n-1}^{(i)}.$$

Removing r_i data units per cycle, it will take

$(M-r_i)\left[\frac{Q_j^{(i)}}{r_i}\right] + Q_j^{(i)}$ slots to transmit these buffered data units with $\lceil a \rceil$ equal to the largest integer contained in a. It takes $Q_j^{(i)}$ slots to transmit $Q_j^{(i)}$ data units, but, due to the periodic availability of the channel, r_i data units are removed and the i^{th} user must wait $M-r_i$ slots before removing another r_i data units, etc. The time spent in waiting for service is given by the first term, and the second term is the time spent in transmitting. At this point the channel is still available to the user to begin transmitting the virtual message.

3. The slots needed to transmit the virtual message. This requires $m + (M-r_i)\left[\frac{m+\beta}{r_i}\right]_-$ slots to complete service to the m virtual data units.

$$\beta = Q_j^{(i)} - r_i\left[\frac{Q_j^{(i)}}{r_i}\right],$$

and $\left[a\right]_-$ is equal to the largest integer strictly less than a. The expression is obtained in a similar manner to that in item 2, but service is completed when the last data unit is transmitted, and we do not have to wait for the channel to be made available again.

The delay, as a random variable, is equal to the queueing time minus the actual service time and is given by

$$M-\alpha + (M-r_i)\left[\frac{Q_j^{(i)}}{r_i}\right] + Q_j^{(i)} + (M-r_i)\left[\frac{m+\beta}{r_i}\right]_- .$$

The stationary expected delay is obtained by averaging over the arrival times ($0 \le \alpha < M$), taking the expectation and letting $j \to \infty$. Unfortunately the delay equation contains integer values of functions of $Q_j^{(i)}$, and hence upper and lower bounds are developed for the stationary delay. By noting that

$$\left[\frac{m+\beta}{r_i}\right]_- \le \frac{m+\beta}{r_i}$$

and averaging over the arrival rate, taking expectation and the limit as $j \to \infty$, the stationary expected delay is bounded above by

$$D_m^{(i)} \leq D_u^{(i)} = \frac{M+1}{2} + \frac{M}{r_i}\left\{ E\left\{ L^{(i)*} \right\} - \frac{M+1}{2}\mu_i \right\} + \left(\frac{M}{r_i} - 1\right) m \cdot \quad \text{(D-17)}$$

For the lower bound, by using

$$\left[\frac{m+\beta}{r_i}\right]_- \geq \left(\frac{m+\beta}{r_i} - 1\right)^+ \geq \left(\frac{m-r_i}{r_i}\right)^+ + \frac{\beta}{r_i}$$

with the function $(a)^+$ equal to the maximum $\{o, a\}$, the stationary expected delay is bounded from below by

$$D_m^{(i)} \geq D_\ell^{(i)} = \frac{M+1}{2} + \frac{M}{r_i}\,\bar{Q}^{(i)} + \left(\frac{M}{r_i} - 1\right)\left(m - r_i\right)^+, \quad \text{(D-18)}$$

where $\bar{Q}^{(i)} = E\left\{ L^{(i)*} \right\} - \frac{M+1}{2}\mu_i$ is the average buffer contents at the arrival time.

Another lower bound can be obtained by assuming that the buffer is empty at the arrival of the virtual message. This lower bound is the absolute minimum delay time. With an empty buffer, the queueing time consists of

- Waiting $M-\alpha$ slots for the next service
- The number of slots needed to complete service to the m virtual data units

$$m + (M-r_i)\left[\frac{m}{r_i}\right]_- .$$

Averaging over the arrival times, we obtain

$$D_\ell^{(i)^1} = \frac{M+1}{2} + (M-r_i)\left[\frac{m}{r_i}\right]_- \quad \text{(D-19)}$$

The difference between the lower bounds (i.e., Eqs. D-18 and D-19) will define the regions where the bounds are better:

$$D_\ell^{(i)} - D_\ell^{(i)^1} = \frac{M}{r_i}\,\bar{Q}^{(i)} + (M-r_i)\left\{\left(\frac{m}{r_i} - 1\right)^+ - \left[\frac{m}{r_i}\right]_-\right\} \quad \text{(D-20)}$$

The bracketed term is negative except when m is an integer multiple of r_i, in which case the term is zero. If Eq. D-20 is positive, depending upon the arrival statistics, the lower bound of Eq. D-18 is a better bound. While in the region where Eq. D-20 is negative, Eq. D-19 yields the better bound.

For the case of allocating one data slot (i.e., $r_i = 1$), the virtual delay can be exactly determined, because the terms of which the integer values were to be taken are integers. The virtual delay with $r_i = 1$ is equal to the lower bound, Eq. D-18, with r_i set to unity.

2. Distributed Allocations

The virtual delay is derived only for an allocation of one data slot at each access. The virtual message, consisting of m data units, is assumed to arrive at the start of the α^{th} slot ($0 \le \alpha < M_j$) of the j^{th} cycle. The queueing time is composed of waiting

- $M_j - \alpha$ slots for the next transmission slot
- The number of slots to completely transmit the buffered data units. The buffer contents at the arrival time of the virtual message are

$$Q = \left(L_{j-1} - 1\right)^+ + \sum_{i=1}^{\alpha} X_i,$$

and these data units are transmitted in $\sum_{k=a}^{b} M_k$

slots, where $a = j + 1$ and $b = a + Q - 1$

- The slots needed to transmit the virtual message.

It subsequently requires

$$1 + \sum_{k=c}^{d} M_k \text{ with } c = Q + j + 1 \text{ and } d = c + m - 2$$

to transmit m data units.

The virtual delay is the queueing time minus the slots actually used to transmit the m data units (m slots). By taking the averages over the arrival slot, the cycle length, and the arrival process, and taking the limit $j \to \infty$, the virtual delay is given by

$$D_m = \frac{\overline{M} + 1}{2} + \overline{M} \left\{ E\left\{ L^* \right\} - \frac{\overline{M} + 1}{2} \mu \right\} + (m-1)\ (\overline{M} - 1), \quad \text{(D-21)}$$

where $\overline{M}$ is the average terminal cycle length.

REFERENCES, APPENDIX D

1. A. G. Konheim and B. Meister, "Service in a Loop System," *Journal of the Association for Computing Machinery*, Vol. *19*, No. 1, January 1972.

2. E. C. Titchmarsh, *The Theory of Functions*, Oxford University Press, 1939.

3. P. E. Boudreau, J. S. Griffin, Jr., and M. Kac, "An Elementary Queueing Problem," *American Mathematics Monthly*, Vol. *69-7*, September 1962.

4. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, 1946.

APPENDIX E

ACCESS WITH RESERVATIONS

APPENDIX E

ACCESS WITH RESERVATIONS

An approximate solution for the average waiting time and average buffer contents for the reservation technique is presented in this appendix. The solutions obtained provide a better approximation to the effects of collisions (blocking) of the reservation or request and allocation messages transmitted on the orderwire channel than the zero blocking approximation of Chapter IV.

The queueing model for reservations separates into two distinct parts: (1) the hypothetical common queue from which messages are allocated transmission time on the data channel(s) and (2) the waiting room, which consists of messages attempting to join the common queue via the orderwire channel. Because of the separation, the queueing parameters associated with both parts are obtained separately from previous derived results under appropriate assumptions and then are combined to obtain the queueing parameters for access with reservations. Under the assumption that the message arrival process at the common queue is Poisson, the desired queueing parameters are given by the available results from queueing theory (e.g., the M/G/1 queue results for geometrically distributed message lengths). Although the Poisson assumption is an approximation, the resultant analytical queueing results agree with the results obtained by simulations (Ref. 1). The input process on the orderwire channel is Poisson with constant-length "control" messages. Therefore, if the access

technique is random access, the results of ALOHA-type systems are applicable with modifications to the delay (the turnaround time) and the contents of the waiting room. The remainder of this appendix is devoted to developing the queueing parameters for the waiting room.

The slotted-ALOHA random-access scheme is used on the orderwire because saturation in a slotted system occurs at a higher utilization than in a purely random system. In a slotted orderwire system, the channel is subdivided into "control" time slots which are equal to the "control" message transmission time. The users are required to synchronize the transmissions with the beginnings of the control slots and are provided with a reference time for this purpose. The slotted system ensures that when transmissions overlap, the overlap is complete and not partial, as in a pure ALOHA system. It is assumed that if transmissions overlap, none of the messages are received correctly and the affected users try again. To prevent a reoccurrence, the messages are randomly delayed before retransmission is attempted. The delay is uniformly distributed over K slots after the unsuccessful transmission is detected.

The time a message spends in the waiting room is the time to effect a capacity assignment, which is the turnaround time. To determine the turnaround time, it is necessary to obtain the "waiting" time for a control message, which is the time to its successful reception measured from the generation time. The "waiting" time includes the time between the arrival and the next transmission slot, the retransmission delay, the satellite propagation time delay, and the control message transmission time. The average waiting time (in control slots) for the i^{th} user is given by

$$W_c^{(i)} = 1.5 + D_p + E_i\left(D_p + \frac{K+1}{2}\right), \qquad \text{(E-1)}$$

where D_p is the propagation time delay in control slots, and E_i is the average number of retransmissions required to obtain a successful transmission. This waiting time neglects the effect of the generation of more than one control message during a control slot at a user. This is a reasonable assumption if the average interarrival time between messages is large in comparison to a control slot.

The average number of retransmissions is determined in the following. The channel input (the control-message arrival process) from a user is Poisson distributed with an average rate of S_i requests/control slot for the i^{th} user. The channel traffic generated by a user consists of the retransmission of previously blocked messages and the transmission of new control messages and is assumed also to be Poisson distributed with rate G_i ($G_i \geq S_i$) for the i^{th} user. The Poisson assumption for the channel traffic is an approximation but has been verified by simulation for K large (Ref. 1). Under stable conditions, the channel throughput rates are equal to the channel input rates, while under unstable conditions, the throughput rates are smaller than the input rates. The average number of transmissions ($1 + E_i$) required for a control message from the i^{th} user under stable conditions is given by (Ref. 1)

$$1 + E_i = \frac{G_i}{S_i} . \qquad \text{(E-2)}$$

Further, again under stable conditions, the traffic rates are related (Ref. 2) to the channel input rates by

$$S_i = G_i \prod_{\substack{j=1 \\ j \neq i}}^{M} (1-G_j) \qquad i = 1, 2, \ldots, M, \qquad \text{(E-3)}$$

where M is the size of the net including the controller. The channel traffic rates are obtained by solving this set of equations. The "waiting" time results compare favorably with the simulation results presented in Ref. 1 for the delay in a slotted-ALOHA data-transmission system with identical users.

The turnaround time is also dependent upon the type of control (i.e., distributed or centralized) that is used in the net. With distributed control, each user in the net forms the hypothetical buffer, and therefore the turnaround time is equal to the average "waiting" time for the reservation message (Eq. E-1). The reservation message is generated by the arrival of a data message, and therefore the channel input rate is equal to

$$S_i = B_1/T_i R_1,$$

where T_i is the average interarrival rate (in seconds) of the data messages, B_1 is the length in bits of a reservation message, and R_1 is the capacity (transmission rate) of the order-wire channel. If only one class of users exists (identical message processes), the traffic-input equation is simplified and is given by

$$S = G\left(1 - \frac{G}{N}\right)^{N-1},$$

where S is the total channel input rate of the net, and N is the number of users in the net.

For the case of centralized control, the controller forms the hypothetical common queue from the user request messages and allocates capacity to the users by an allocation message.

The channel input from the controller is equal to the request message and throughput, and, under stable conditions, the rate is S allocations/control slot. The allocation message length is equal to the request message length B_1. The waiting time for a successful reception of an allocation message is

$$1 + D_p + E_{N+1}\left(D_p + \frac{K+1}{1}\right),$$

where $1 + E_{N+1} = \frac{G_{N+1}}{S}$. The turnaround time is the sum of the average waiting times for request and allocation messages. The equations for the turnaround time b_i are presented in Table E-1. Finally, the contents of the waiting room (in data units) is $b_i S_i \bar{m}$ by Little's theorem, where $\bar{m}$ is the average number of data units in a data message.

TABLE E-1. TURNAROUND TIME

Type of Control	Turnaround Time b_1, control slots
Distributed	$1.5 + D_p + E_i\left(D_p + \frac{K+1}{2}\right)$
Centralized	$2.5 + 2D_p + \left(E_i + E_{N+1}\right)\left(D_p + \frac{K+1}{2}\right)$

Before presentation of numerical results, saturation and its effects are discussed. The orderwire channel using slotted ALOHA saturates at a channel utilization of 0.368, and when the orderwire channel saturates, allocation and reservation messages are blocked, and hence the data throughput also becomes blocked (decreases with time and becomes zero). The system may also become unstable even when the utilization is less than 0.368 because of statistical fluctuations. Both of these facts imply that a dynamic control of the traffic on

the orderwire may be required (possibly on the value of the retransmission delay or a suppression of the input). The orderwire capacity should be carefully selected to ensure that the resultant orderwire utilization is smaller than the saturation value (possibly no larger than 0.25, for example) to try to minimize the danger of instability. Other accessing techniques are also possible for the orderwire channel. For example, fixed assignment is attractive because of its inherent stability even though the delays may be increased. The resultant delays with a fixed-assignment orderwire are left for future work.

Numerical results for the parameter values of Table E-2 are presented in Table E-3 for the turnaround times and in Table E-4 for the buffer contents and queueing times with the average data message length of 26 data units. In Chapter IV, the turnaround time was assumed to be 3 data slots or 0.76 second, which, from Table E-3, is underestimated. The results of Table E-4 demonstrate that the queueing parameters are more sensitive to the reduction in the data channel capacity than to the turnaround time. Therefore, the capacity allocations to the orderwire and data channels must be properly selected. Table E-5 presents the results for the case of pooling four nets and dynamically allocating capacity to the users on three channels, with one channel reserved for the orderwire. In this case, the orderwire capacity is seen to be excessive.

TABLE E-2. ORDERWIRE PARAMETER VALUES

Control:	Distributed
Number of Users:	11 (one class)
Request Message Length:	304 bits
Retransmission Delay (K):	10 control slots

TABLE E-3. TURNAROUND TIME AND AVERAGE NUMBER OF RETRANSMISSIONS

Interarrival Rate, minutes	Average Number of Retransmissions		Turnaround Time, seconds	
	A[a]	B[b]	A[a]	B[b]
27	0.01	0.003	3.9	1.3
10	0.03	0.01	4.1	1.3
2	0.16	0.05	5.4	1.4
1.34	0.28	0.11	6.5	1.6

[a] A: R_1 = 200 b/s.

[b] B: R_1 = 600 b/s.

TABLE E-4. EFFECTS OF ORDERWIRE AND DATA CHANNEL CAPACITIES ON THE BUFFER CONTENTS AND QUEUEING TIME

Interarrival Rate, minutes	Buffer Contents, data units			Queuing Time, minutes		
	A[a]	B[b]	C[c]	A[a]	B[b]	C[c]
27	0.2	0.2	0.1	0.19	0.18	0.13
10	0.5	0.5	0.4	0.21	0.20	0.14
2	5.6	9.9	3.7	0.44	0.76	0.29
1.34	127.3	∞[d]	21.2	6.57	∞[d]	1.09

[a]A: R_1 = 200 b/s, data rate (R_2) = 2.2 kb/s.

[b]B: R_1 = 600 b/s, R_2 = 1.8 kb/s.

[c]C: R_2 = 2.4 kb/s, turnaround time = 3 data slots.

[d]Saturation on data channel.

TABLE E-5. BUFFER CONTENTS AND QUEUEING TIMES FOR THE CASE OF 44 USERS TIME-SHARING THREE 2.4-kb/s CHANNELS WITH A 2.4-kb/s ORDERWIRE AND CENTRALIZED CONTROL

Interarrival Rate, minutes	Buffer Contents, data units		Queueing Times, minutes	
27	0.1	(0.1)[a]	0.13	(0.12)
10	0.3	(0.3)	0.13	(0.12)
2	3.3	(3.2)	0.25	(0.25)
1.34	∞[b]	(∞)	∞	(∞)

[a]Zero blocking assumption, turnaround time = 3 data slots.

[b]Saturation on the data channel.

REFERENCES, APPENDIX E

1. University of California, Los Angeles, *Packet Switching in a Multi-Access Broadcast Channel with Application to Satellite Communication in a Computer Network*, Technical Report UCLA-ENG-7429, S. S. Lam, April 1974.

2. N. Abramson, "Packet Switching with Satellites," *Proc. National Computer Conference*, 1973. (Also in Ref. 3)

3. W. W. Chu, *Advances in Computer Communications*, Artech House, Incorporated, Dedham, Massachusetts, 1974.