Columbia University in the City of New York
DEPARTMENT OF ELECTRICAL ENGINEERING
NEW YORK, N.Y. 10027

1. Report No. CR-167903	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

NEXT GENERATION COMMUNICATIONS SATELLITES: MULTIPLE ACCESS AND NETWORK STUDIES

Second Year Final Report

5. Report Date: September 1982

6. Performing Organization Code

7. Author(s): H.E. Meadows, M. Schwartz, T.E. Stern, S. Ganguly, B. Kraimeche, K. Matsuo, I. Gopal

8. Performing Organization Report No.

10. Work Unit No.

9. Performing Organization Name and Address

Department of Electrical Engineering
Columbia University
New York, N.Y. 10027

11. Contract or Grant No.: NAS 3-22464

13. Type of Report and Period Covered: Contractor Report

12. Sponsoring Agency Name and Address

National Aeronautics and Space Agency
NASA Lewis Research Center
Cleveland, Ohio 44135

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report deals with the second phase of a two year study program dealing with efficient resource allocation and network design for satellite systems serving heterogeneous user populations with large numbers of small "direct-to-user" earth stations. The work focuses on TDMA systems involving a high degree of frequency reuse by means of satellite-switched multiple beams (SSMB) with varying degrees of onboard processing.

The overall objective of the second phase was to develop and evaluate algorithms for the efficient utilization of the satellite resources under reasonably realistic constraints. This report deals with the following issues:

1) Effect of skewed traffic, overlapping beams and batched arrivals in packet-switched SSMB systems. (Results obtained via simulation)

2) Integration of stream and bursty traffic. (Results obtained using approximate analytical models.)

3) Optimal circuit scheduling in SSMB systems: Performance Bounds and Computational Complexity. (Optimal algorithms and suboptimal heuristics are presented.)

17. Key Words (Suggested by Author(s))

Communications Satellites
Multiple Access
Satellite-Switched/Time Division Multiple Access
Skewed traffic
Integrated Services

18. Distribution Statement

Unclassified-Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified		

* For sale by the National Technical Information Service, Springfield, Virginia 22161

TABLE OF CONTENTS

Symbol List by Chapters

Symbol List, Chapter 2

Symbol	Definition
N_T	Number of Transponders
N_Z	Number of zones
D_i	Data packet
R_i	Request of data packet on order wire
T_i	Trigger signal
τ	Time interval of one segment
T	Time interval of a single run
m	Number of segments
μ	Mean of population
$\overline{W}^{(j)}$	Sample mean of j^{th} segment
S_W	Sample variance
$\overline{W}$	Sample mean
$1-\alpha$	Confidence coefficient or confidence level
ρ	Traffic intensity (packets/slot/transponder)
$\overline{Q}$	Average queue length (packets)
$\overline{W}$	Mean packet waiting time (slots)
Q_{max}	Maximum queue length (packets)
N_{max}	Maximum number of Transponders used
$\overline{N}$	Average number of Transponders used
P_{ij}	Probability [source zone i to destination zone j]

Symbol List, Chapter 3

Symbol	Definition
CS	Circuit-Switched
PS	Packet-Switched
FFFB	Fixed-frame, fixed-boundary
FFMB	Fixed-frame, movable-boundary
M	Number of earth stations per zone
N_z	Number of zones
TDMA	Time Division Multiple Access
N_T	Number of transponders
λ_1	Rate of CS arrival process per station (Poisson)
λ_2	Rate of PS arrival process per station (Poisson)
$\frac{1}{\mu_1}$	Average circuit holding time (exponential)
$\frac{1}{\mu_2}$	Average packet service time (exponential)
F	Number of slots per TDM frame
N_S	Number of slots in the CS portion of the frame
N_B	Number of slots in the PS portion of the frame
S	Number (random) of unused CS slots
$Q_S=i$	Number of circuits in the system
$Q_B=j$	Number of packets in the system
P_{ij}	joint CS/PS probabilities
B	packet buffer size
P_o	Probability of buffer overflow
$P_i(z)$	Conditional generating function of the number of packets in buffer
ρ_1	CS traffic intensity, $\rho_1=\frac{\lambda_1}{\mu_1}$

Symbol	Definition
Δ	Slot duration (= packet transmission time)
ρ_2	PS traffic intensity $\rho = \lambda_2/\mu_2$
σ	Average proportion of the frame available for PS transmission
ε	Average number of circuits in packet
$Q(z)$	z-transform of the PS arrival process
$L, \frac{T}{\Delta}$	Average number of packets in the buffer and normalized delay under the randomized M/D/1 model.
τ	Frame duration (in seconds)
T_1, T_2, T	Parameters used for delay analysis in the approximate model (b)
γ	Effective boundary in the FFMB
C	Satellite channel capacity (bps)
N	Number of slots dedicated to a user
U	Number of slots in the unassigned pool
R	Number of shared channels
N_A	Number of slots per shared channel
R_i	bit rate of i-th CS class
α_i	Number of slots allocated to i-th CS class
β	Traffic mix parameter
P	Packet length in bits
$\lambda_2(i), \mu_2(i)$	Input and output rates of system at state i, i=0,1
Π_i	Probability of system being in state i, i=0,1
ρ	Constant system utilization ($=\lambda_2(i)/\mu_2(i)$
EL	Average packet queue
K	Number of priority traffic classes

Symbol	Definition
λ_r	Arrival rate of class r
σ	Pr. (server available)
FIFO	First-In, First-Out discipline
W_r	Waiting time of class r
V_r^t	Virtual waiting time process for class r
Gr, Br, G'r	Random variables used in delay analysis
EU(x), ED(x)	Average number of up and down crossings of level x

Symbol list, Chapter 4

Symbol	Definition
D	Traffic demand matrix
d_{ij}	Element of traffic demand matrix
N	Number of up/down links
R_i	Row sum of D
C_j	Column sum of D
L	Frame size
K	Number of transponders
N_{min}	Minimum number of switchings
N_S	Number of switchings
B_i	Number of subchannels in uplink i
α_j	Number of subchannels in downlink j
M	Number of uplinks
D_i	Member of a feasible decomposition of D

1. INTRODUCTION

1.1 CONTEXT OF THE WORK

This report covers the second year of a two year study program, initiated on July 9, 1979, conducted at Columbia University under NASA support, dealing with problems of multiple access, resource allocation and network design for next generation communication satellite systems. Two basic sets of requirements shaped our effort. First was the recognition that the next decades will see increasing demand for the two limited resources at the disposal of the satellite system designer: suitable positions in geostationary orbit, and available radio frequency spectrum. Second was the fact that the user populations for these systems will be highly heterogeneous both in their statistical characteristics and their service requirements. Thus the general thrust of the work has been to attempt to identify the problems associated with the allocation of scarce communication resources in as efficient a manner as possible while satisfying the requirements of a heterogeneous and continually varying community of users. Particular attention has been directed toward systems involving large numbers of small "direct to user" earth stations.

During the two year effort, we focussed on TDMA systems involving a high degree of frequency reuse by means of satellite switched multiple beams with varying degrees of onboard processing. The need for a system that can provide integrated service to a disparate user group ranging from interactive computer terminals to television channels suggests that a highly sophisticated set of multiplexing, multiple access and switching procedures will be required if user needs are to be met economically while using the available bandwidth as efficiently as possible. Therefore, in the first year of the study we examined a variety of communication procedures, choosing those which showed promise of achieving these objectives, even though they might appear to be highly complex by today's operational standards.

The goal of the first year of the project was to develop and analyze the performance of multiple access techniques for satellite switched multiple beam (SSMB) systems operating in a packet switched mode. The multiple access problem was viewed at two levels - the station access problem and the switch assignment problem. Station access was considered on a fixed (FA), demand (DA), and random access (RA) basis. For the switch assignment problem, fixed (FS), demand (DS), and store-and-forward (SF) switching was considered. Six access protocols were selected for study, and delay-throughput analysis was performed for each. One of the protocols, a demand-access/demand-switched (DA/DS) "list scheduling" scheme, showed promise in a number of contexts, and was selected for further study in the second phase of our work.

1.2. PROBLEMS ADDRESSED AND APPROACH USED

As is typical in any detailed analytical study, our first phase work raised at least as many questions as it answered. In the second year we addressed some of these outstanding issues. A major goal of this second phase was to examine performance questions in a more realistic traffic environment, using more realistic system configurations. In particular, we dealt with the following issues.

1) Effect of Skewed Traffic, Overlapping Beams and Batched Arrivals in Packet-Switched SSMB Systems.

2) Integration of Stream and Bursty Traffic.

3) Optimal Circuit Scheduling in SSMB Systems. Performance Bounds and Computational Complexity.

The overall objective was to develop and evaluate algorithms for the efficient utilization of the satellite resources under reasonably realistic constraints.

1.3 SUMMARY OF RESULTS

Simultaneous studies were undertaken to deal with the three problems mentioned above. The results of these three studies are reported in Chapters 2,3 and 4 respectively. We summarize them in this section.

1.3.1 Effects of Skewed Traffic, Beam Overlap, Batched Arrivals

Having observed from our first phase studies that the DA/DS "list scheduling" packet-switching protocol performed well in simple configurations, it was chosen as a basis for further studies in more realistic environments. Three series of simulation studies were undertaken. (Due to the complexity of the system configurations analysis was out of the question.) A description of the protocol, the simulation approach and the techniques of validation are presented in Section 2.1, with details and interpretation of results appearing in Section 2.2. The system parameters that were varied were:

Number of zones - N_z

Traffic skew ratio - S

Message length - { S, single packet
M, mixed single and multipacket

Overlap probability - { High
Low

In each case the number of transponders was fixed at $N_T = 5$.

The quantities of interest for performance evaluation were

Average packet queue length, $\bar{Q}$

Max queue length, Q_{max}

Average packet waiting time $\bar{W}$

Average number of transponders used, $\bar{N}$

Max number of transponders used, N_{ms};

Each of these quantities was studied as a function of traffic intensity, ρ. The results of these runs are presented in a series of tables and figures in Section 2.2.

The first study dealt with a small system ($N_Z = N_T = 5$), with no zone overlap and single packet messages, the objective being to study the effect of traffic skew in isolation.

For ease of reference, we list below the cases studied as well as the table and figure numbers where the results can be found.

Case No.	Skew ratio	Table	Figs.
1	1	2.5	2.22-2.37
2	2	2.5	
3	4	2.6	

In the second study, we increased N_z to 20, keeping $N_T = 5$. In this way we were able to compare systems with $N_z = N_T$ (the aforementioned study) to those with "N_z N_T". Basically, we were moving from a system which was "switch-limited" to one which was transponder-limited". Here, we considered the combined effects of skew and variable message lengths.

Case No.	Skew ratio	Message length	Table	Figs.
1	5	M	2.7	
2	1	M	2.8	2.41-2.47
3	5	S	2.9	2.59-2.60
4	1	S	2.10	

In the third study, beam overlap was considered in the context of a system with high traffic skew (S = 5), with $N_z = 20$, $N_T = 5$. High or low overlap probability was created by skewing the traffic matrix more toward the overlapping or non-overlapping beams. (See Figs. 2.39-2.40)

Case No.	Overlap	Message length	Table	Figs.
1	High	M	2.14	
2	Low	M	2.13	2.48-2.58
3	High	S	2.12	2.61-2.62
4	Low	S	2.11	

As expected, the packet queue lengths and average delay increased with traffic skew, beam overlap and message length. However, because of the interaction of the many parameters involved, the exact dependence of these quantities on the system configuration and traffic intensity was rather complex. One important result was that the combination of traffic skew and overlapping zones caused the system to saturate at lower traffic intensities than in systems with uniform traffic and no overlap. The most severe case studied (S = 5 with high overlap probability) resulted in $\rho_{saturation} \approx .55$.

1.3.2 Integration of Stream and Bursty Traffic

In Chapter 3 we present an integrated circuit-switched and packet-switched multiple access scheme to handle combinations of stream and bursty traffic. The access problem is treated using a three-level hierarchical structure: level 1 is the concentration of the various information sources at their earth stations, level 2 is the multiplexing of the competing population of earth stations into an integrated TDMA frame, and level 3 is the allocation of satellite capacity and scheduling of stations to achieve some desired performance.

The level 1 concentration problem has already been studied at length in the context of terrestrial networks. In this study we deal only with levels 2 and 3. An approximate analytic approach for predicting the behavior of the access scheme at level 2 is presented in Section 3.2, and it is used as a basis for performance analysis of the capacity assignment strategies discussed in Section 3.3. Four schemes for assigning circuit-switched (CS) slots are examined and compared:

a) Fixed assignment: each station is allocated a fixed number of slots.

b) Shared pool: some slots are allocated to the stations on a fixed basis, and others are assigned from a common pool on a demand basis.

c) Demand assigned: all slots allocated on a demand basis from a common pool.

d) Shared channels: a refinement of (b).

Superimposed on the circuit-switched assignment is the packet-switched allocation. We stress here the movable boundary approach, in which the remainder of the frame (after CS capacity allocation) is reserved for packets. In addition, any CS slots not currently being used for active circuits are allocated to packets.

In the integrated system, the parameters of interest are circuit blocking probability and packet time delay, which we examine and compare as a function of CS and PS traffic intensity. A numerical example is presented in Section 3.3.2 which illustrates the comparative performance of the four CS allocation schemes. The best scheme from the point of view of blocking probability is (c). On the other hand, since (c) leads to the highest channel occupancy by circuit traffic, it gives the poorest packet delay performance . Thus, there is a trade-off between CS and PS performance. In all cases, the movable boundary scheme offers considerable enhancement of PS performance over fixed boundary schemes.

In concluding this part of our study we also examined certain additional issues associated with on board processing and storage. These are discussed in Section 3.4, and include performance of downlinks with onboard buffering, and packet flow control based on CS slot occupancy.

1.3.3 Optimal Circuit Scheduling

Focussing attention on a purely circuit-switched system we have attempted to draw some general conclusions concerning the feasibility of optimal scheduling in the face of adjacent beam interference, as well as other constraints caused by special system configurations. While this work had not been completed at the conclusion of the contract, we did obtain a number of interesting preliminary results. A complete account of this study, which was carried to conclusion after the termination of the NASA contract, appears in [GOP 82].

The preliminary results, presented in Chapter 4, deal with optimal switching algorithms for systems with single and multi-channel beams and with different uplink/downlink zone configurations. In the case of systems with adjacent beam interference, the scheduling problem is too complex to be solved optimally in an operational system. We have therefore devised some good heuristics for dealing with interference by a combination of proper switch scheduling and the use of two orthogonal polarizations.

REFERENCE

[GOP 82] Gopal, I.S., "Scheduling Algorithms for Multi-Beam Communication Satellites," Ph.D. Dissertation, Department of Electrical Engineeing, Columbia University, May, 1982.

2. DEMAND ASSIGNED/DEMAND SWITCHED PACKET

2.1 SIMULATION: DESCRIPTION AND VALIDATION

2.1.1 Introduction

This section addresses the effects of skewed vs. uniform traffic on packet delay of satellite-switched, time-division multiple-access (SS/TDMA) systems using the demand station access DA and demand switch assignment DS system as shown in Fig. 2.1 . In the DA/DS system* all stations send request messages to the onboard scheduler for transmission slots for arriving data packets. Each request for slot allocation for the data includes information about the desired uplink-downlink connection. The scheduler makes the switch assignment on demand on a slot-by-slot basis, based on the list of requests currently awaiting scheduling, and triggers packet transmission from the stations whose requests have been scheduled. The operation of the DA/DS protocol is illustrated in the timing diagram of Fig. 2.2 . A data packet D_i arrives at a station which transmits a request R_i on the order wire, indicating the desired destination. The scheduler forms a queue of all requests. The request waits on this queue for a time $W_{DA/DS}$ until it is scheduled. A signal T_i is transmitted by the scheduler on the downlink control channel to trigger transmission of the packet. On receiving the trigger the station transmits the packet. The average queuing delay is a function of the algorithm used by the scheduler for switch assignment. Simulation results are presented for the list scheduling algorithm briefly described in the next paragraph. The switch assignments must be made subject to the switching constraints which state that any one uplink zone may not be connected to more than one downlink zone at the same time. Similarly, any one downlink zone may not be connected to more than one uplink zone. Also, the total number of connections at any given time cannot exceed the number of transponders, N_T.

*See [STE 80] for further description of the DA/DS protocol, using the list assignment algorithm for processing requests.

There can be many different algorithms for DS system, e.g., maximum matching scheduling, cyclic permutations scheduling and list scheduling. An efficient algorithm is one which is simple to implement and which also reduces request waiting times. For our simulation we have considered the list scheduling algorithm, which is simple to implement though it may not completely minimize the average waiting time in the request queue. We consider first a simple system in which the number of zones N_Z is equal to the number of transponders $N_T = 5$. The contents of the queue of all pending requests at the controller is shown in Fig. 2.3 in matrix form. Each cell in the matrix corresponds to a pair of source and destination zones; an entry in the cell indicates a request for a packet transmission between those zones. The total number of entries in the matrix indicates the length of the request queue, and the entry number indicates the position of the request in the queue. The switching constraints require that no more than one entry be scheduled in a given time slot from the same row or column of the matrix. In the list scheduling algorithm, the scheduler scans the lists of requests in the queue and schedules them if consistent with switching constraints. The scanning of requests is done on a first-come, first-served basis and terminates if either N_T requests have been scheduled or if all requests have been scanned. For the example of Fig. 2.3 a list scheduling of requests 1,2,8 is illustrated in Fig. 2.4 and Fig. 2.5.

We consider now a system with antenna arranged to form N_Z spot beams. Each beam comprises an uplink and a downlink at different carrier frequencies, with connections made between up and downlinks through a switch matrix in the satellite shown in Fig. 2.6. There are N_T frequency translating transponders, and each connection represents a path through one of them. Messages

arrive randomly at the stations and we assume that the message arrival processes at the stations are independent Poisson process. Each message consists of a single packet requiring one time slot for transmission. The N_Z spot beams can be overlapping or non-overlapping as shown in Fig. 2.7 and 2.8. In this section we consider systems with non-overlapping spot beams.

Messages are assumed to arrive according to Poisson distribution with different mean interarrival time. Traffic can be uniform or non-uniform (skewed) for the zones of interest. The time slot length is assumed to be 400 units of time.

2.1.2 Simulation Technique

Discrete event simulation by digital computer offers many conveniences that make it an attractive tool for analysis of SS/TDMA systems. Simulation is essentially the technique of conducting sampling experiments on a model of the system. [KOB 78, DAN 81, FIS 78, SCH 74]. Thus simulated experiments may be regarded as similar to ordinary physical experiments and should be based on sound statistical techniques. Many simulation programs today are written in one of the existing simulation programming languages rather than in a general purpose programming language such as ALGOL, FORTRAN, PL/1 or APL. Some of the best known languages for discrete event system simulation are GPSS, SIMSCRIPT, SIMULA, CSL & SIMPL/1. Simulation language (1) provide a convenient representation of elements that commonly appear in simulation models, (2) expedite configuration changes of the systems model, (3) provide an internal timing and control mechanism required to execute a simulation run, and (4)

facilitate collection of data and statictics on the aggregate behavior of the simulated system. In our simulation we have used the most frequently used language - GPSS. One of the special features of GPSS is that it describes directly the functional flow of the jobs (called transactions) through the system. The static structure of the system is described by entities such as facilities and storages. Facilities are entities that can be occupied by only one transaction at a time, while storages are entities that may be occupied by more than one transaction simultaneously.

Transactions have numerical characterisitics called parameters. Facilities and storages have numerical attributes. These parameters and attributes are collectively termed standard numerical attributes. The dynamic mechanism of the system is modeled in terms of block commands: Each block command represents an operation that acts on individual transactions or entities. The operation is accomplished by means of field information or a collection of operands that accompany each block statement.

Corresponding to each block command is a block-type subroutine. A GPSS simulation is the interpretive execution of these subroutines. In order to execute simulated events in the correct time sequence, GPSS maintains a clock that records the current simulation clock time. All clock times in the simulator are integer values. The GPSS clock is updated to indicate the time of occurrence of the most imminent event.

2.1.3 Simulation Model

The basic flow chart for simulation is shown in Fig. 2.9. The model is composed of a major segment and two supporting segments shown in Figs. 2.10-2.12. These segments are described below.

Model Segment 1: Taking into consideration the most useful modeling approach, a transaction was used to simulate a packet. First we bring a packet into the model through use of the GENERATE block. The interarrival time is indirectly specified through halfword savevalue UTIL. This makes it possible to vary the traffic rate by redefining the value of UTIL between consecutive runs. Two byte parameters are associated with each transaction (packet), one assigned to a source zone and the other one to a destination zone. The function EXPON represents a continuous, 24 point GPSS function, used to sample from the exponential distribution with an expected value of 1. The function ADDR represents a discrete function for traffic distribution. The arriving packets form a queue. We use a QUEUE block in order to gather statistics. Since the time slot comes at regular discrete intervals of time and the packets can come at any time, we use a LOGIC GATE to let requests be scheduled only when a time slot arrives. The gate opens when a LOGIC SWITCH is set and the LOGIC SWITCH is set when a time slot arrives. Arrival of a time slot is shown in model segment 2. When a time slot arrives, we test for switch assignments of the requests in the queue. We use 11 Byte savevalues for this testing purpose. If a packet cannot be sent, it goes back into the queue with priority ahead of latter packet arrivals and waits for the next time slot. Since there are 5 transponders available, we use the ENTER block to seize storage representing the transponder.

Model Segment 2: In this segment each transaction represents a time slot. As a time slot arrives, LOGIC SWITCH 1 is set, which causes the gate in model segment 1 to open. After a delay of 1 unit of time, the logic switch is reset. Since the assignment is done on a slot by slot basis, we initialize the Byte savevalues at the end of each time slot.

Model Segment 3: This is a timer segment. Since the steady-state behavior is been considered here, statistics obtained during an initialization period have been suppressed. The simulation is run for 2,000,000 units of time.

Table 2.1 summarizes the definitions of GPSS entities used in the simulation, which is described further in Figs. 2.10-2.12.

Table 2.1 DEFINITIONS FOR SIMULATION MODEL

GPSS entity	Interpretation
Transactions	
Model Segment 1	Packets
Model Segment 2	Time slot
Model Segment 3	A timer
Storages	
1	Storage used to simulate switch connection
Queues	
1	The queue used to gather statistics for the requests to be scheduled.
Functions	
EXPON	Fucntion for sampling from exponential distribution with a mean value of 1, used to generate a Poisson arrival process.
ADDR	Function describing the distribution of traffic.
Savevalues	
6	Counter used to record the number of packets which have already been scheduled in a time slot.
1-5, 7-11	Used to store source zone and destination zone addresses.
Variables	
UTIL	Variable describing the mean interarrival time of packets.
Logic Switches	
1	A logic switch simulating the arrival of a time slot.

In order to achieve valid simulation results, it is necessary to assure that steady-state conditions are approached and that the resulting statistics are statistically significant at a certain level of confidence. These considerations are treated below.

2.1.4 Steady State

A system has reached stable or steady state conditions when successive observations of the system performance are statistically indistinguishable. A system whose behavior does not satisfy steady-state conditions is usually described as being in a transient state. [KOB 78, DAN 81, FIS 78] For some kinds of systems no steady-state conditions are expected.

Most simulation models are developed and run under the assumption that there are steady-state operating conditions. If this is the case, transients occur because the simulation run was begun with atypical values for its state variables. Therefore it is necessary first to select starting condition that keep the transient period short and then to minimize the effects of transients that do occur, so that the simulator performance is judged only on its steady-state behavior.

For a single run of a simulator, a simple and basic strategy for setting starting conditions is to begin in an empty and idel status; that is, assume that the system is completely clean of activity and run the simulator until the transient effects are insignificant. It is easy to start the simulator under these conditions, but the transient period is likely to be quite long.

The effects of transients must be removed when performance measures are recorded in a simulation run. The most common method and the one used here is to introduce a nonrecording interval to put the simulator into

steady-state conditions before collecting data. This is done by running the simulator until steady-state conditions are achieved, then clearing all statistical accumulations (but leaving the state of the simulated system as it is) and then continuing the run. The conditions at the end of the transient period become an a priori estimate of the steady-state conditions and, in effect, are used to start a new run. This method is often easier to program than its alternative: to explicitly insert steady-state starting conditions into the model. In GPSS, a RESET control operator facilitates the clearing and restarting process [GPS 71]. This operator, which clears statistical and other output accumulations without changing the model, is used in our program.

To determine whether the model has reached a steady-state condition is generally not a simple matter, because a steady state of a stochastic system or its model must be defined in terms of the probability distribution of the system state (rather than a particular state). That is, the notion of the transient in a stochastic system is different, for instance, from that of a deterministic system (such as an electrical circuit), where the transient behavior is completely characterized by its impulse response function. It should also be recognized that the equilibrium is a limiting condition that may be approached but never attained exactly. There is no single point in simulation time beyond which that system is in equilibrium, but we can choose some reasonable point beyond which we are willing to neglect the error that is made by considering the system to be in equilibrium.

The choice of reasonable starting conditions can diminish the transient interval but cannot eliminate it completely. No simple rules can be given to decide how long an interval should be eliminated. It is advisable to use some pilot runs starting from the idle state to judge how long the initial bias remains. This can be done by plotting the measured statistic against

run length as has been done in Figs. 2.13-2.15. It is highly desirable, however, that the pilot investigation be done by repeating runs. Needless to say, the initial values (often called "seeds") of the pseudo-random generator (used to simulate random arrival of packets and random source destination pairs) should be different in different runs; otherwise exactly the same system output would always be observed, Figs. 2.13-2.15 show how difficult it is to judge from a single run when the measured value has approached its steady state. To minimize simulation cost, not too many runs were made to determine steady state. We have used only one random generator and since run length is proportional to computer cost, we have taken measurements for a certain limited interval of time only instead of making the run length too long.

2.1.5 Confidence Intervals

The use of confidence intervals is an important technique for measuring the degree of uncertainity in making inferences about population characteristics from sample data. The sample mean waiting time in queue $\overline{W}$ is different from the actual population mean μ and we estimate an interval around $\overline{W}$ such that we can predict with some level of confidence that μ falls within that interval. Out of the three main methods [KOB 78], for determining confidence intervals, we have used "the single run method" or "method of batch means". It makes consecutive runs to obtain the confidence intervals shown in Figs. 2.16-2.18, using the final condition of one run as the steady-state starting condition for the next run. That is, one long continuous simulation run is divided into contiguous segments or batches, each of which has the same length. The sample mean $\overline{W}^{(j)}$ of the j^{th} segment is then treated as an individual observation. The method of batch means is shown below. A simulation program was

written and the initial 50,000 time units portion of output data starting form an empty state was discarded for the transient period. Fig. 2.19 shows the effect of simulation time on packet delay. A single run of $T = 2{,}000{,}000$ time units was divided into $m = 20$ segments of equal length $\tau = \frac{T}{m} = 100{,}000$ units.

The observed process is the waiting time in queue $W(t)$ as a function of simulated time t. More precisely, let the waiting time for a packet arriving at $t = t_{ji}$ during the j^{th} segment be denoted $W_i^{(j)}$. The sample mean of this quantity in the j^{th} segment is then

$$\bar{W}^{(j)} = \frac{\sum_{i=1}^{n_j} W_i^{(j)}}{n_j} \qquad (2.1.1)$$

where n_j is the number of packets processed during the j^{th} segment, the overall sample mean of the waiting time is

$$\bar{W} = \frac{\sum_{j=1}^{m} \bar{W}^{(j)}}{m} \qquad (2.1.2)$$

and the sample variance is

$$S_W^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{W}^{(j)} - \bar{W})^2 \qquad (2.1.3)$$

The random variates $\bar{W}^{(j)}$ are approximately normally distributed. Thus, the t-distribution of $m-1=19$ degrees of freedom [KOB 78, BAS 66] is our model for the variable

$$t_{m-1} = \frac{\bar{W} - \mu}{\sqrt{S_W^2/m}} \qquad (2.1.4)$$

where μ is the (unknown) population mean waiting time $E[W(t)]$.

The distribution of the variable t_{m-1} is student's t-distribution with $m-1$ degrees of freedom. Fig. 2.20 shows the distribution curves for various

degrees of freedom. From these curves it can be seen that with small sample size, the distribution is considerably more spread out than the normal distribution, but as sample size increases the t-distribution approaches the normal distribution. By letting $t_{\alpha/2}$ denote the upper $\frac{\alpha}{2}$ x 100 percentile of the t-distribution we have,

$$P\left[-t_{\alpha/2;m-1} \leq \frac{(\bar{W} - \mu)\sqrt{m}}{S_W} \leq t_{\alpha/2;m-1}\right] = 1 - \alpha \qquad (2.1.5)$$

$$\text{or} \quad P\left[\bar{W} - t_{\alpha/2;m-1}\frac{S_W}{\sqrt{m}} \leq \mu \leq \bar{W} + t_{\alpha/2;m-1}\frac{S_W}{\sqrt{m}}\right] = 1-\alpha \qquad (2.1.6)$$

The random variable $\bar{W} \pm t_{\alpha/2;m-1} S_W/\sqrt{m}$ is called a confidence interval and $1-\alpha$ is the corresponding confidence coefficient or the confidence level.

The sample mean and variance based on the m(=20) data, uniform traffic with $\rho = 0.2$, are

$$\bar{W} = 300{\cdot}5 \qquad (2.1.7)$$

$$\text{and} \qquad S_W^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{W}^{(j)} - \bar{W})^2 = 1127$$

Thus, the confidence interval for this case is

$$\therefore \quad \bar{W} \pm t_{\alpha/2;m-1}\frac{S_W}{\sqrt{m}} = 300{\cdot}5 \pm 7.5 \; t_{\alpha/2;19} \,. \qquad (2.1.8)$$

Taking 0·95 as a confidence coefficient, for which $t_{0.025;19} = 2{\cdot}093$, the 95 per cent $(= 1 - \alpha)$ confidence interval is calculated as

$$\bar{W} \pm t_{\alpha/2;m-1}\frac{S_W}{\sqrt{m}} = 300{\cdot}5 \pm 15.69$$

$$= 316{\cdot}2, \quad 284{\cdot}8 \qquad (2.1.9)$$

$$\text{or} \quad 0.79, 0.712 \text{ slots}$$

Similarly the 75 per cent confidence intervals for nonuniform traffic are also calculated. The results are given in Table 2.2, which shows that the mean waiting time obtained by simulation may be considered as reliable estimates. For the case $N_T = N_Z = 5$, with single packets only and uniform traffic, Table 2.3 and Fig. 2.21 shows the CPU time necessary to achieve reliable simulation results, as a function of traffic intensity. It will be noted that simulation of a large number of cases can become expensive.

Table 2.2 CONFIDENCE INTERVALS

Traffic	Traffic intensity ρ	95% confidence interval (slots)	Mean from simulation, $\bar{W}$ (slots)
Uniform	0.2	0.71, 0.79	0.76
"	0.4	1.12, 1.23	1.19
"	0.6	2.03, 2.25	2.17
Nonuniform (2:1)	0.2	0.77, 0.87	0.83
" (2:1)	0.4	1.35, 1.52	1.45
" (2:1)	0.6	3.24, 3.63	3.49
Nonuniform (4:1)	0.2	0.82, 0.92	0.88
" (4:1)	0.4	1.66, 1.94	1.82
" (4:1)	0.6	6.32, 7.63	7.23

Table 2.3

UNIFORM TRAFFIC (N_T = N_Z = 5, single packets only)

TRAFFIC INTENSITY ρ	CPU TIME IN MINS.
0.16	0.39
0.26	0.56
0.4	0.85
0.6	1.50
0.8	3.04

2.2 EFFECTS OF SKEWED TRAFFIC, OVERLAPPING BEAMS, BATCHED PACKET ARRIVALS

Tables 2.4. through 2.6 give the output statistics obtained for $N_Z = N_T = 5$ and Tables 2.7 through 2.14 give the output statistics obtained for $N_Z = 20$, $N_T = 5$ from simulation. In general, various different kinds of traffics have been considered,

i) uniform traffic

ii) nonuniform traffic

iii) nonoverlapping zone traffic

iv) overlapping zone traffic

v) single packet traffic

vi) multiple packet traffic

For nonuniform traffic, in the $N_Z = N_T = 5$ case, two types of traffic skew (2:1) and (4:1) have been considered, whereas for the $N_Z = 20$, $N_T = 5$ case only one type of traffic skew (5:1) has been dealt with. With these different skews, the probability of source and destination zones are given in the corresponding tables. Nonoverlapping and overlapping zones for $N_Z = 20$, $N_T = 5$ case are shown in Figs. 2.38 and 2.39, respectively. Two different kinds of overlapping zones have been considered. In Fig. 2.39(a) zones 1,2 and 3 overlap as well as zones 6 and 7. This means that no two packets with source or destination address from the overlapping zones can be sent in the same time slot, e.g., if zone 1 is sending or receiving a packet, zones 2 and 3 cannot send or receive, respectively. On the other hand, if zone 2 is sending or receiving a packet, zone 3 can send or receive but not zone 1, since zone 2 overlaps with zone 1 and not zone 3. However, if zone 2 is sending, zone 1 may receive because the uplink and downlink frequencies are different. The probability distribution of the nonuniform traffic (5:1) for

$N_Z = 20$, $N_T = 5$ is shown in Figure 2.40. In Fig. 2.39(b) zones 13, 14 and 15 overlap as well as zones 18 and 19. The difference between Figs. 2.39(a) and 2.39(b) is that the traffic from overlapping zones in Fig. 2.39(a) is more probable than traffic from overlapping zones in Fig. 2.39(b). In the simulation results shown in various Tables and figures, traffic as in Figs. 2.39(a) and 2.39(b) has been designated as overlapping zones (high probability) and overlapping zones (low probability), respectively. Messages may consist of various numbers of packets. However, for simplicity, only two different kinds of message length statistics have been considered -messages consisting of only single packets and messages consisting of either single and or multiple packets each with equal probability of occurrence. In general, multiple packet messages might have many different lengths, but for our purposes, only messages with length eight times that of a single packet have been considered. In the simulation results we have represented messages with a combination of equiprobable single and multiple packets as "multiple packets."

Tables 2.4 through 2.6 and 2.7 through 2.14 show the queue statistics and switch matrix statistics in parts (a) and (b) respectively. The results of these various tables are plotted in Figs. 2.22 through 2.37 and in Figs. 2.41 through 2.62. In any table for queue statistics, the most important parameters of interest, the average number of packets waiting in queue $\bar{Q}$, mean packet waiting time in slots $\bar{W}$ and maximum number of packets waiting in queue Q_{max}, have been shown as a function of traffic intensity ρ. Traffic skew s:1 implies that the probability of having packets from a source

or to a destination is s times higher than that from other sources or to other destinations. In the case of uniform traffic, skew is 1:1. For the case of $N_Z = N_T = 5$ we have considered traffic skew 2:1 and 4:1; Traffic skew 2:1 means that Prob. [zone 1 to 3] = 2 Prob. [zone 4 to 5] and skew 4:1 means Prob.[zone 1 to 3] = 4 Prob.[zone 4 to 5]. For $N_Z = 20$, $N_T = 5$, traffic skew 5:1 has been considered, which means Prob.[zone 1 to 4] = 5 Prob.[zone 17 to 20], as shown in Fig. 2.40.

The average number of packets waiting in queue $\bar{Q}$ is given by the ratio of cumulative time integral to relative clock time [GPS 71]. The cumulative time integral for a queue is the sum of all of the clock units spent by all of the packets in the queue. The mean packet waiting time $\bar{W}$ is given as the cumulative time integral for the queue divided by total entries into the queue during the run. The maximum queue length Q_{max} gives the maximum number of packets in queue at any time during the run.

In any table for switch matrix statistics, the most important parameters of interest, maximum number of transponders used N_{max} and average number of transponders used $\bar{N}$, have been shown as a function of traffic intensity ρ. N_{max} represents the maximum number of transponders used at any time during the run. The cumulative time integral for storage is the sum total of all of the clock units spent by all of the packets in storage. The average number of transponders $\bar{N}$ is given by the ratio of cumulative time integral to relative clock time. The traffic intensity ρ is defined as mean packet arrivals/slot/transponder.

A DA/DS system with list scheduling has been considered. Now, according to the switching constraints, no more than one packet with same source (or same destination) zone can be transmitted in a given time slot. As

a result, in general, for nonuniform traffic, the queue lengths and packet delays are expected to be larger than those for uniform traffic. In cases of nonuniform traffic, transmissions from the source or destination zones having higher probability mass functions are more likely to occur, which implies that higher numbers of packets have repeating source (or destination) zones as compared to the case of uniform traffic. Due to our switching constraint, the larger the number of packets with the same source (or destination) zones, the more the queue builds up and and the average delay of the packets increases. Packets from high probability source or destination zones remain unprocessed with higher probability. Tables 2.4(a) to 2.6(a) and Tables 2.7(a) to 10(a) verify what has been discussed above. Comparing Tables 2.4-2.6, it is obvious that the more nonuniform the traffic, the higher is the average queue content and the average delay.

Tables 2.4 through 2.6 deal with $N_Z = N_T = 5$, and Tables 2.7 through 2.10 deal with $N_Z = 20$, $N_T = 5$. Figs. 2.22 and 2.41 show a comparison of average number of waiting packets $\bar{Q}$ vs. traffic intensity for various traffic distributions. Fig. 2.22 deals with $N_Z = 5$, $N_T = 5$ whereas Fig. 2.41 deals with $N_Z = 20$, $N_T = 5$. Figs 2.23, 2.42, 2.43, and 2.44 show comparisons of average waiting time $\bar{W}$ for packets vs. traffic intensity ρ for various traffic distributions. Fig. 2.23 has $N_Z = N_T = 5$ while the others have $N_Z = 20$, $N_T = 5$. Q_{max} denotes the maximum number of packets waiting in queue at any time during the simulation run, and since we have assumed Poisson arrival message statistics, we expect Q_{max} to increase with traffic intensity ρ. This is verified in Tables 2.4-2.6 and also in Tables 2.7-2.10. Plots of maximum queue length Q_{max} vs traffic intensity ρ for

various traffic distributions are shown in Figs. 2.24-2.29 and Fig. 2.45. The first five figures deal with $N_Z = N_T = 5$ and the last one with $N_Z = 20$, $N_T = 5$. The maximum queue length can have only discrete values since it represents a number of packets. There is not much difference in queue statistics for low values of traffic intensity , because for low traffic intensity queue occupancy is low and the effect of the switching constraint is less. As a result it does not matter much whether the traffic is uniform or nonuniform. But for higher values of traffic intensity there are many packet arrivals and the switching constraints have a more significant effect, resulting in a stronger dependence of the results on different traffic distributions.

For uniform and nonuniform traffic saturation occurs at different values of traffic intensity ρ. For uniform traffic $\rho_{sat} \approx 1$, for the nonuniform (2:1) case $\rho_{sat} \approx 0.8$ and for nonuniform (4:1), $\rho_{sat} \approx 0.7$. These are all for $N_Z = N_T = 5$.

Tables 2.4 -2.6 and Tables 2.7-2.10 also give comparisons of the average number of transponders used N and the maximum number of transponders used N_{max} for different traffic distributions. These results are plotted in Figs. 2.30-2.37 and Figs. 2.46-2.47. Like the maximum number of waiting packets in queue Q_{max}, the maximum number of transponders used N_{max} can have only discrete values. In general, we expect all the transponders to be busy when the traffic intensity is high for $N_Z = N_T = 5$ because for high traffic intensity, the chances of having waiting packets from all different source and destination zones are better, which causes more transponders to be used. But for $N_Z = 20$, $N_T = 5$, we expect the maximum number of transponders used

N_{max} to be 5 for all values of traffic intensity ρ because in the same time slot, five packets out of $N_Z = 20$ can easily be found with nonrepeating source or destination address. But if $N_Z = 5$, the chances of having 5 packets with different source or destination zone address are low. Thus N_{max} stays constant at 5 in Fig. 2.47 and slowly rises to 5 in Figs. 2.34-2.37. Of course, the maximum number of transponders used can never exceed N_T, the number of transponders available. From the plots we may conclude that the distribution of traffic does not have much influence on the maximum or average number of transponders used . The two results most strongly affected by traffic distribution are mean packet waiting time W and average queue length $\bar{Q}$.

Now considering traffic from overlapping and nonoverlapping zones, we expect traffic from overlapping zones to have larger mean waiting time and average queue length than traffic from nonoverlapping zones, unless the overlapping zones have very low probability mass function. For traffic from the low probability regions, not many packets from overlapping zones arrive and it does not matter whether the traffic is from overlapping or from nonoverlapping region. But if the traffic is from overlapping zones with high probability then we may expect several packets from the overlapping zones to contend for service. No two packets from overlapping zones can be sent in the same time slot. In addition, every packet must also satisfy the condition that the overlapping zones are not sending (or receiving) any packet in the same time slot. The more constraints on the packet to be sent, the fewer the packets that can be sent in the same time slot.

Tables 2.11-2.14 give the queue statistics for traffic from different kinds of overlapping zones and verify the conclusions above. These results are plotted in Figs. 2.48-2.52. From these simulation results it may be

concluded that for traffic from high probability overlapping zones, the mean packet waiting time and the average queue content is higher than that of low probability overlapping zones . However, by comparing Figs. 2.42 and 2.48 it may be concluded that the mean packet waiting time for nonuniform traffic (5:1), nonoverlapping zones, and traffic from overlapping zones with low probability are very close, especially for low traffic intensity. A similar conclusion about the average queue length may be drawn.

Tables 2.11-2.14 also give comparisons of the average number of transponders used and the maximum number of transponders used vs. traffic intensity for traffic from different kinds of overlapping zones. These results are plotted in Figs. 2.53-2.58. In these figures, the maximum number of transponders used does not depend on the type of overlapping zones. Since $N_Z = 20$, $N_T = 5$, there are always at least 5 packets satisfying all the constraints. The average number of transponders used is the same for different kinds of overlapping zones provided that all the traffic is served.

However, saturation occurs at different values of ρ as can be found from Tables 2.13 and 2.14. For multipacket traffic from overlapping zones saturation occurs at $\rho_{sat} \approx 0.55$ for high probability overlapping zones and at $\rho_{sat} \approx .65$ for low probability overlapping zones .

Next the effect of message length on the queue of waiting packets and on the transponder usage has been studied for $N_Z = 20$, $N_T = 5$. Two types of message length statistics have been considered: only single packet messages and a mix consisting of a combination of single and multiple packet messages. Since according to the switching constraints, two packets with the same source or destination addresses cannot be scheduled in the same time slot,

we expect the queue of packets to build up for the multiple packet case and cause the mean packet waiting time $\bar{W}$ and the average queue content $\bar{Q}$ to increase. Tables 2.7-2.14 giving the queue statistics for single and multiple packet traffic show this behavior. These results are plotted in Figs. 2.42, 2.48, 2.59-2.62, which show mean packet waiting time $\bar{W}$ in slots for different message lengths and different traffic distribution, in Figs. 2.41 and 2.51, which show the effect of multiple packets on the average number of packets waiting for various traffic distributions, and Figs. 2.45 and 2.52, which give the maximum number of packets waiting on the queue for different message mixes and traffic distributions. In general, the following conclusions may be drawn from the results; for multiple packets the average queue content and the mean packet waiting time is much higher than that of single packet traffic whether the traffic is uniform, nonuniform, overlapping or nonoverlapping. For the particular case of multiple packet and traffic distributions considered, multiple packet arrivals have much more influence on mean packet waiting time and average queue content than nonuniformity of traffic.

Tables 2.7-2.14 also give the maximum number and average number of transponders used for messages of various packet lengths and different traffic distribution . These results are shown in Figs. 2.46, 2.47, 2.53 - 2.58. There is not much difference in average number of transponders used N for various packet lengths and the maximum number of transponders used N_{max} remains the same as before, since all packets are served below saturation.

Table 2.4(a) QUEUE

UNIFORM TRAFFIC

$N_Z = N_T = 5$

P_{ij} = Prob. [source zone i] . Prob. [destination zone j]

$= \frac{1}{5} \cdot \frac{1}{5}$ for $1 \leq i \leq 5$, $1 \leq j \leq 5$

Traffic intensity ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
.07	.21	.56	4
.2	.77	.76	11
.4	2.37	1.19	16
.6	6.54	2.17	25
.8	21.82	5.43	61

Table 2.4(b) SWITCH MATRIX

UNIFORM TRAFFIC

$N_Z = N_T = 5$

P_{ij} = Prob.[source zone i] · Prob. [destination zone j]

$= \frac{1}{5} \cdot \frac{1}{5}$ for $1 \leq i \leq 5$, $1 \leq j \leq 5$

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
.07	3	.37
.2	4	1.01
.4	5	2
.6	5	3.01
.8	5	4.01

Table 2.5(a) QUEUE

NONUNIFORM TRAFFIC ; Traffic skew 2:1

$N_Z = N_T = 5$

Traffic skew s:1

P_{ij} = Prob.[source zone i] · Prob.[destination zone j]

$$= \frac{s^2}{(3s+2)^2} \quad \text{for} \quad 1 \le i \le 3, \quad 1 \le j \le 3$$

$$= \frac{s}{(3s+2)^2} \qquad 1 \le i, \ j \le 3, \ 4 \le j, \ i \le 5$$

$$= \frac{1}{(3s+2)^2} \qquad 4 \le i \le 5, \ 4 \le j \le 5$$

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
.07	.22	.59	5
.2	.84	.83	11
.4	2.91	1.45	16
.6	10.54	3.5	34
.63	12.97	4.12	41
.66	18.34	5.48	51
.73	37.5	10.26	82

Table 2.5(b) SWITCH MATRIX

NONUNIFORM TRAFFIC; Traffic skew 2:1

$N_Z = N_T = 5$

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
.07	3	.37
.2	4	1.01
.4	5	2
.6	5	3.01
.63	5	3.15
.66	5	3.34
.73	5	3.65

Table 2.6(a) QUEUE

NONUNIFORM TRAFFIC; Traffic skew 4:1

$N_Z = N_T = 5$

Traffic skew s:1

p_{ij} = Prob.[source zone i] · Prob.[destination zone j]

$$= \frac{s^2}{(3s+2)^2} \quad \text{for} \quad 1 \le i \le 3, \quad 1 \le j \le 3$$

$$= \frac{s}{(3s+2)^2} \qquad 1 \le i, \quad j \le 3, \quad 4 \le j, \quad i \le 5$$

$$= \frac{1}{(3s+2)^2} \qquad 4 \le i \le 5, \quad 4 \le j \le 5$$

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
.07	.23	.61	5
.2	.89	.88	11
.4	3.66	1.83	19
.6	21.8	7.23	60
.612	26.58	8.62	67

Table 2.6(b) SWITCH MATRIX

NONUNIFORM TRAFFIC; Traffic skew 4:1

$N_Z = N_T = 5$

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
.07	2	.37
.2	3	1.01
.4	4	2
.6	5	3.01
.612	5	3.08

Table 2.7(a) QUEUE

NONOVERLAPPING ZONES

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob.[source zone i] · Prob[destination zone j]

Prob[source zone i] = Prob[destination zone j]

$= 0.083 \quad \text{for } 1 \leq i, j \leq 4$

$= 0.066 \quad \text{for } 5 \leq i, j \leq 8$

$= 0.049 \quad \text{for } 9 \leq i, j \leq 12$

$= 0.033 \quad \text{for } 13 \leq i, j \leq 16$

$= 0.016 \quad \text{for } 17 \leq i, j \leq 20$

MULTIPLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	3.88	4.02	32
0.4	9.22	4.57	43
0.6	16.10	5.40	67
0.8	30.50	7.59	108

Table 2.7(b) SWITCH MATRIX

NONOVERLAPPING ZONES

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

MULTIPLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$.
0.2	5	0.97
0.4	5	2.01
0.6	5	2.97
0.8	5	4.01

Table 2.8(a) QUEUE

NONOVERLAPPING ZONES

UNIVORM TRAFFIC

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob[source zone i]. Prob[destination zone j]

Prob[source zone i] = Prob[destination zone j]

$= \frac{1}{20}$ for $1 \leq i, j \leq 20$

MULTIPLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	.79	3.92	32
0.4	8.79	4.36	46
0.6	14.87	4.99	59
0.8	27.30	6.80	107

Table 2.8(b) SWITCH MATRIX

NONOVERLAPPING ZONES

UNIFORM TRAFFIC

$N_Z = 20, \quad N_T = 5$

MULTIPLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used $\bar{N}$
0.2	5	0.97
0.4	5	2.02
0.6	5	2.97
0.8	5	4.00

Table 2.9(a) QUEUE

NONOVERLAPPING ZONES

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob. [source zone i]· Prob.[destination zone j]

Prob.[source zone i] = Prob.[destination zone j]

$= 0.083 \quad \text{for} \quad 1 \leq i,j \leq 4$

$= 0.066 \quad \text{for} \quad 5 \leq i,j \leq 8$

$= 0.049 \quad \text{for} \quad 9 \leq i,j \leq 12$

$= 0.033 \quad \text{for} \quad 13 \leq i,j \leq 16$

$= 0.016 \quad \text{for} \quad 17 \leq i,j \leq 20$

SINGLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	0.59	0.58	7
0.4	1.32	0.66	12
0.6	2.37	0.78	16
0.8	4.38	1.09	20

Table 2.9(b) SWITCH MATRIX

NONOVERLAPPING ZONES

NONUNIFORM TRAFFIC; Traffic skew (5:1)

$N_Z = 20, \quad N_T = 5$

SINGLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
0.2	5	1.00
0.4	5	2.00
0.6	5	3.01
0.8	5	4.02

Table 2.10(a) QUEUE

NONOVERLAPPING ZONES

UNIFORM TRAFFIC

$N_Z = 20$, $N_T = 5$

P_{ij} = Prob.[source zone i]. Prob[destination zone j]

Prob[source zone i] = Prob[destination zone j]

$= \frac{1}{20}$ for $1 \leq i, j \leq 20$

SINGLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	0.56	0.55	7
0.4	1.23	0.61	12
0.6	2.16	0.72	17
0.8	4.05	1.01	20

Table 2.10(b) SWITCH MATRIX

NONOVERLAPPING ZONES

UNIFORM TRAFFIC

$N_Z = 20$, $N_T = 5$

SINGLE PACKET

Traffic intenstiy, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
0.2	5	1.00
0.4	5	2.00
0.6	5	3.01
0.8	5	4.02

Table 2.11 (a) QUEUE

OVERLAPPING ZONES (overlapping in low probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20$, $N_T = 5$

P_{ij} = Prob. [source zone i]. Prob.[destination zone j]

Prob. [source zone i] = Prob.[destination zone j]

= 0.083 for $1 \leq i,j \leq 4$

= 0.066 for $5 \leq i,j \leq 8$

= 0.049 for $9 \leq i,j \leq 12$

= 0.033 for $13 \leq i,j \leq 16$

= 0.016 for $17 \leq i,j \leq 20$

SINGLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	0.59	0.58	7
0.4	1.33	0.66	12
0.6	2.43	0.8	16
0.8	4.57	1.14	21

Table 2.11 (b) SWITCH MATRIX

OVERLAPPING ZONES (overlapping in low probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20$, $N_T = 5$

SINGLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
0.2	5	1.00
0.4	5	2.00
0.6	5	3.01
0.8	5	4.02

Table 2.12 (a) QUEUE

OVERLAPPING ZONES (overlapping in high probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob.[source zone i]. Prob[destination zone j]

Prob[source zone i] = Prob.[destination zone j]

= 0.083 for $1 \leq i,j \leq 4$

= 0.066 for $5 \leq i,j \leq 8$

= 0.049 for $9 \leq i,j \leq 12$

= 0.033 for $13 \leq i,j \leq 16$

= 0.016 for $17 \leq i,j \leq 20$

SINGLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Meand packet waiting time (slots), $\bar{W}$	Maximum queue queue length, Q_{max}
0.2	0.66	0.65	9
0.4	1.59	0 79	14
0.6	3.26	1.08	18
0.8	7.32	1.82	31

Table 2.12.(b) SWITCH MATRIX

OVERLAPPING ZONES (overlapping in high probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

SINGLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponder used, $\bar{N}$
0.2	5	1.00
0.4	5	2.00
0.6	5	3.01
0.8	5	4.02

Table 2.13 (a) QUEUE

OVERLAPPING ZONES (overlapping in low probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob.[source zone i]. Prob.[destination zone j]

Prob[source zone i] = Prob.[destination zone j]

$= 0.083 \quad \text{for} \quad 1 \leq i,j \leq 4$

$= 0.066 \quad \text{for} \quad 5 \leq i,j \leq 8$

$= 0.049 \quad \text{for} \quad 9 \leq i,j \leq 12$

$= 0.033 \quad \text{for} \quad 13 \leq i,j \leq 16$

$= 0.016 \quad \text{for} \quad 17 \leq i,j \leq 20$

MULTIPLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	3.9	4.04	32
0.4	5.4	4.67	45
0.6	16.66	5.60	70
0.65	19.69	6.02	78

Table 2.13 (b) SWITCH MATRIX

OVERLAPPING ZONES (overlapping in low probabililty region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

MULTIPLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
0.2	5	0.96
0.4	5	2.01
0.6	5	2.97
0.65	5	3.27

Table 2.14 (a) QUEUE

OVERLAPPING ZONES (overlapping in high probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

P_{ij} = Prob.[source zone i]. Prob.[destination zone j]

Prob[source zone i] = Prob.[destination zone j]

$= 0.083 \quad$ for $\quad 1 \leq i,j \leq 4$

$= 0.066 \quad$ for $\quad 5 \leq i,j \leq 8$

$= 0.049 \quad$ for $\quad 9 \leq i,j \leq 12$

$= 0.033 \quad$ for $\quad 13 \leq i,j \leq 16$

$= 0.016 \quad$ for $\quad 17 \leq i,j \leq 20$

MULTIPLE PACKET

Traffic intensity, ρ	Average queue length, $\bar{Q}$	Mean packet waiting time (slots), $\bar{W}$	Maximum queue length, Q_{max}
0.2	4.16	4.3	34
0.4	10.84	5.4	59
0.45	12.63	5.65	60
0.51	16	6.21	64
0.55	18.24	6.58	70

Table 2.14 (b) SWITCH MATRIX

OVERLAPPING ZONES (overlapping in high probability region)

NONUNIFORM TRAFFIC; Traffic skew 5:1

$N_Z = 20, \quad N_T = 5$

MULIPLE PACKET

Traffic intensity, ρ	Maximum no. of transponders used, N_{max}	Average no. of transponders used, $\bar{N}$
0.2	5	0.96
0.4	5	2.01
0.45	5	2.23
0.51	5	2.57
0.55	5	2.77

[STE 80] T.E. Stern, M. Schwartz, H.E. Meadows, H.K. Ahmadi, J.G. Gadre, I.S. Gopal and K. Matsuo, "Next generation Communications Satellites: Multiple Access and Network Studies," " Columbia University, Annual Report prepared for National Aeronautics and Space Administration, Contract NAS 5-25759, July 1980.

[KOB 78] H. Kobayashi, Modeling and Analysis, Addison-Wesley Publishing Company, Inc. 1978.

[BAS 66] J. Bass, Elements of probability theory, Academic Press Inc. 1966.

[DAN 81] D. Dannenlering and M. Starr, Management Science, An Introduction, McGraw-Hill, Inc. 1981.

[FIS 78] G.S. Fishman, Principles of Discrete Event Simulation, John Wiley & Sons, Inc. 1978.

[SCH 74] T.J. Schriber, Simulation using GPSS, John Wiley & Sons, Inc. 1974.

[GPS 71] General purpose Simulation System V User's Manual, IBM Corp., SH20-0851-1 (Second Edition, Aug. 1971).

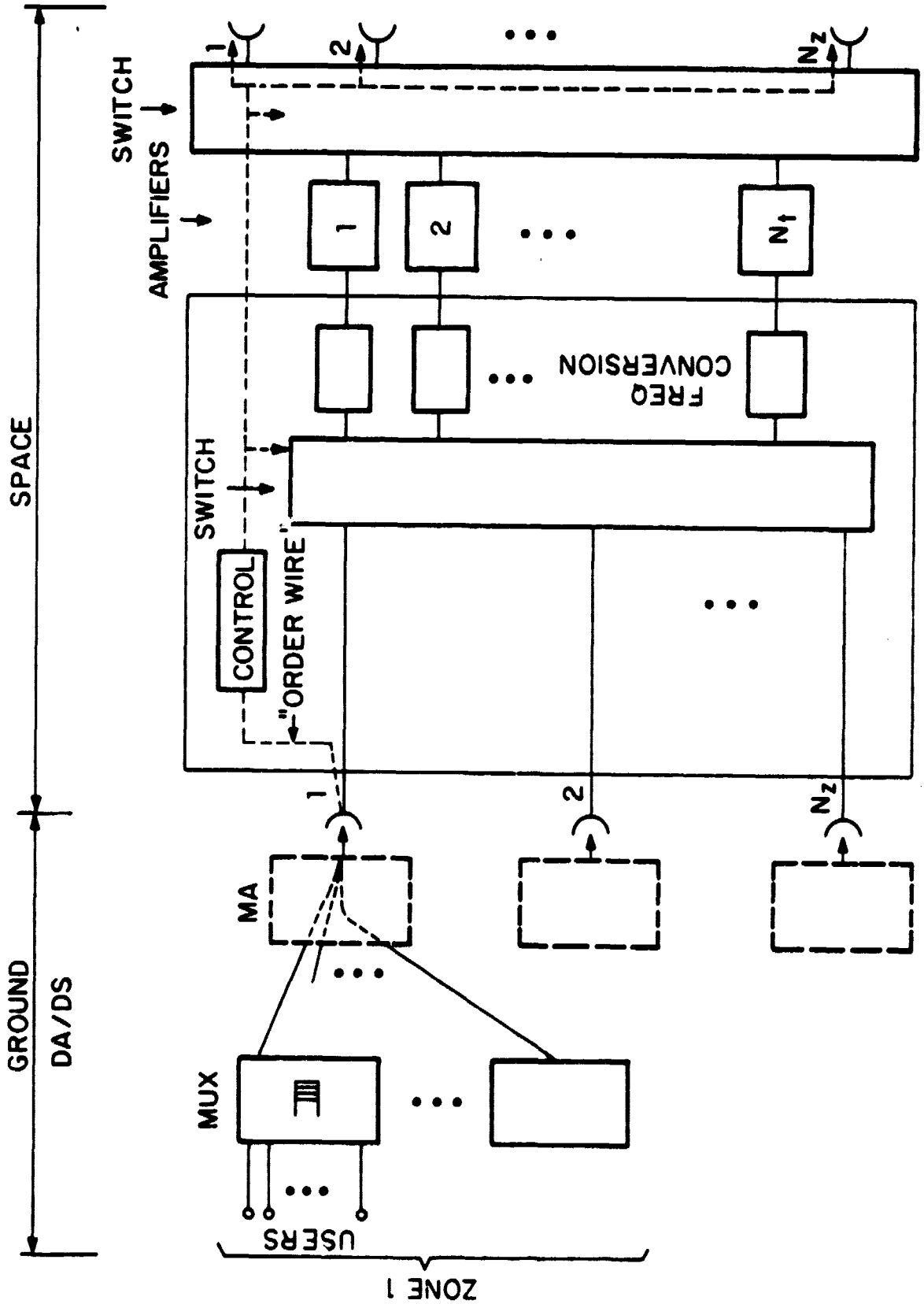
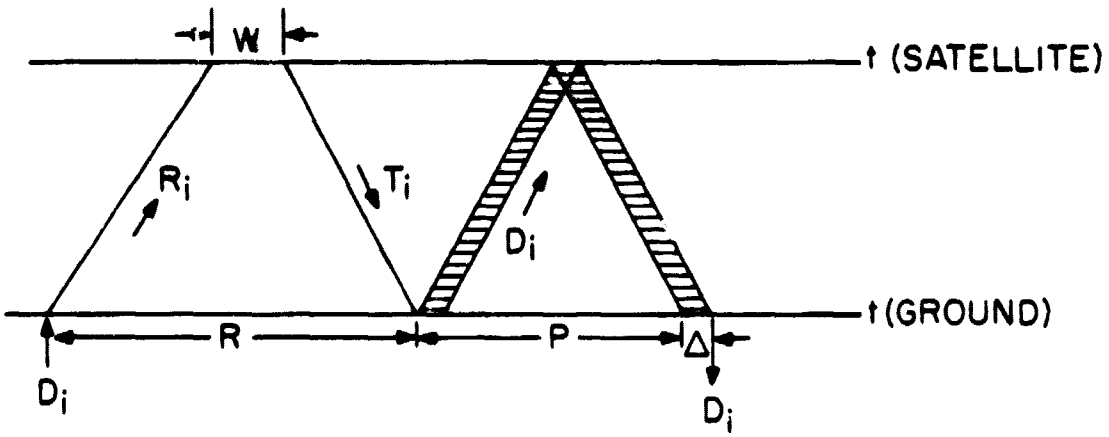


FIG. 2.1 DA/DS SYSTEM

PACKET SCHEDULING (DA/DS)



$$\overline{T} = \overline{R} + P + \Delta$$

$$\overline{R} = \overline{W} + P$$

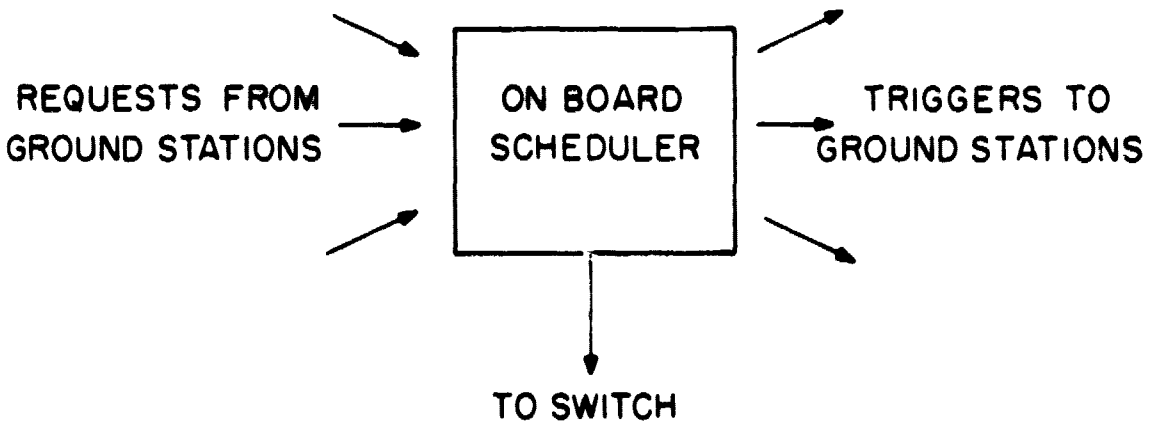


FIG. 2.2 PACKET SCHEDULING IN A DA/DS SYSTEM

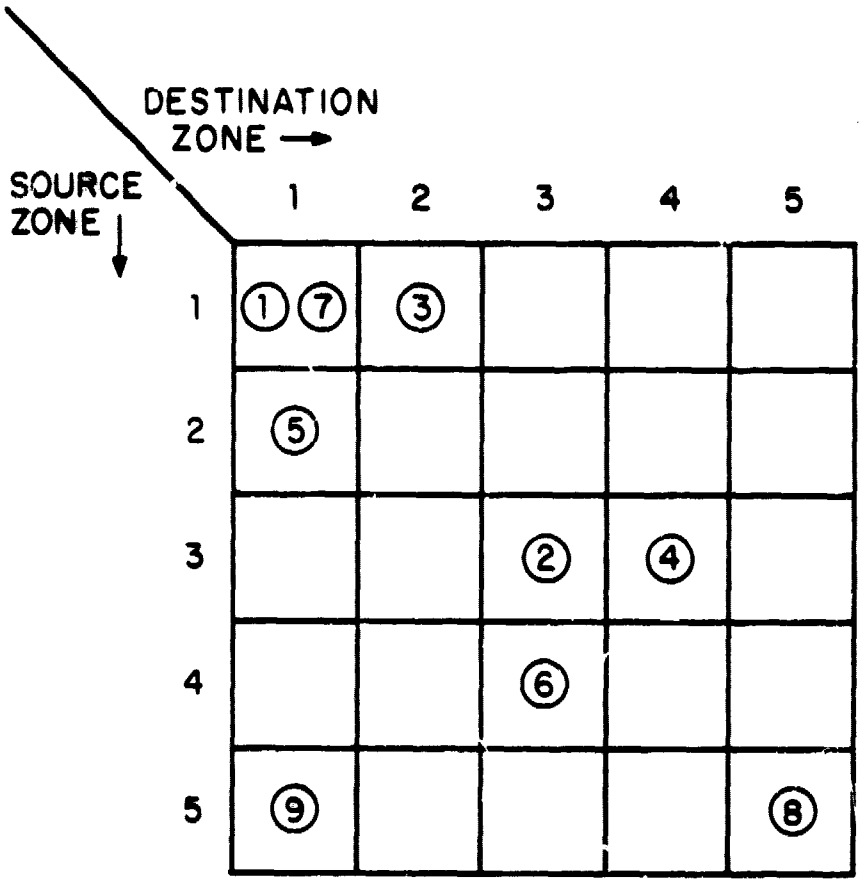


FIG. 2.3 CONTENTS OF THE REQUEST QUEUE
ORDER OF REQUESTS RECEIVED SHOWN ENCIRCLED

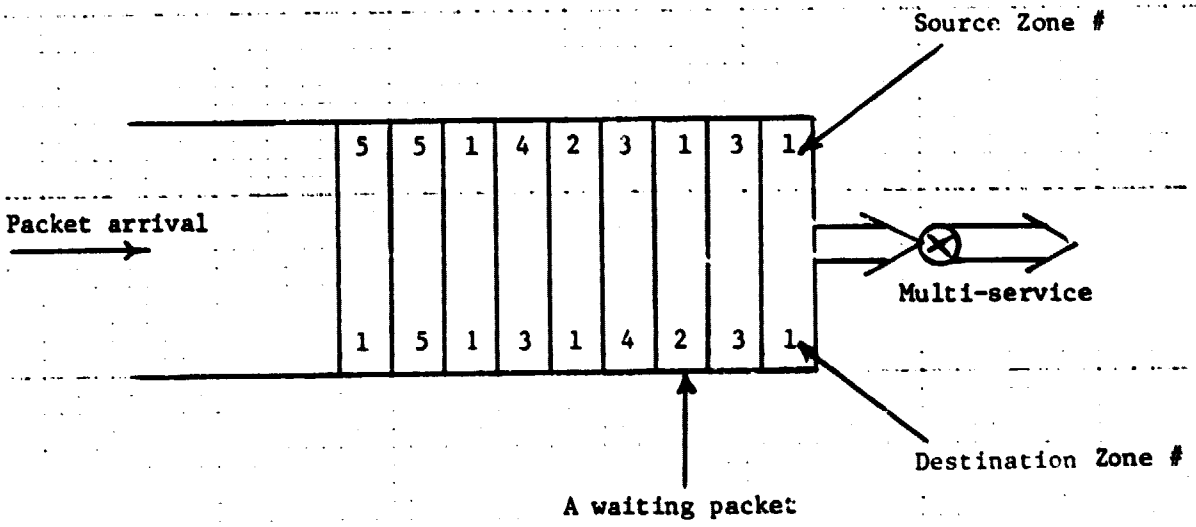


Fig. 2.4 Request Queue

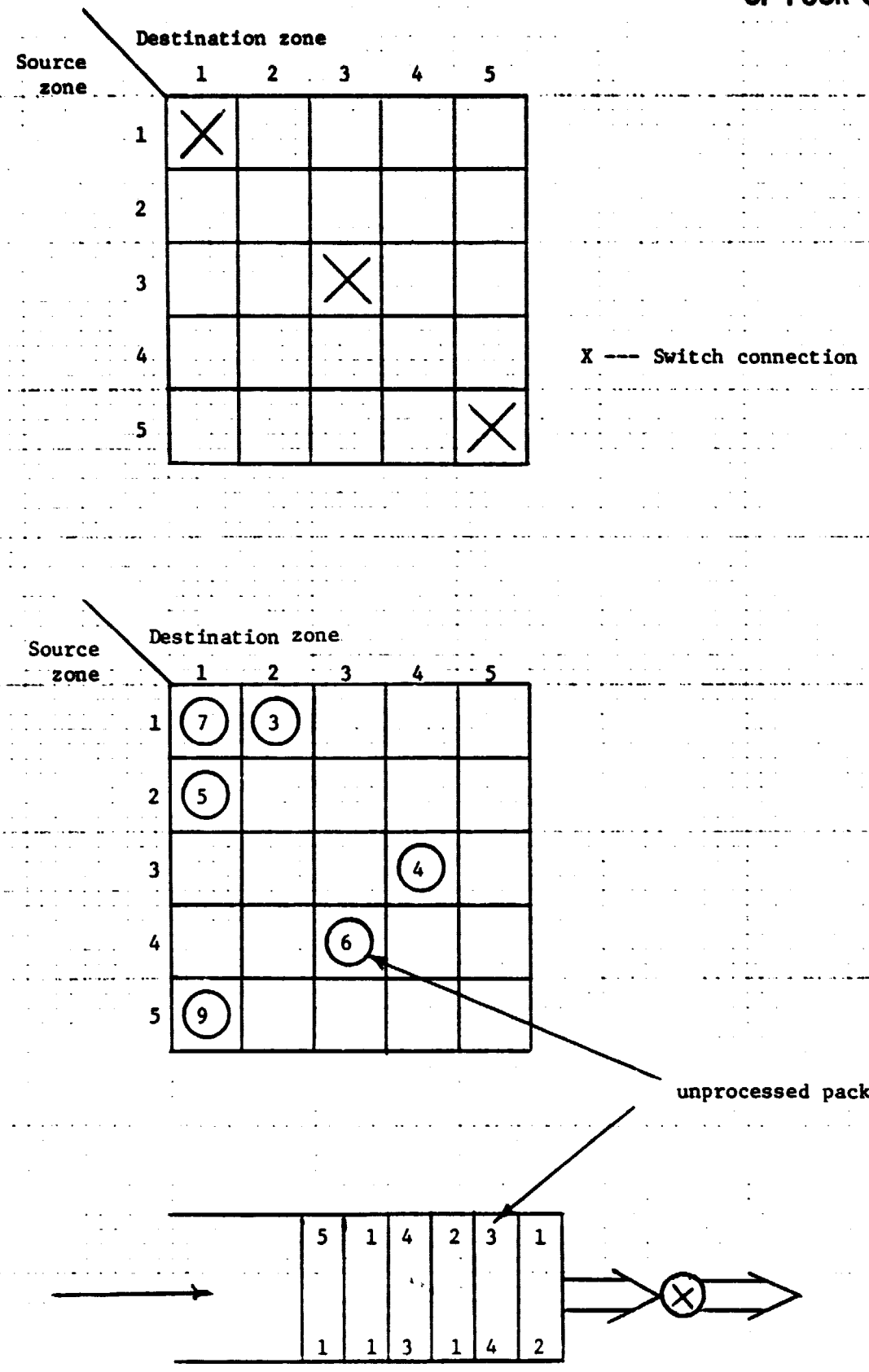
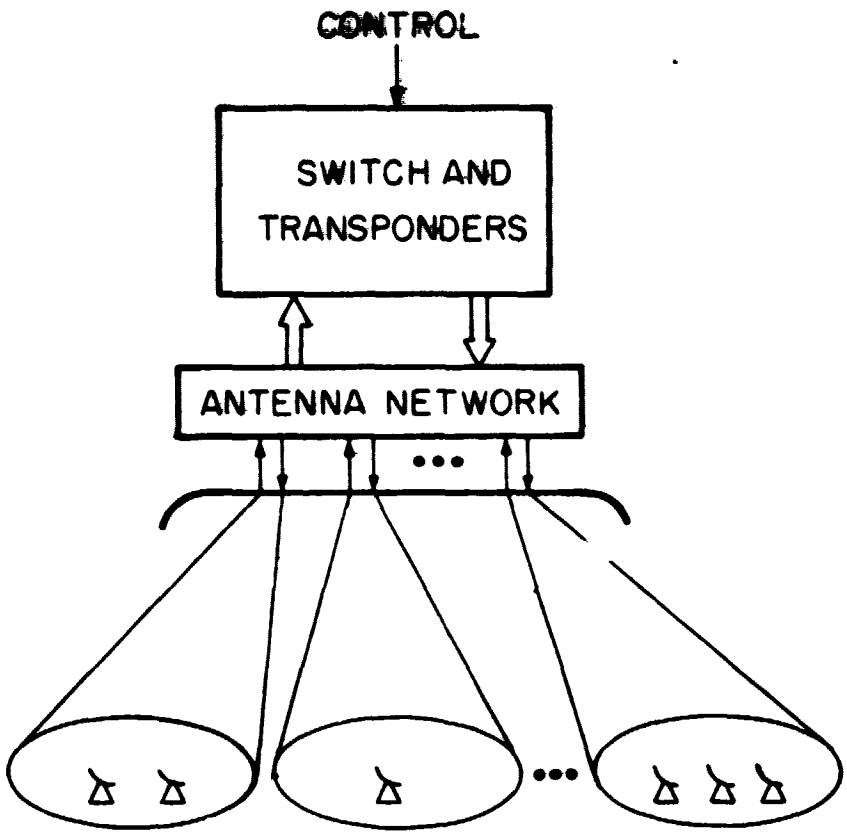


Fig.2.5 Scheduled Packets and Pending Requests.



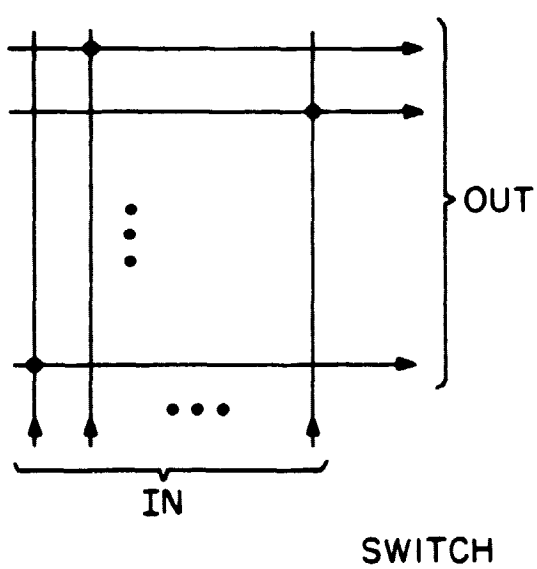


FIG. 2.6 SPACE SEGMENT

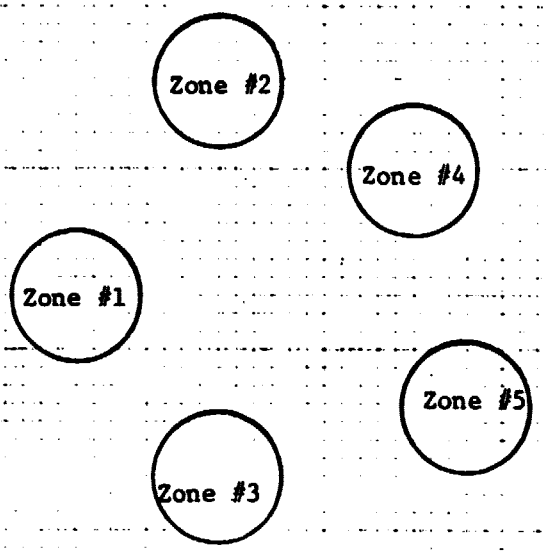


Fig. 2.7 Non-Overlapping Zones

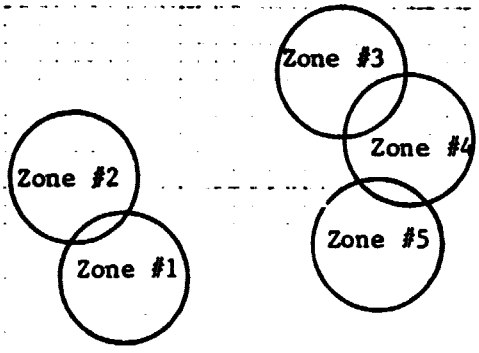


Fig. 2.8 Overlapping Zones

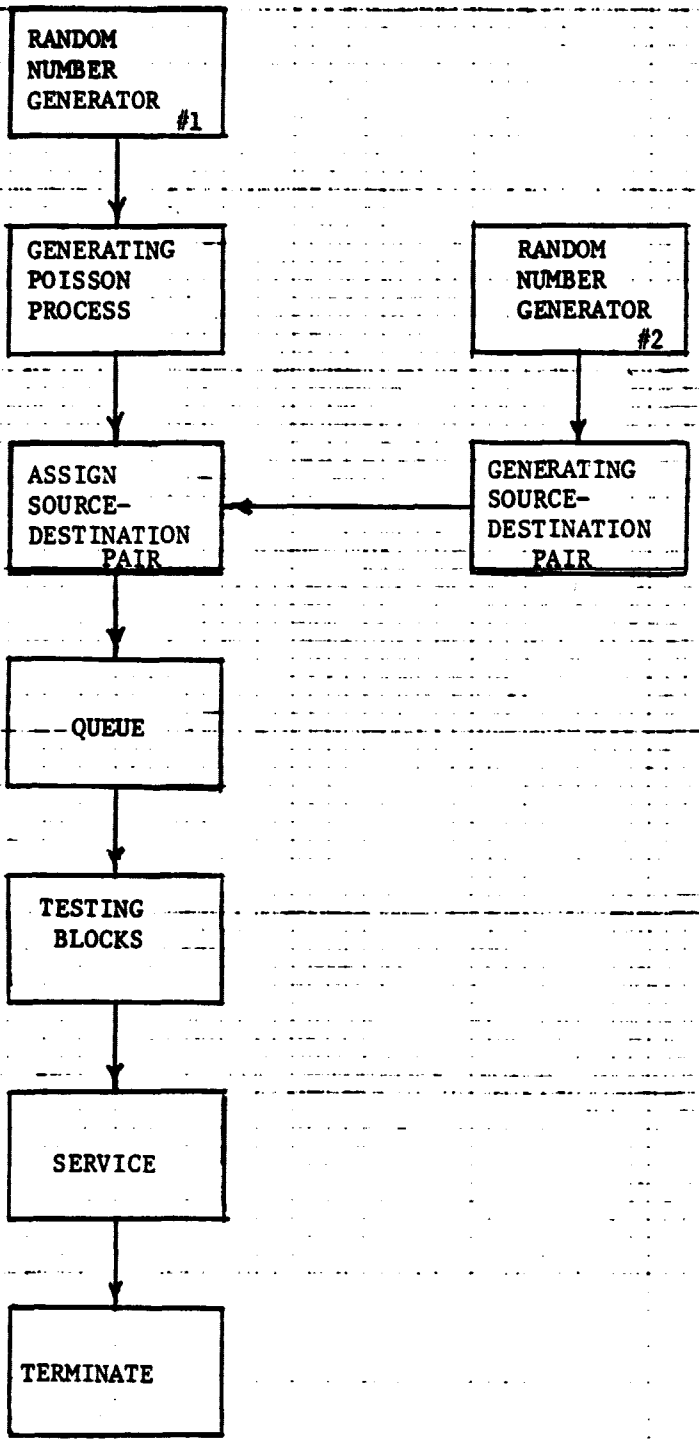


Fig. 2.9 BASIC FLOW CHART FOR SIMULATION.

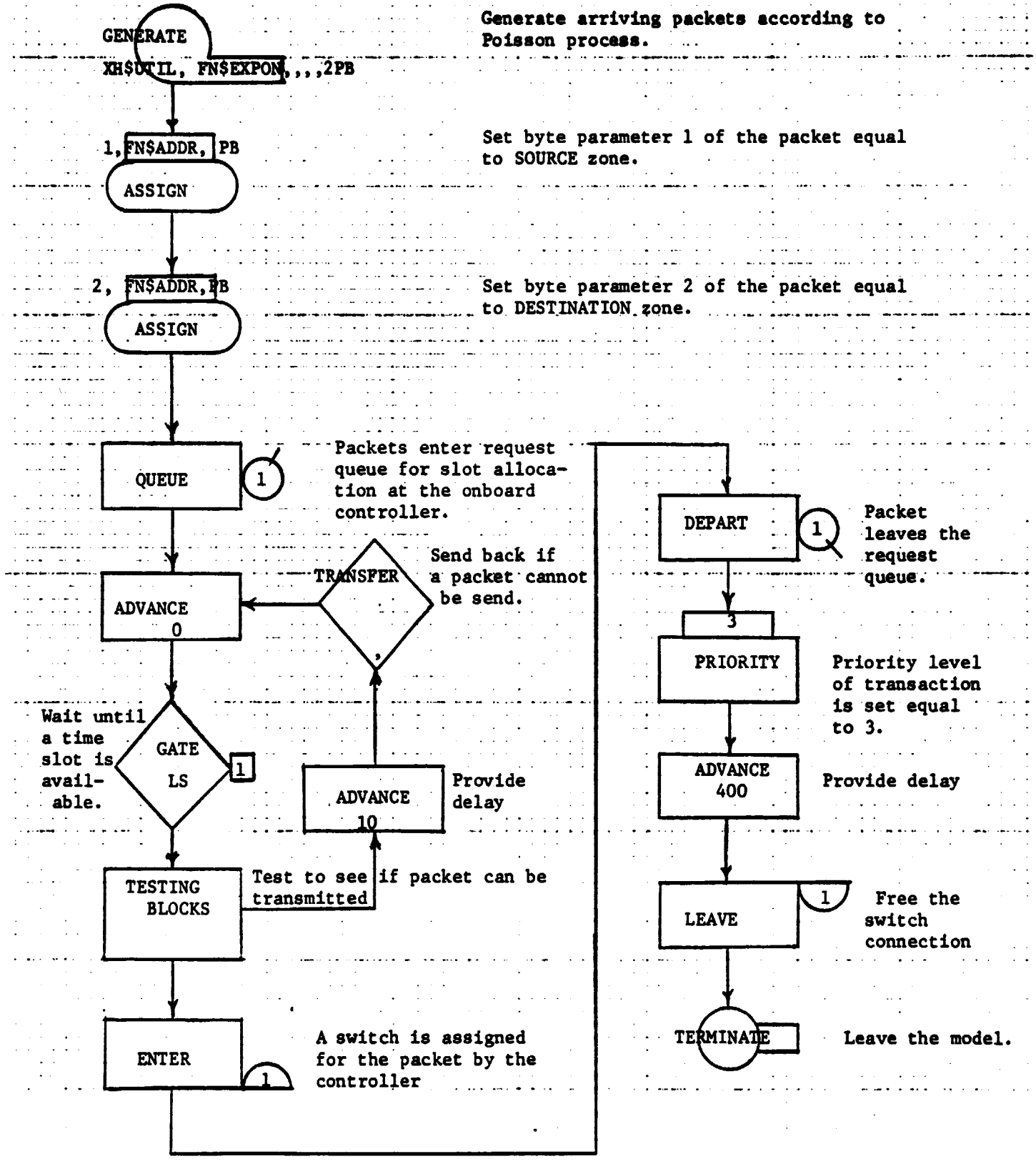
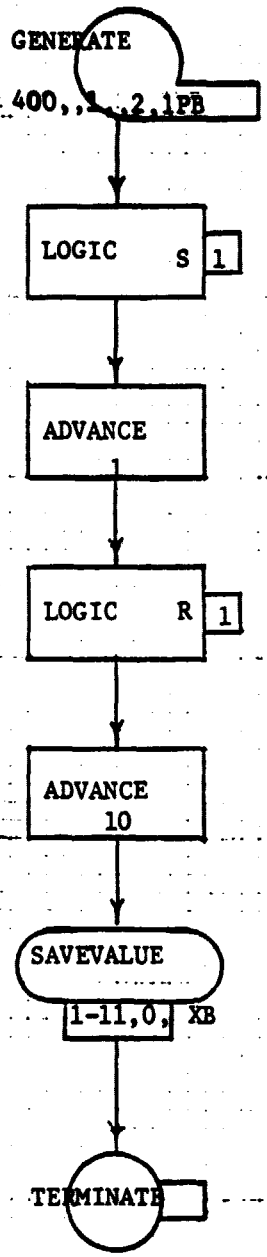


Fig. 2.10 Model Segment 1



Transactions are created at the block every 400 time units.

Open gate for packets.

Provide delay

Close the gate.

Provide delay

Update the byte savevalues containing source-destination zone addresses & no. of packets served.

Leave the model.

Fig. 2.11 Model Segment 2

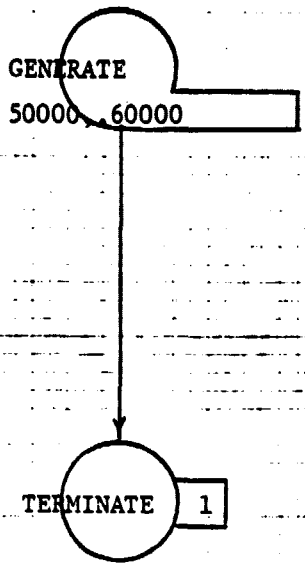


Fig. 2.12 Model Segment 3

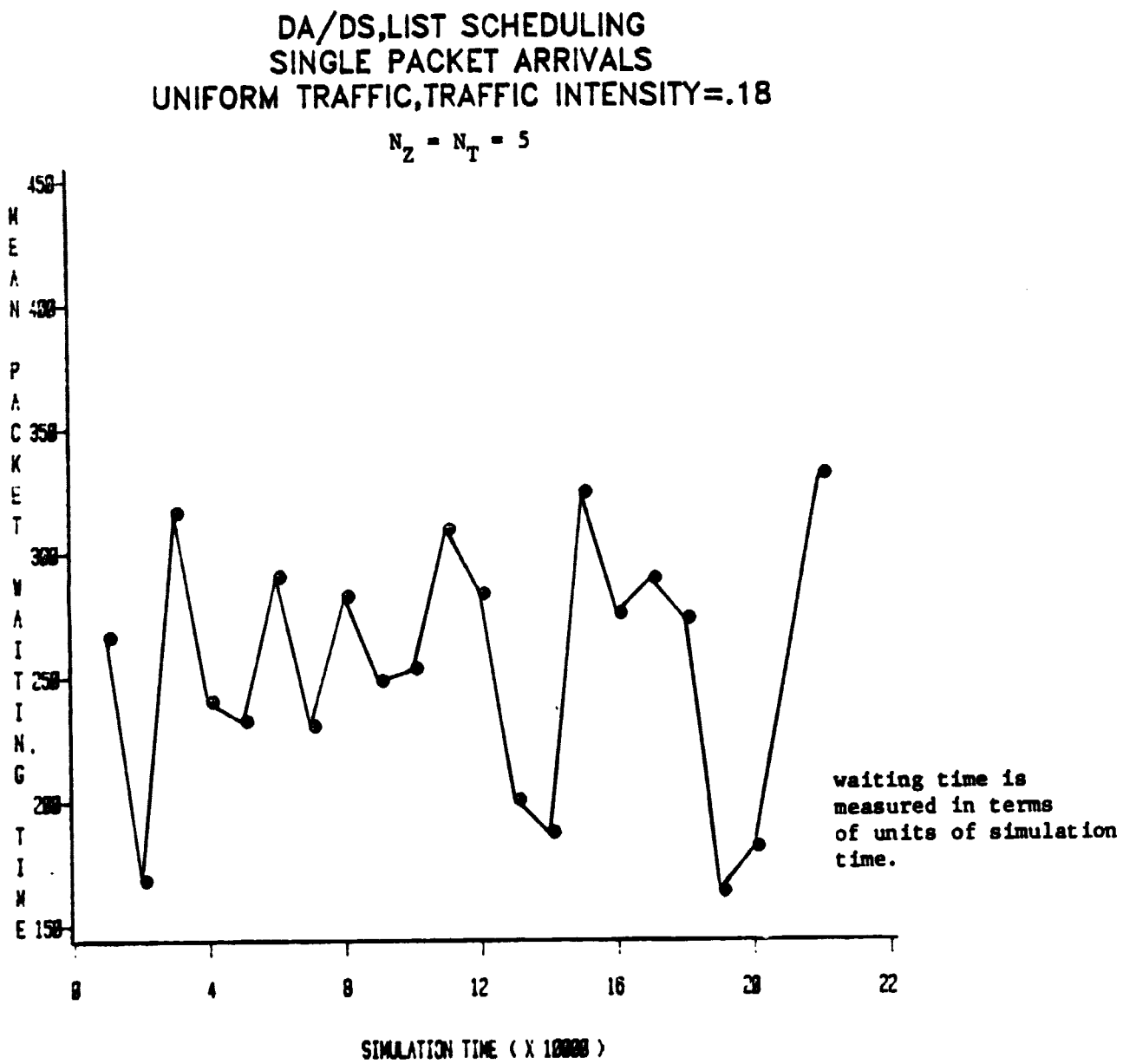


Fig. 2.13 Effect of simulation time on packet delay for uniform traffic.

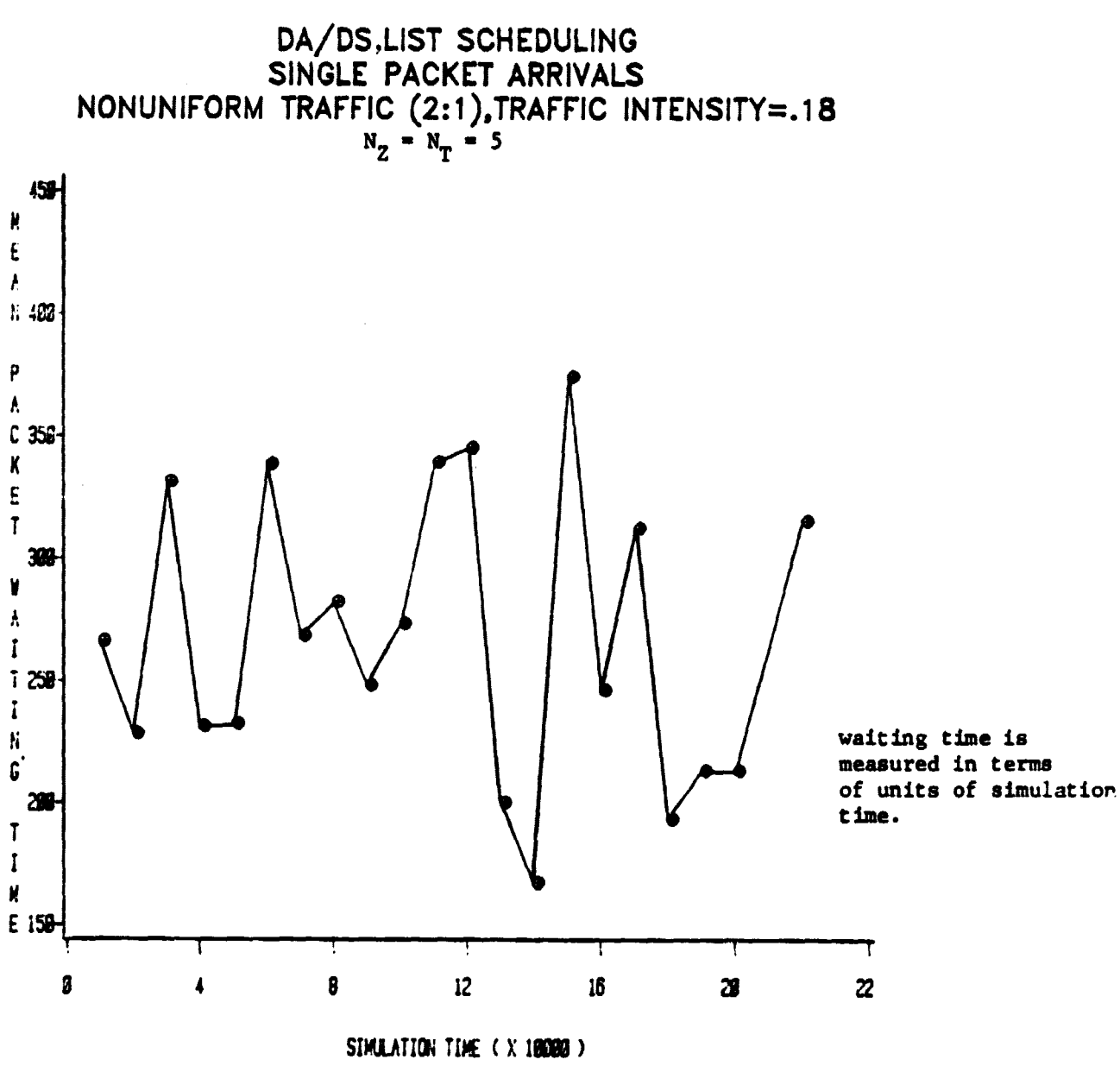


Fig. 2.14 Effect of simulation time on packet delay for nonuniform traffic (2:1).

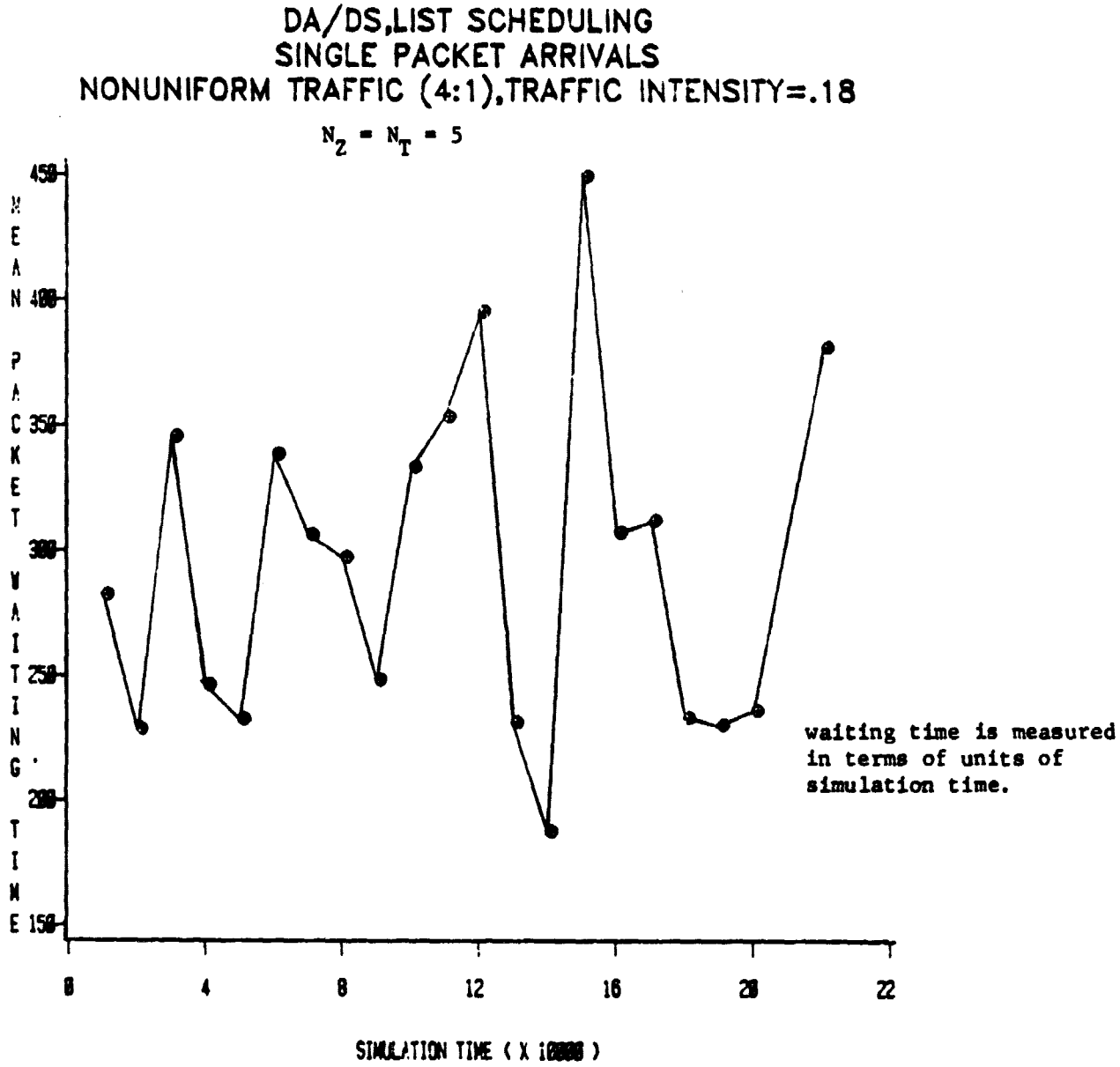


Fig. 2.15 Effect of simulation time on packet delay for nonuniform traffic (4:1).

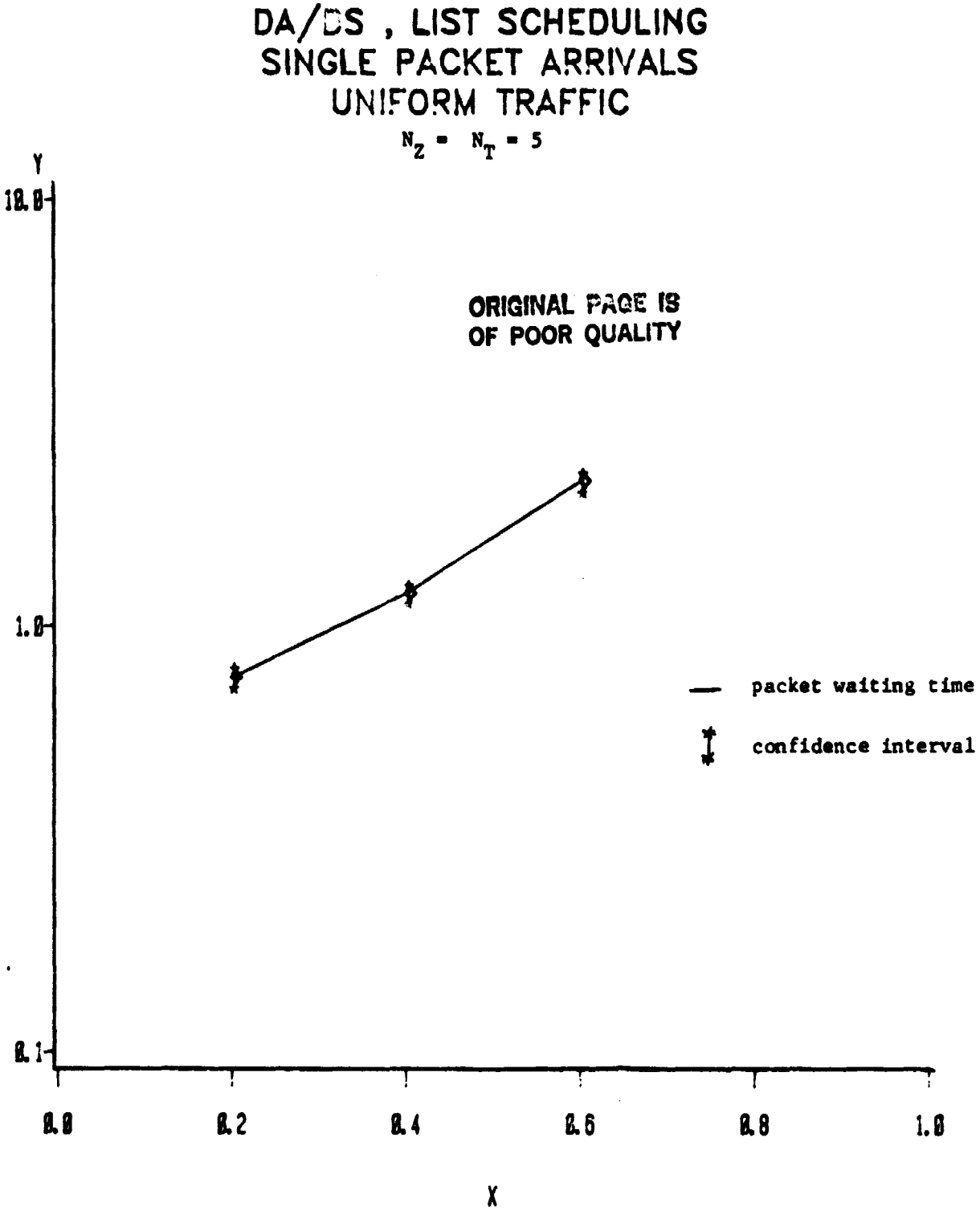


Fig. 2.16 Confidence interval on waiting time for uniform traffic.

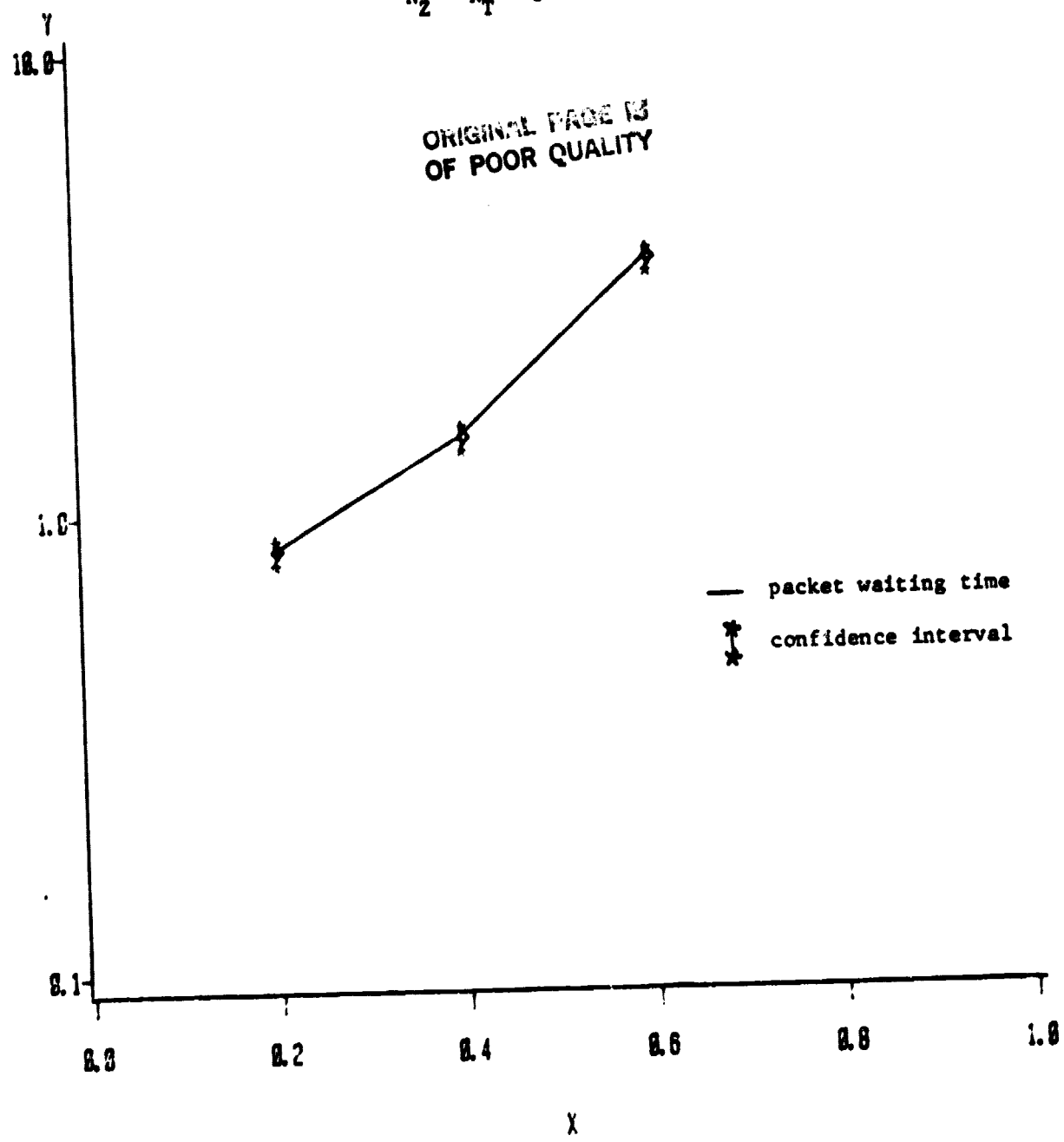


Fig. 2.17 Confidence interval on waiting time for nonuniform traffic (2:1).

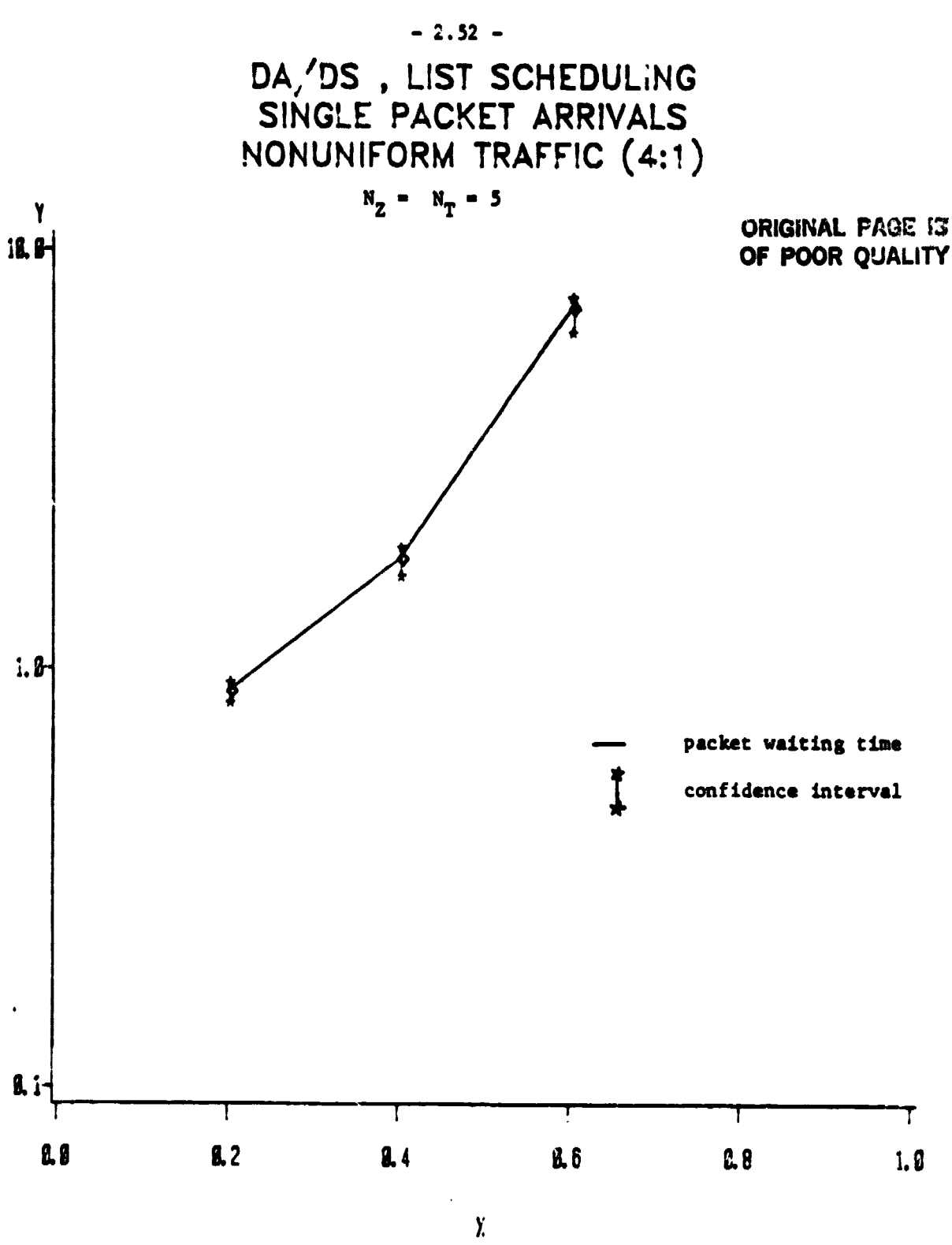


Fig. 2.18 Confidence interval on waiting time for nonuniform traffic (4:1)

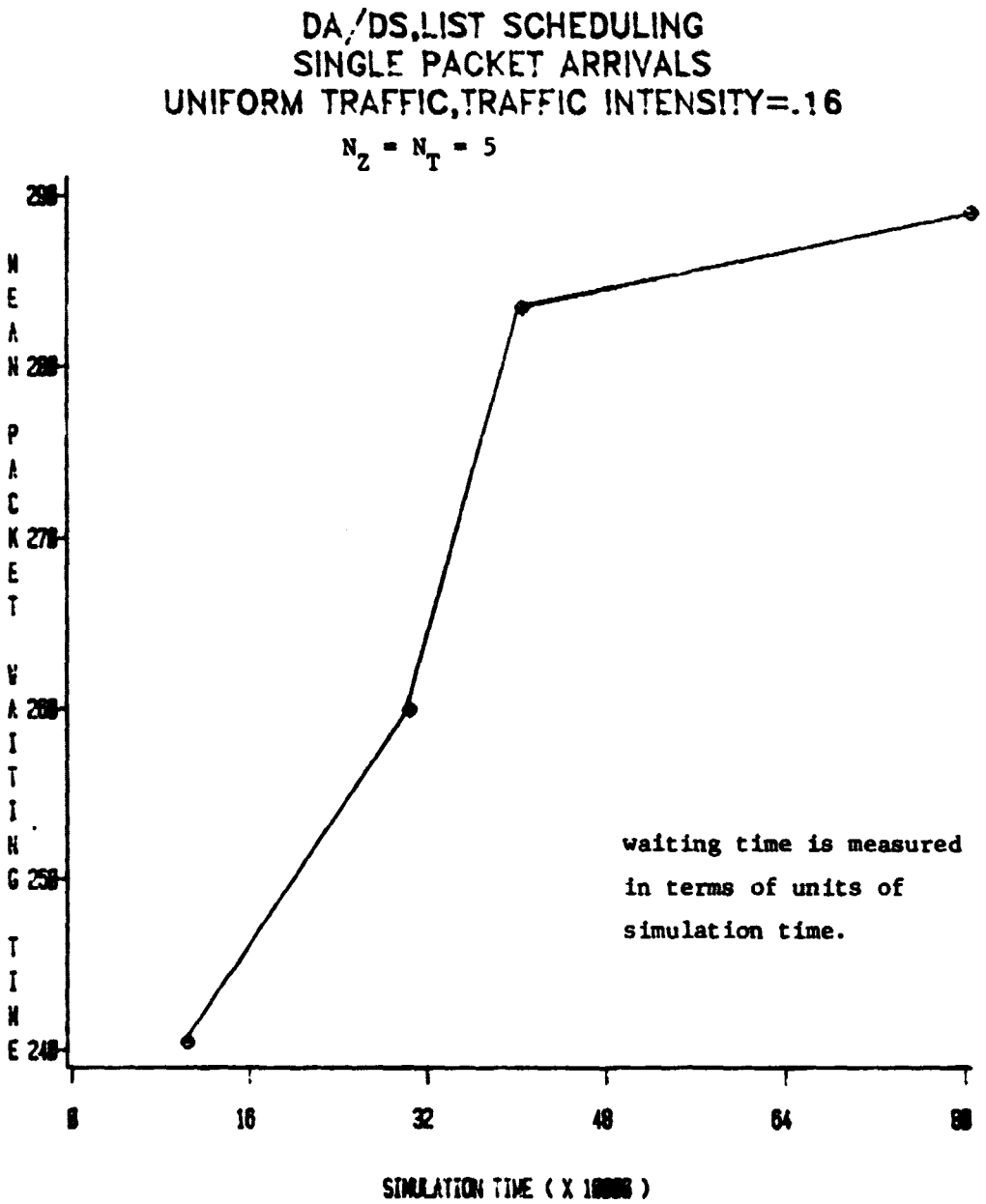


Fig. 2.19 Effect of simulation time on packet delay with initial 50,000 units of simulation time as the nonrecording period.

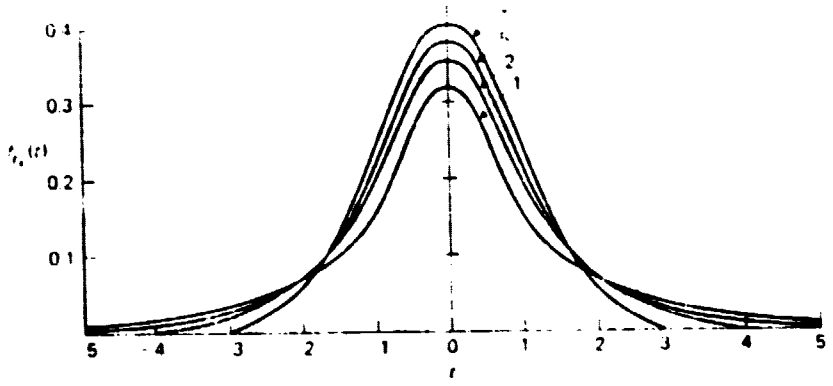


Fig. 2.20 The student's t-distribution with k degrees of freedom ($k = 1,2,5,\infty$).

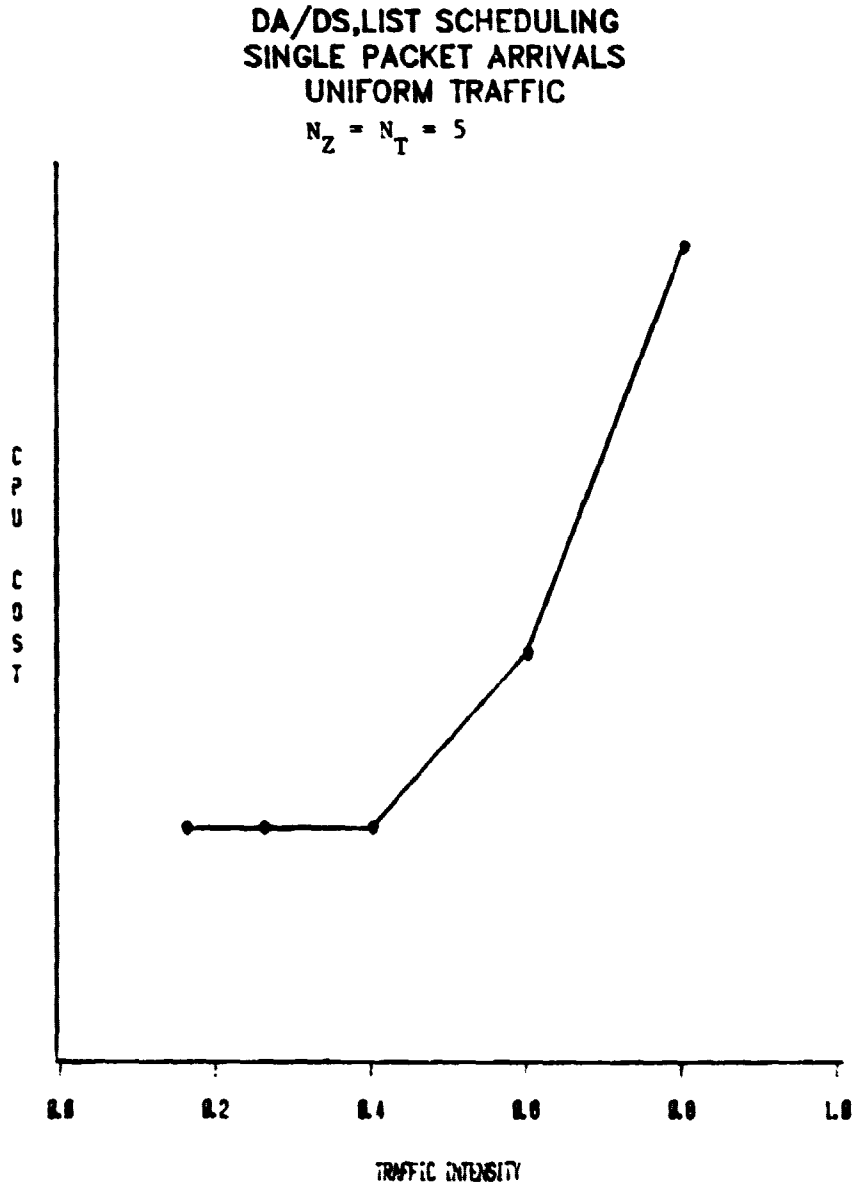


Fig. 2.21 Cost vs traffic intensity curve.

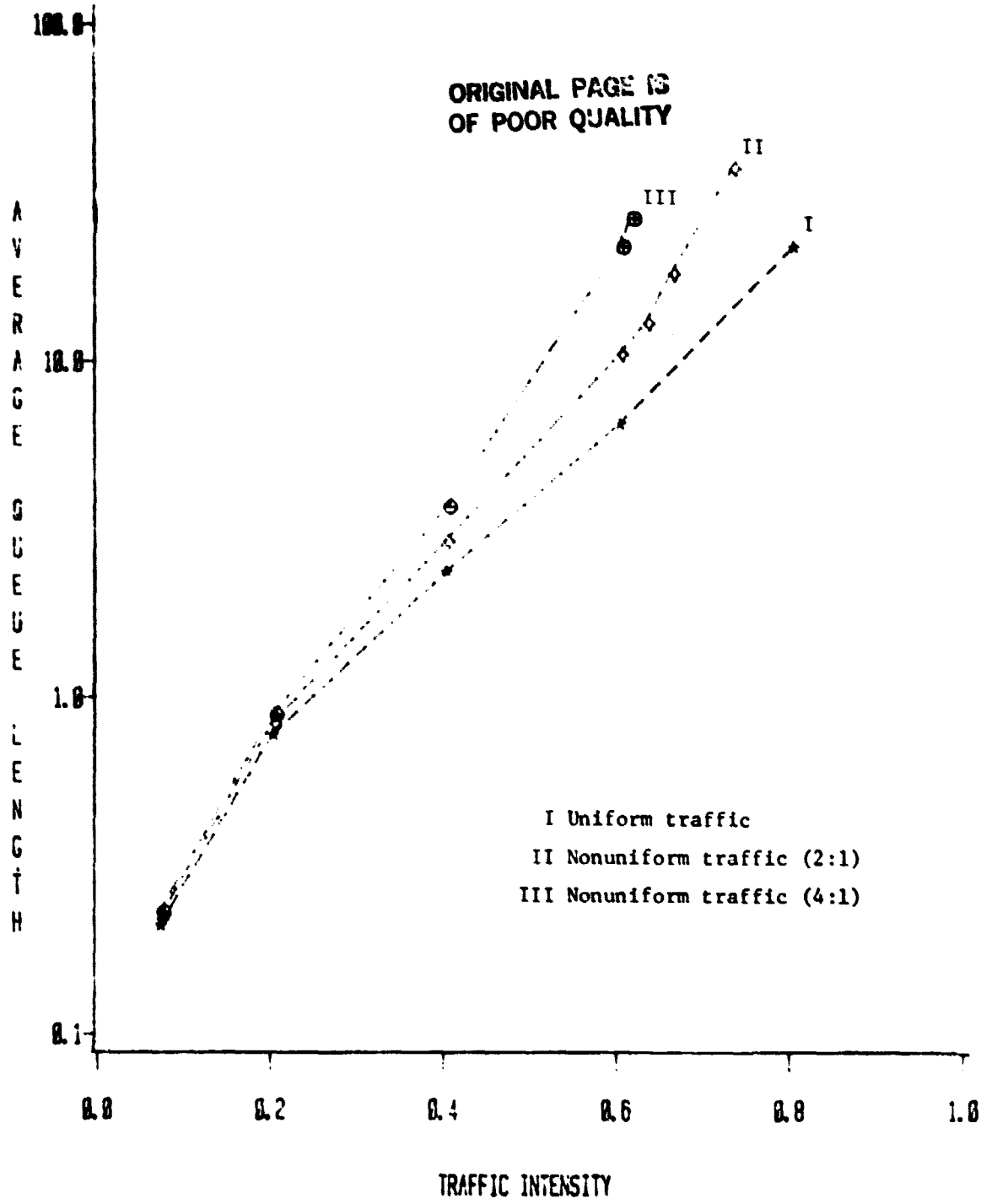
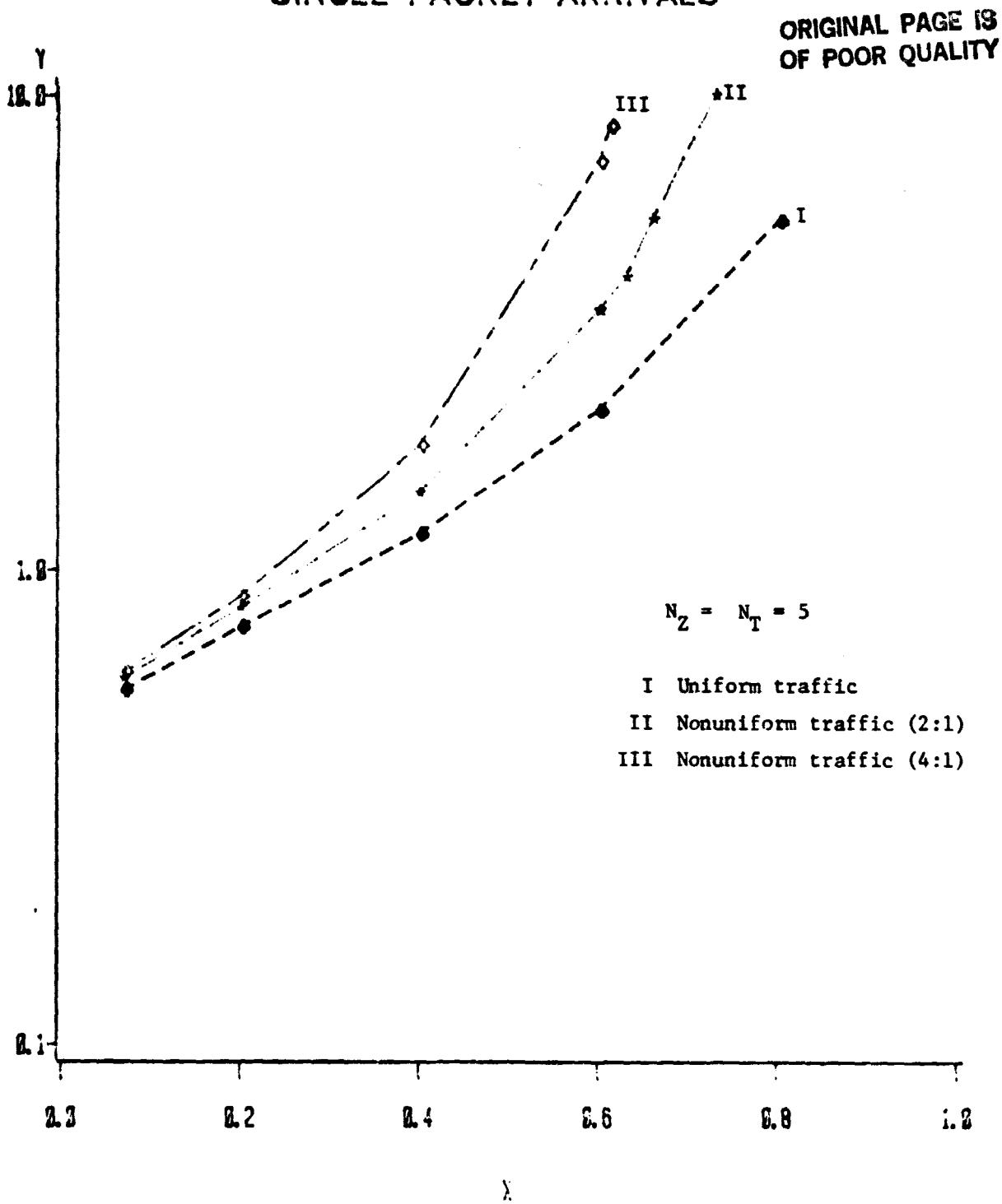


Fig. 2.22 Comparison of average number of waiting packets for different traffic distribution.



X=TRAFFIC INTENSITY
Y=MEAN PACKET WAITING TIME (SLOTS)

Fig. 2.23 Comparison of waiting time for different traffic distribution.

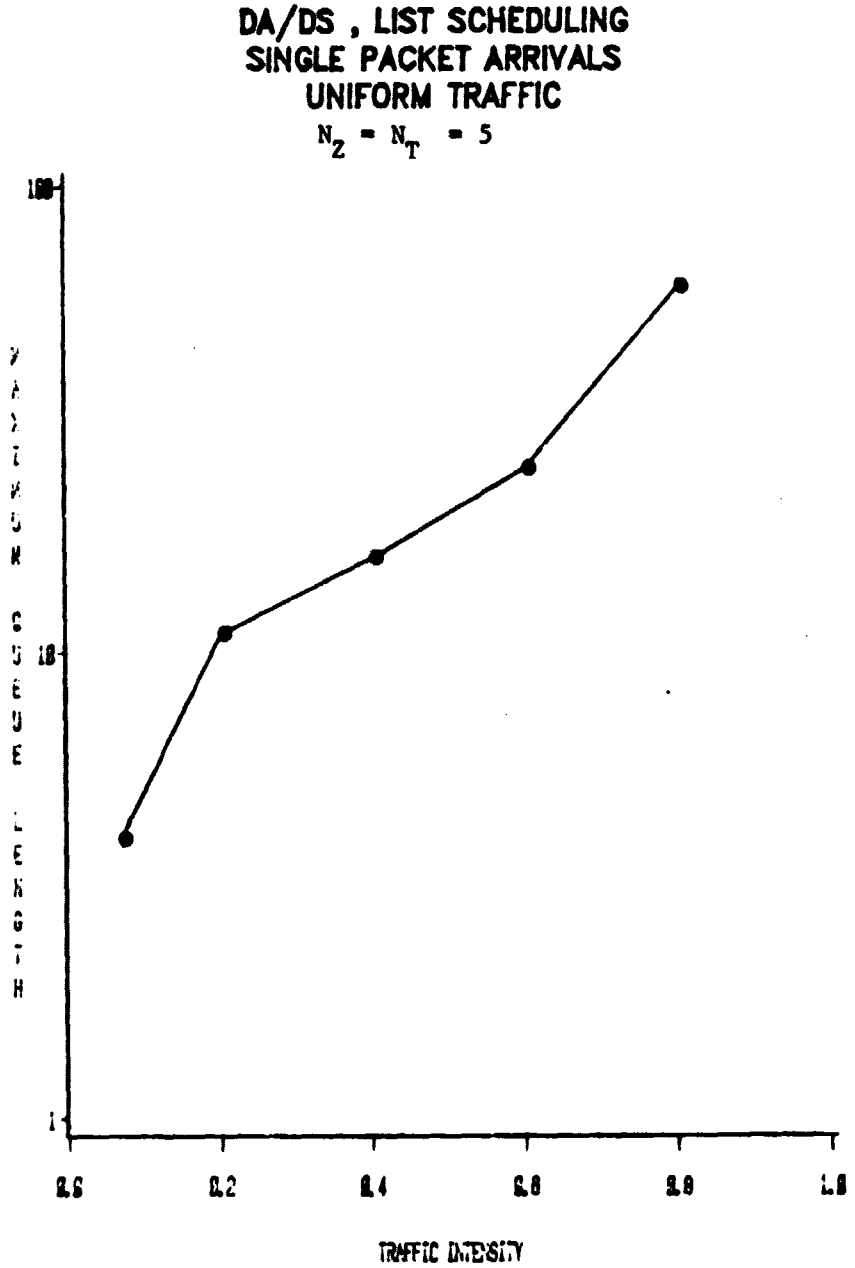


Fig. 2.24 Maximum number of waiting packets (Q_{max}) for uniform traffic.

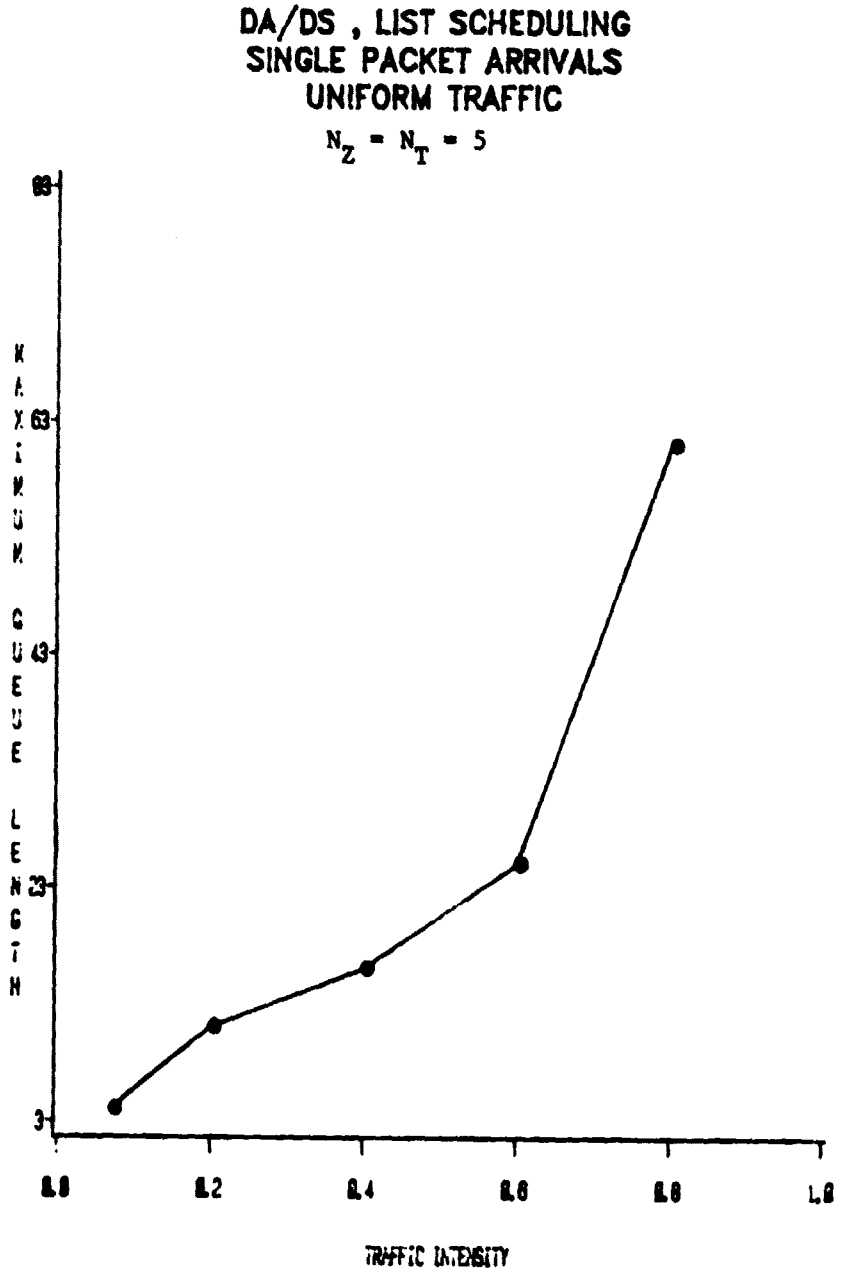


Fig 2.25 Maximum number of waiting packets (Q_{max}) for uniform traffic.

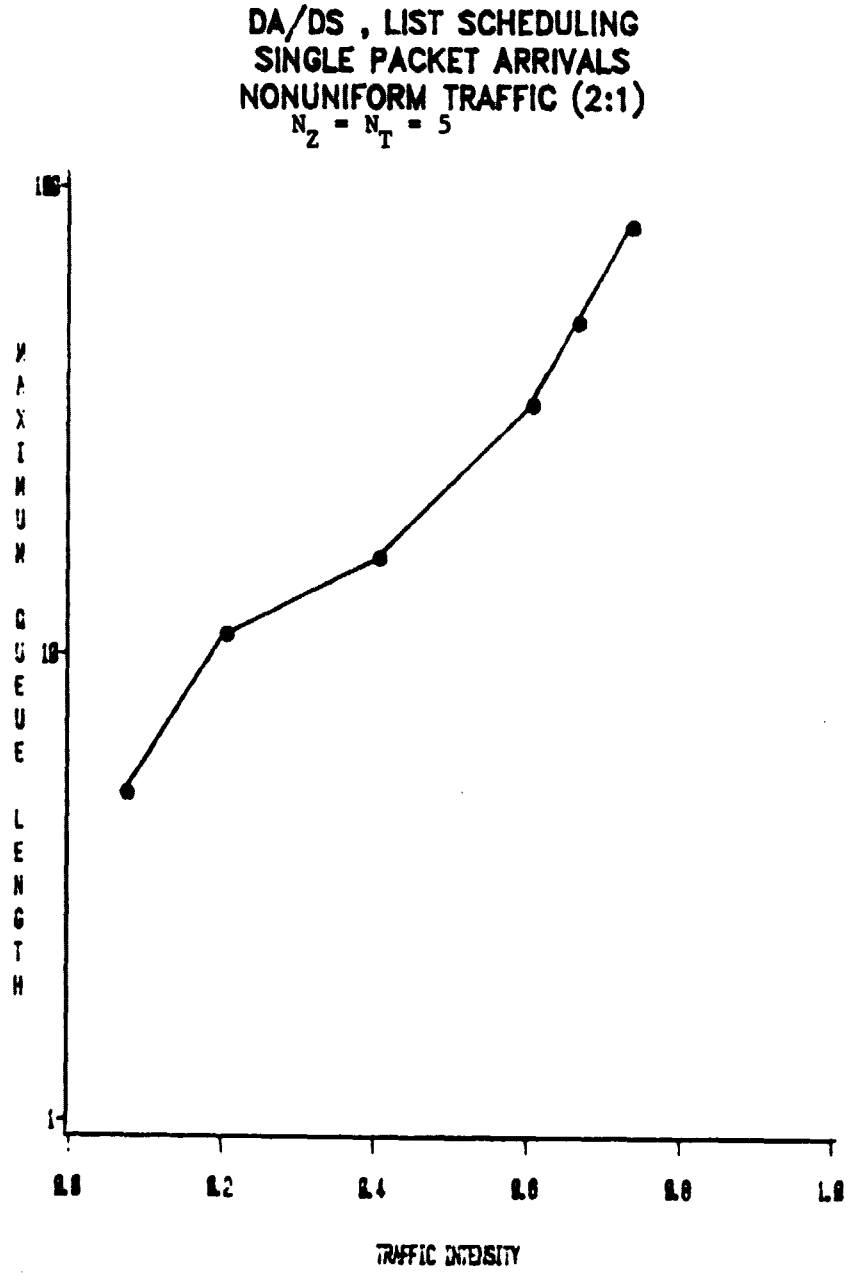


Fig. 2.26 Maximum number of waiting packets (Q_{max}) for nonuniform traffic (2:1).

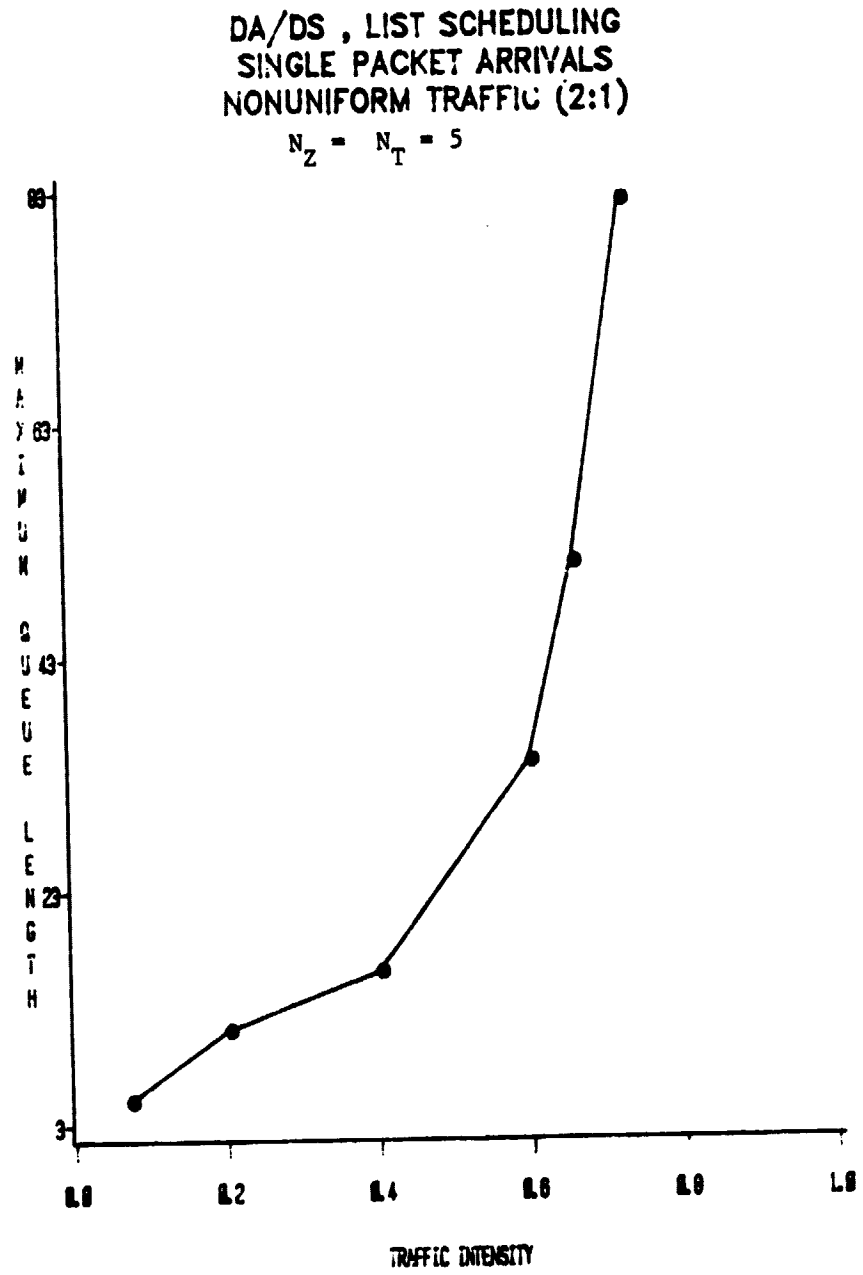


Fig. 2.27 Maximum number of waiting packets (Q_{max}) for nonuniform traffic (2:1).

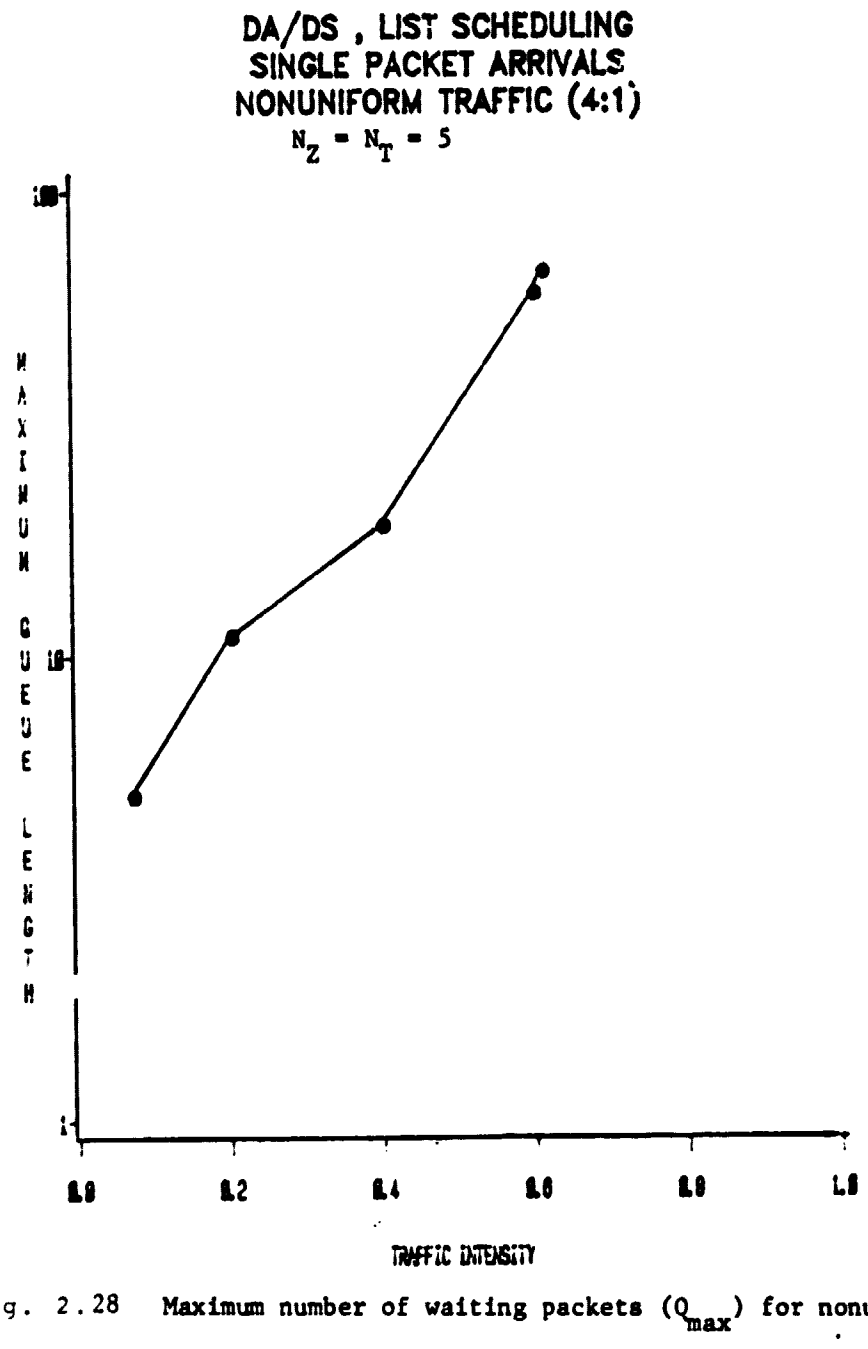


Fig. 2.28 Maximum number of waiting packets (Q_{max}) for nonuniform traffic (4:1).

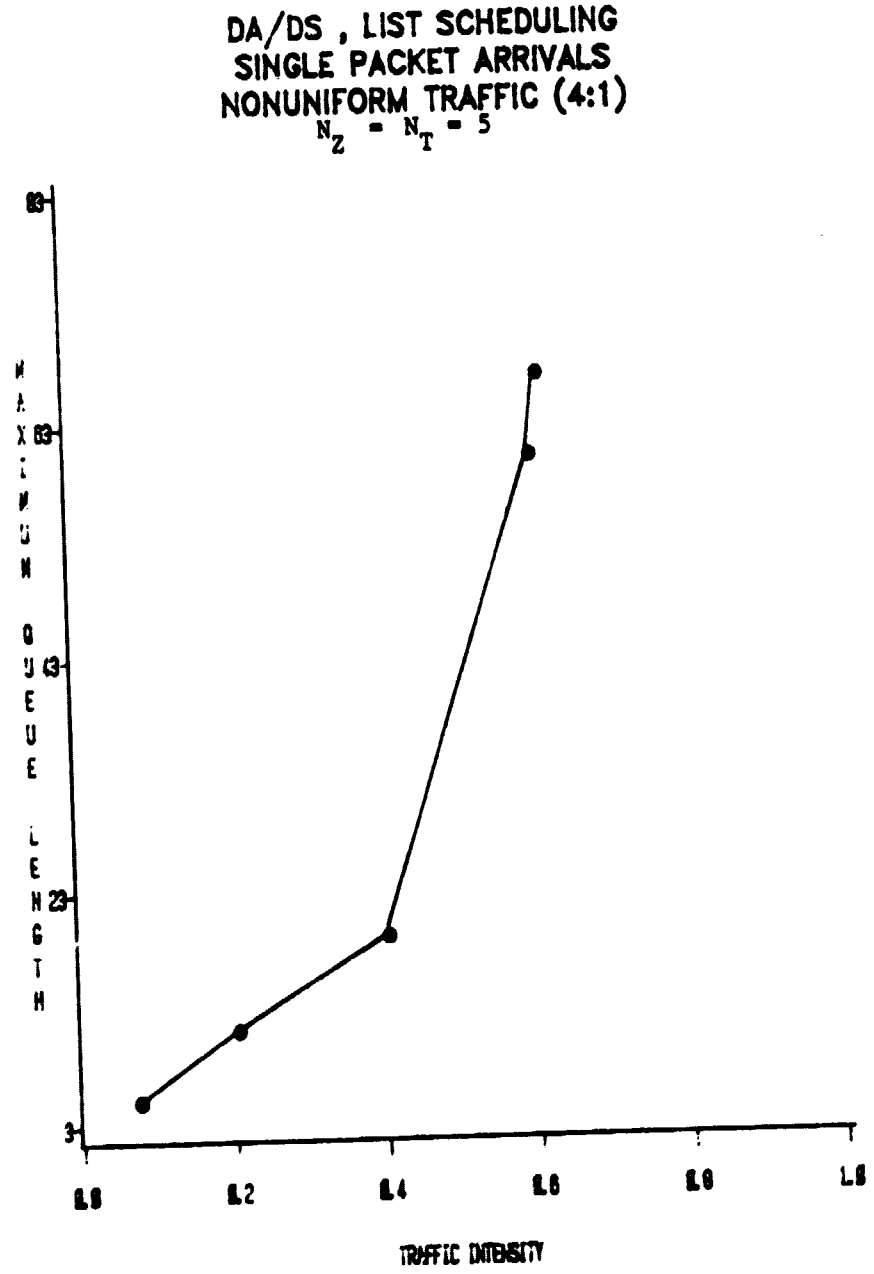
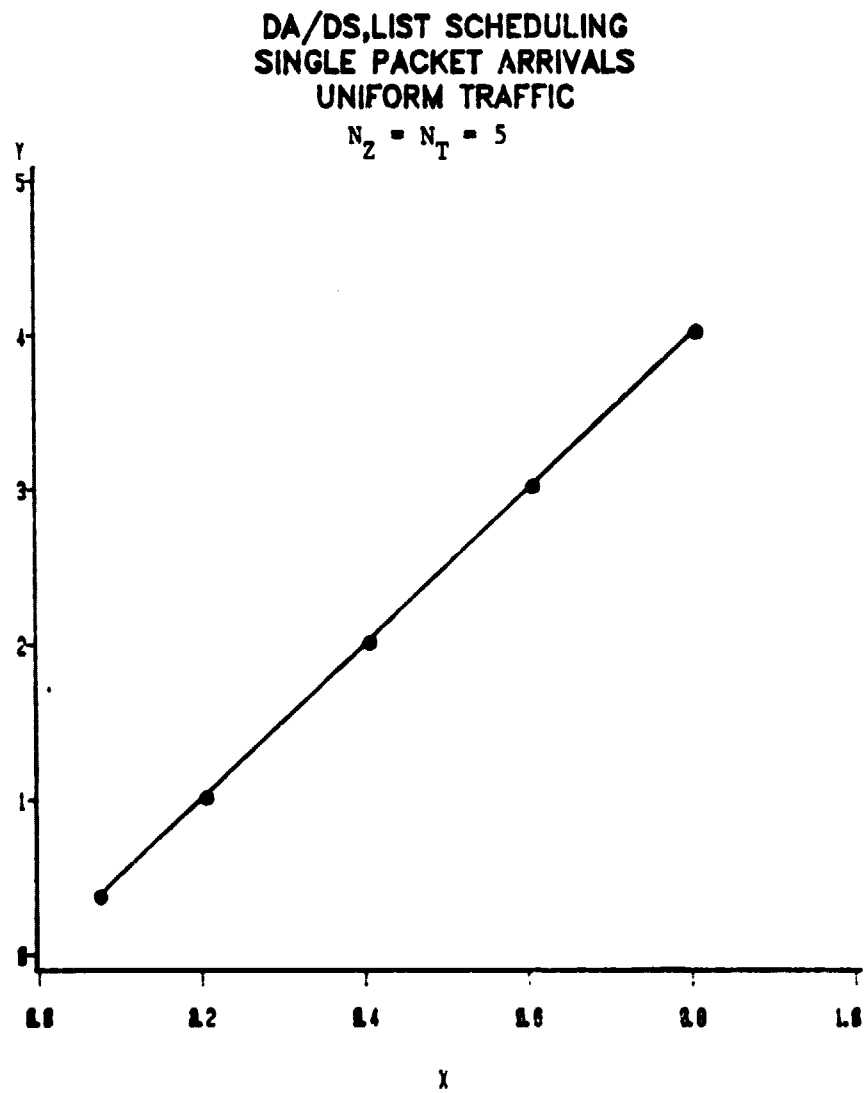
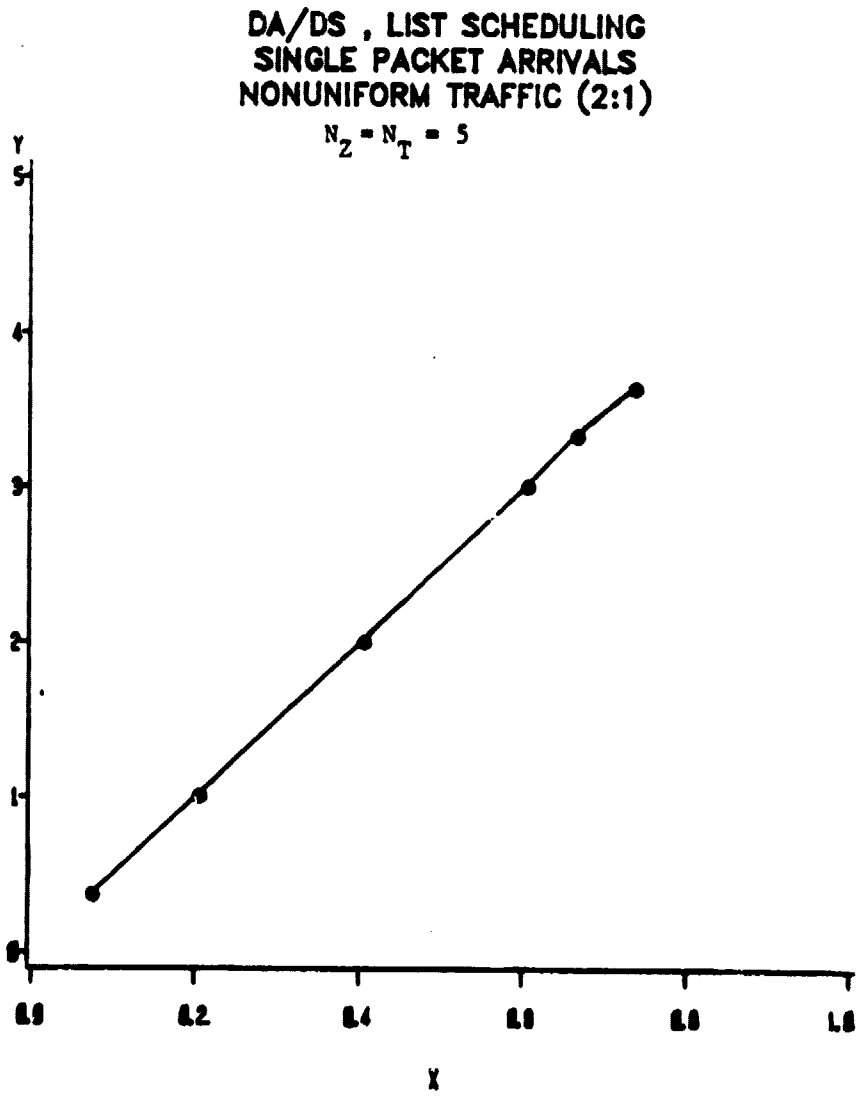


Fig. 2.29 Maximum number of waiting packets (Q_{max}) for nonuniform traffic (4:1).



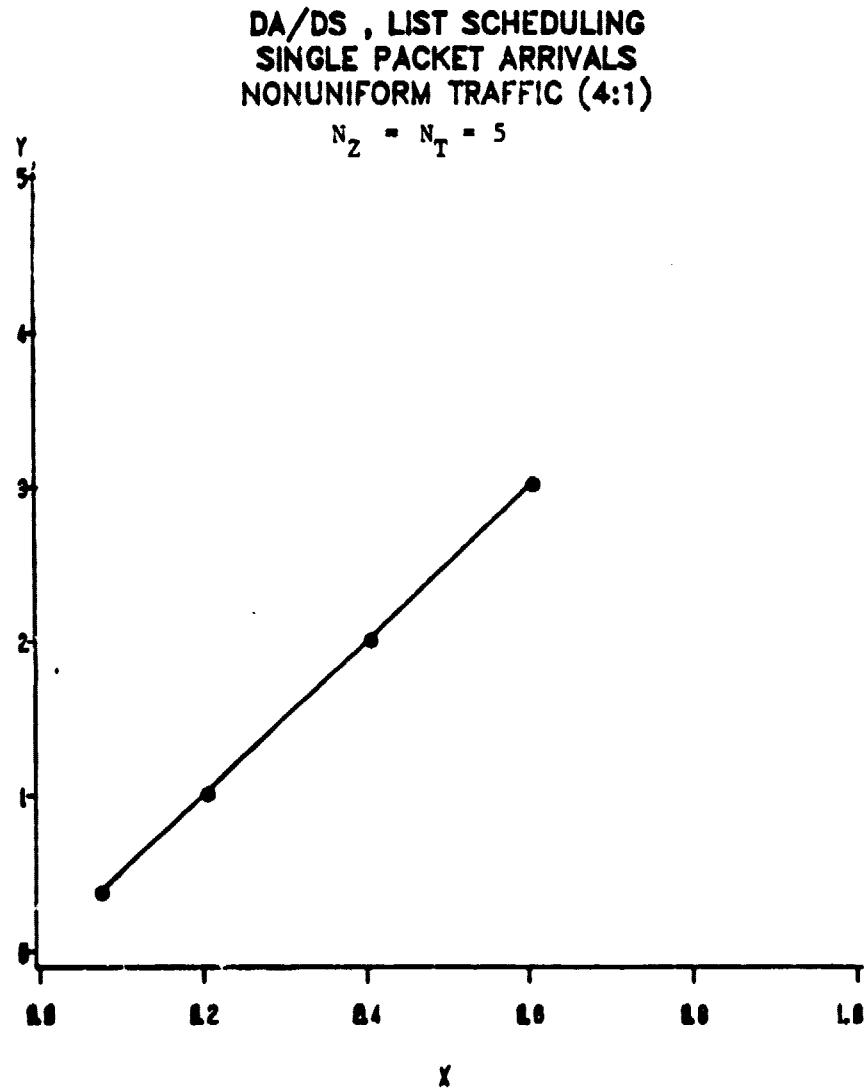
X=Traffic intensity
Y=Average number of transponders used, $\overline{N}$

Fig. 2.30 **Average no. of transponders used for uniform traffic.**



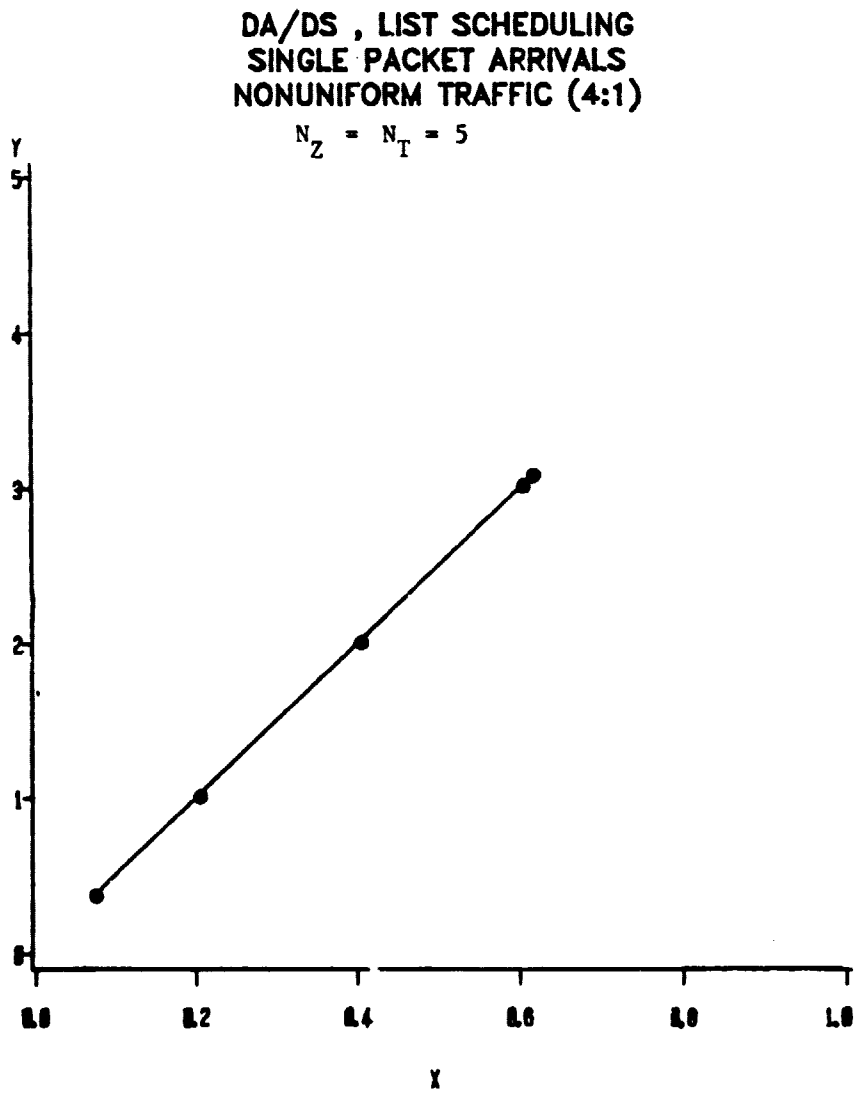
X=Traffic intensity
Y=Average number of transponders used, $\bar{N}$

Fig. 2.31 **Average number of transponders used for nonuniform traffic (2:1).**



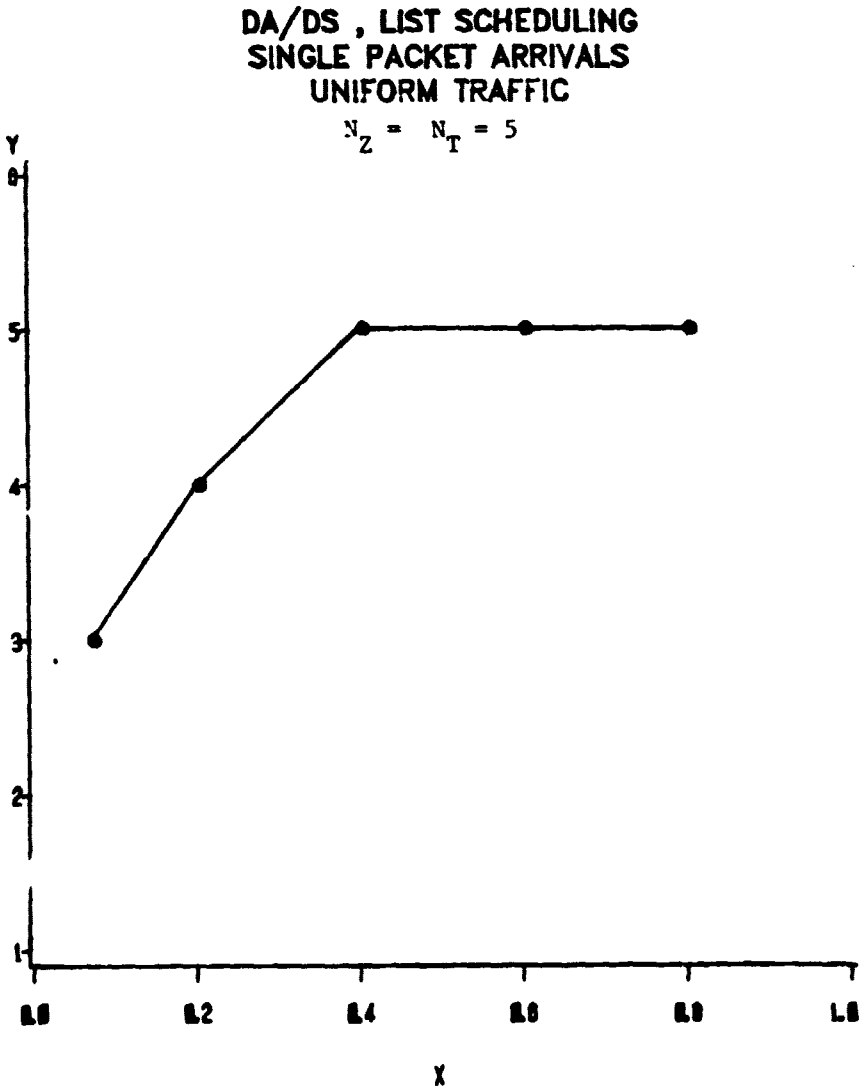
X=Traffic intensity
Y=Average number of transponders used, $\bar{N}$

Fig. 2.32 Average number of transponders used for nonuniform traffic (4:1).



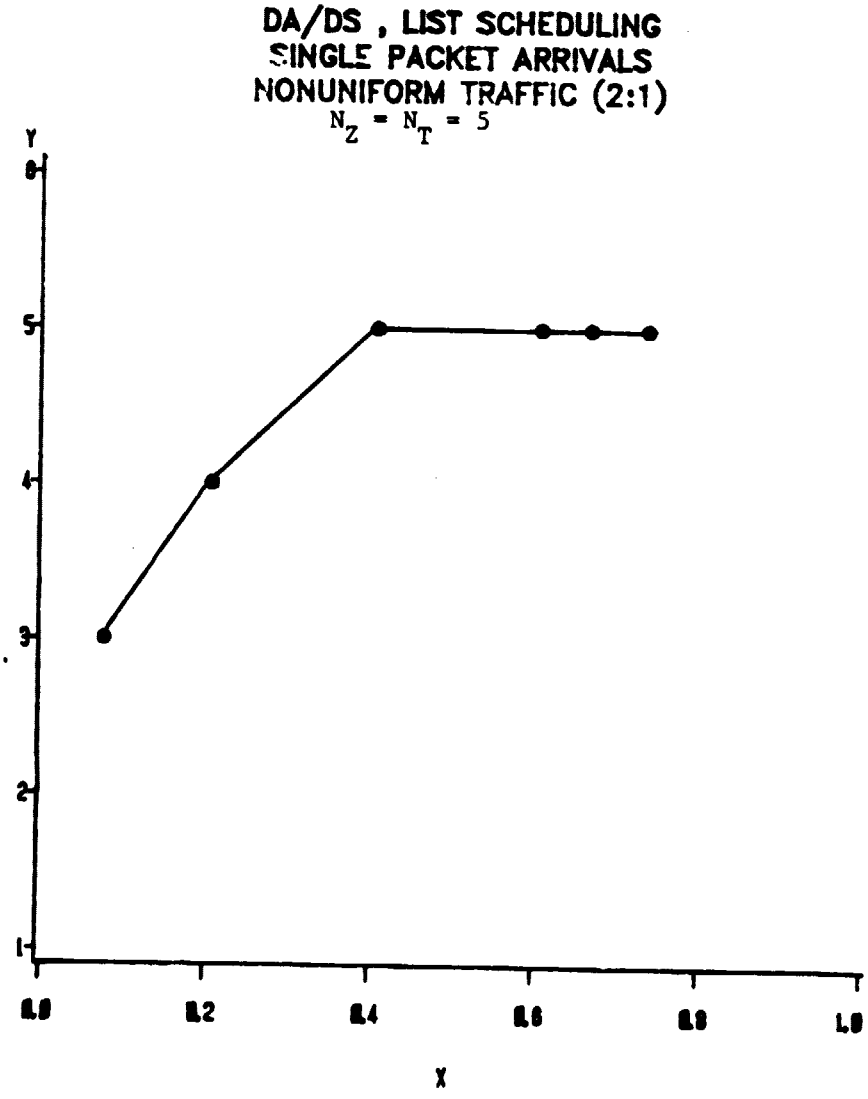
X=Traffic intensity
Y=Average number of transponders used, $\overline{N}$

Fig. 2.33 Average number of transponders used for nonuniform traffic (4:1).



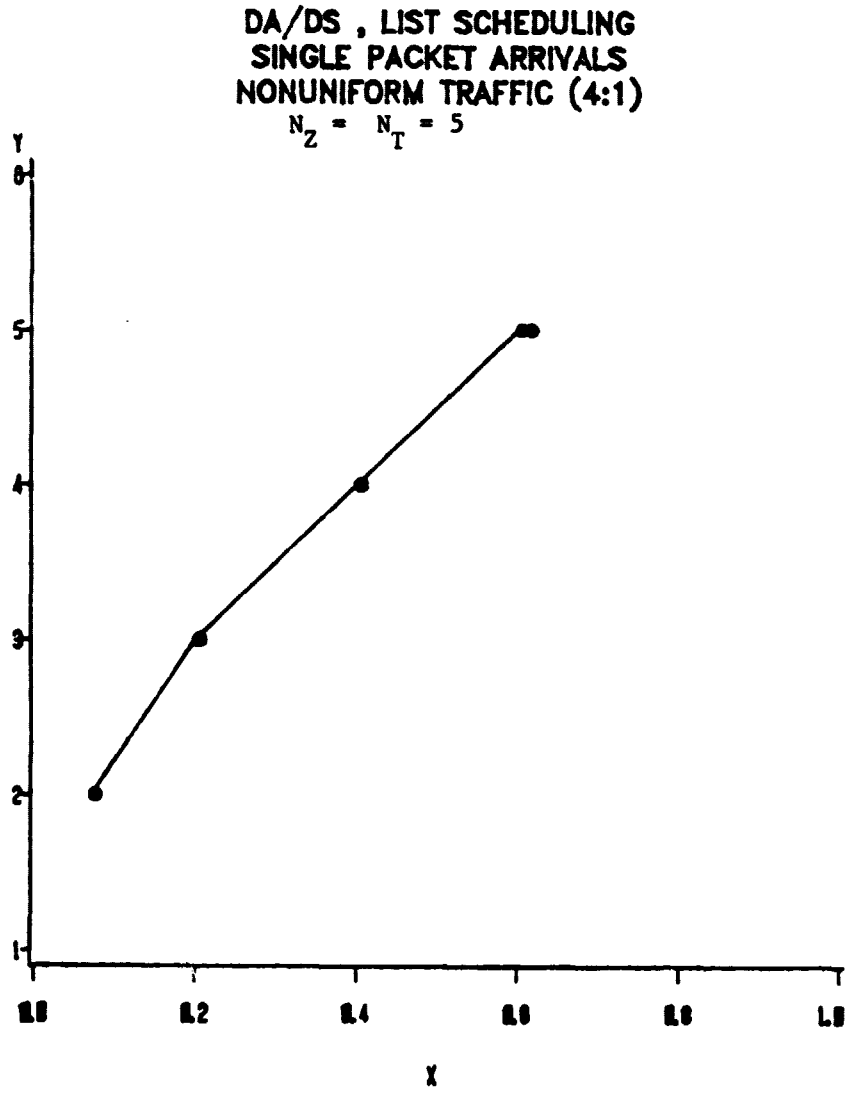
X=Traffic intensity
Y=Maximum number of transponders used, N_{max}

Fig. 2.34 Maximum number of transponders used for uniform traffic.



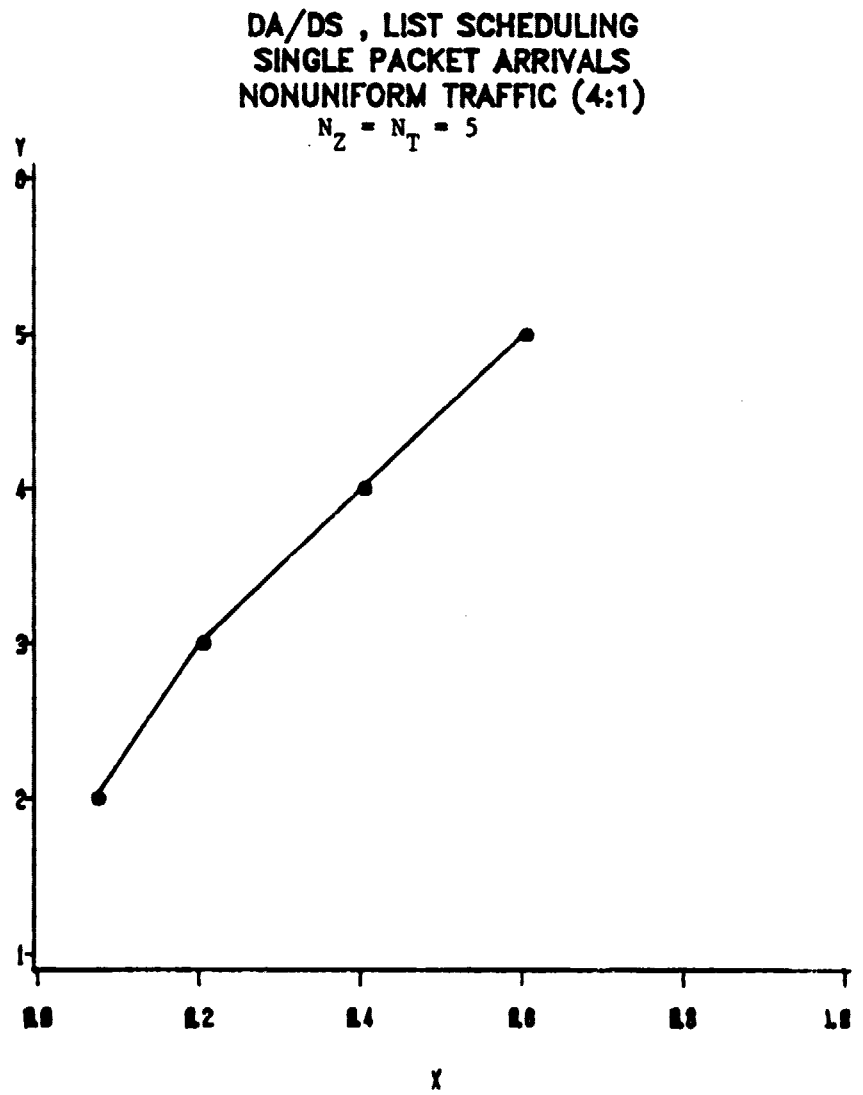
X=Traffic intensity
Y=Maximum number of transponders used, N_{max}

Fig. 2.35 Maximum number of transponders used for nonuniform traffic (2:1).



X=Traffic intensity
Y=Maximum number of transponders used, N_{max}

Fig. 2.36 **Maximum number of transponders used for nonuniform traffic (4:1).**



X=Traffic intensity
Y=Maximum number of transponders used, N_{max}

Fig. 2.37 **Maximum number of transponders used for nonuniform traffic (4:1).**

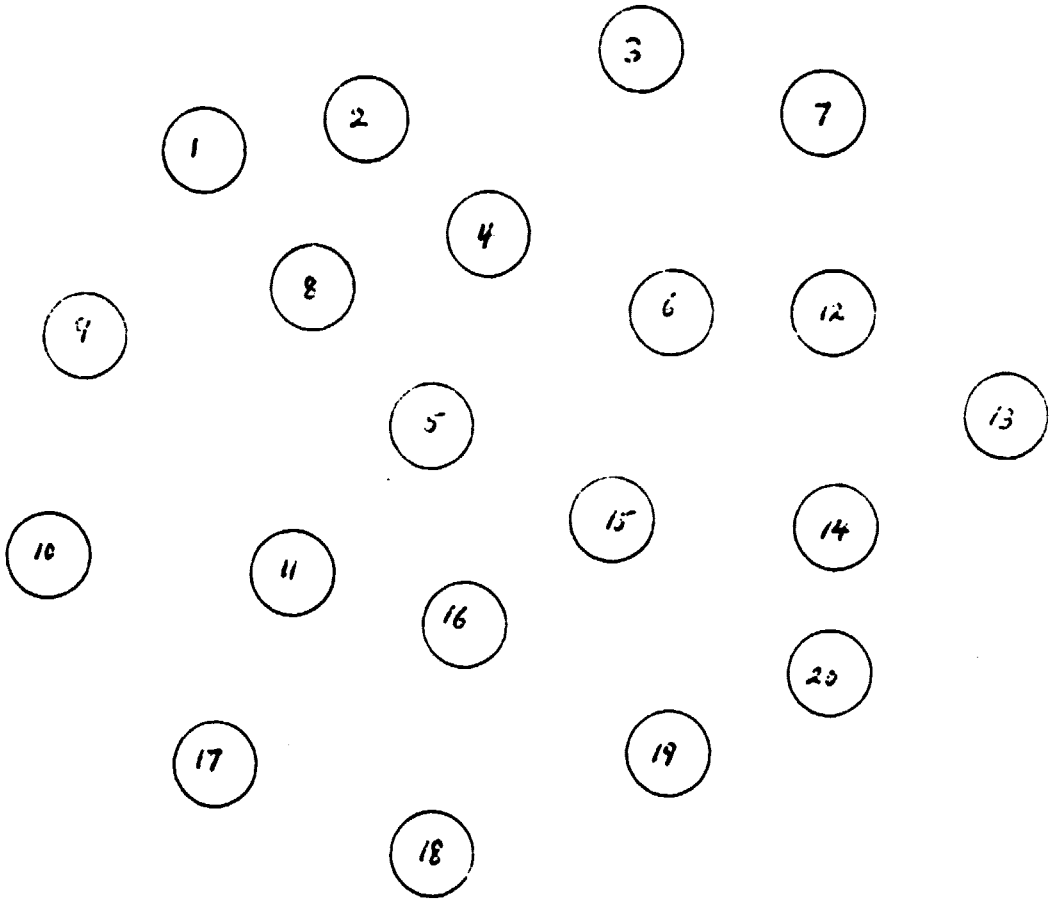
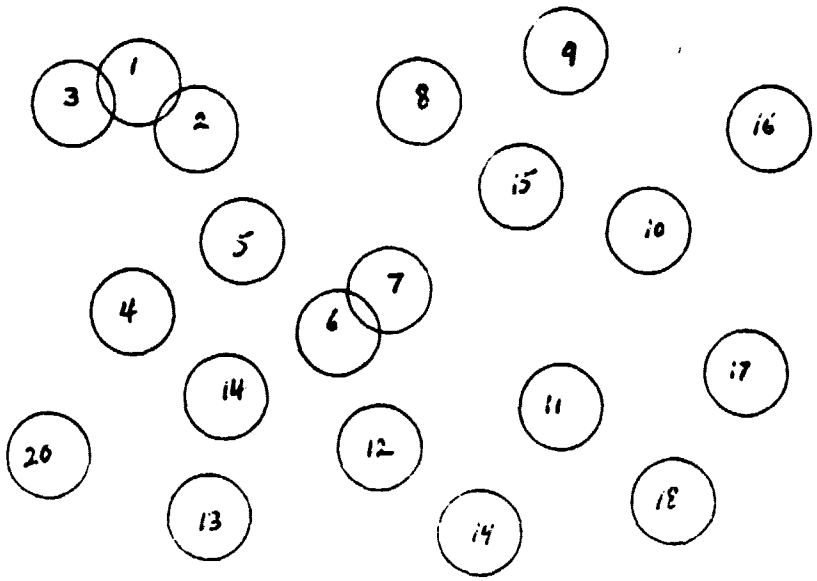
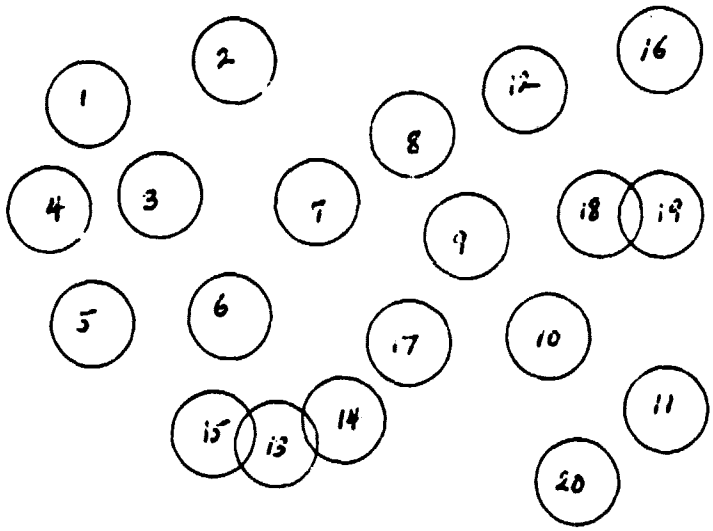


Fig. 2.38 Nonoverlapping Zones



(a) Overlapping High Probability Zones



(b) Overlapping Low Probability Zones

Fig. 2.39 Overlapping Zones

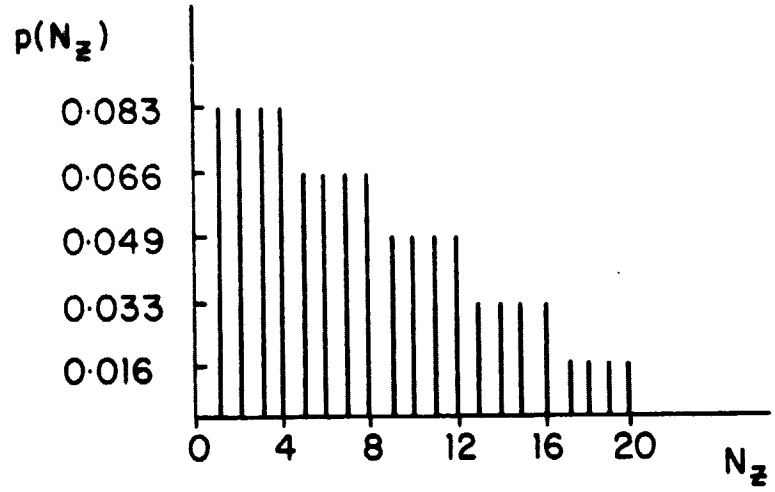


Fig. 2.40 Traffic Distribution for $N_z = 20$

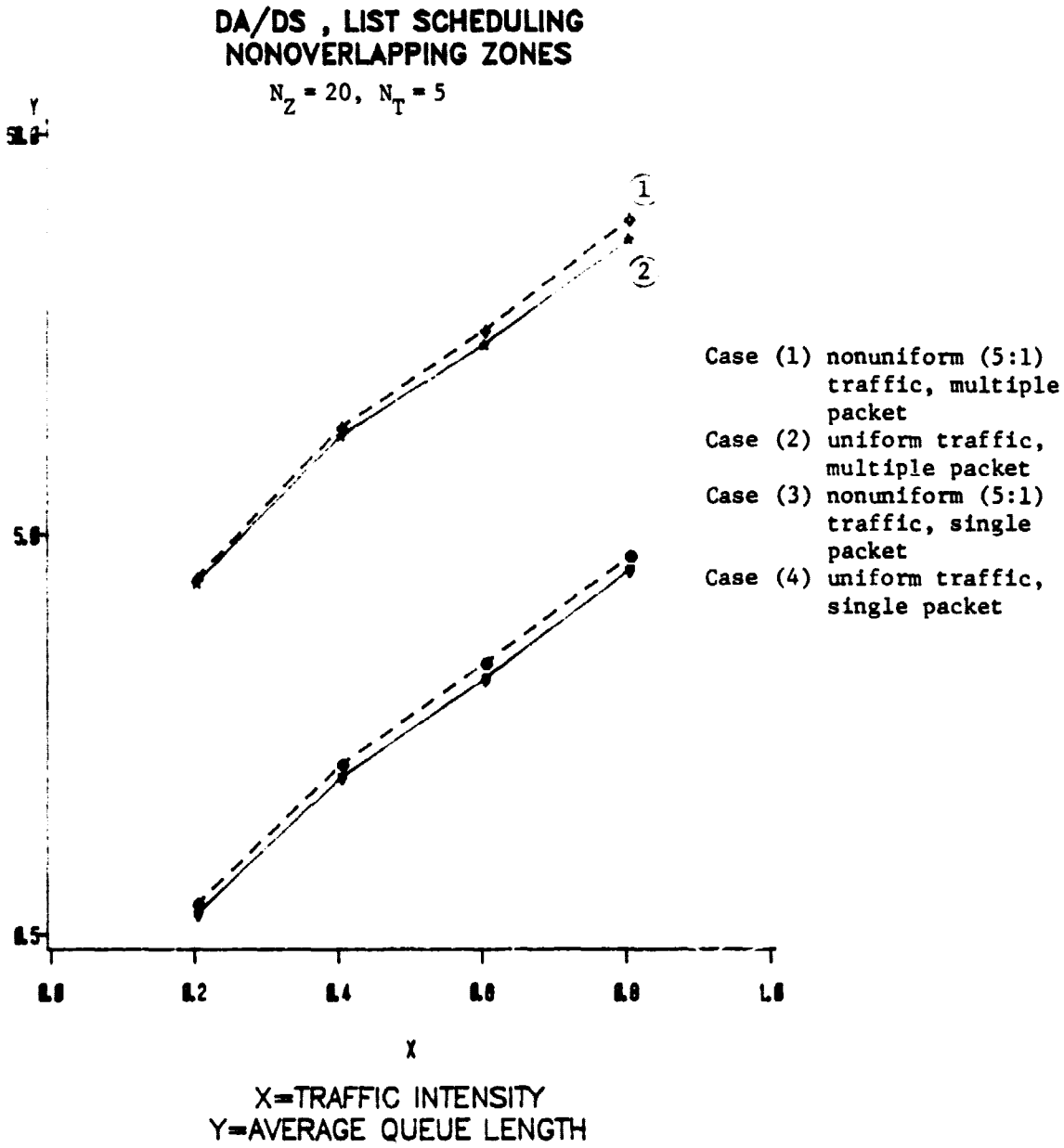


Fig. 2.41 Comparison of average number of waiting packets for different traffic distributions and message lengths.

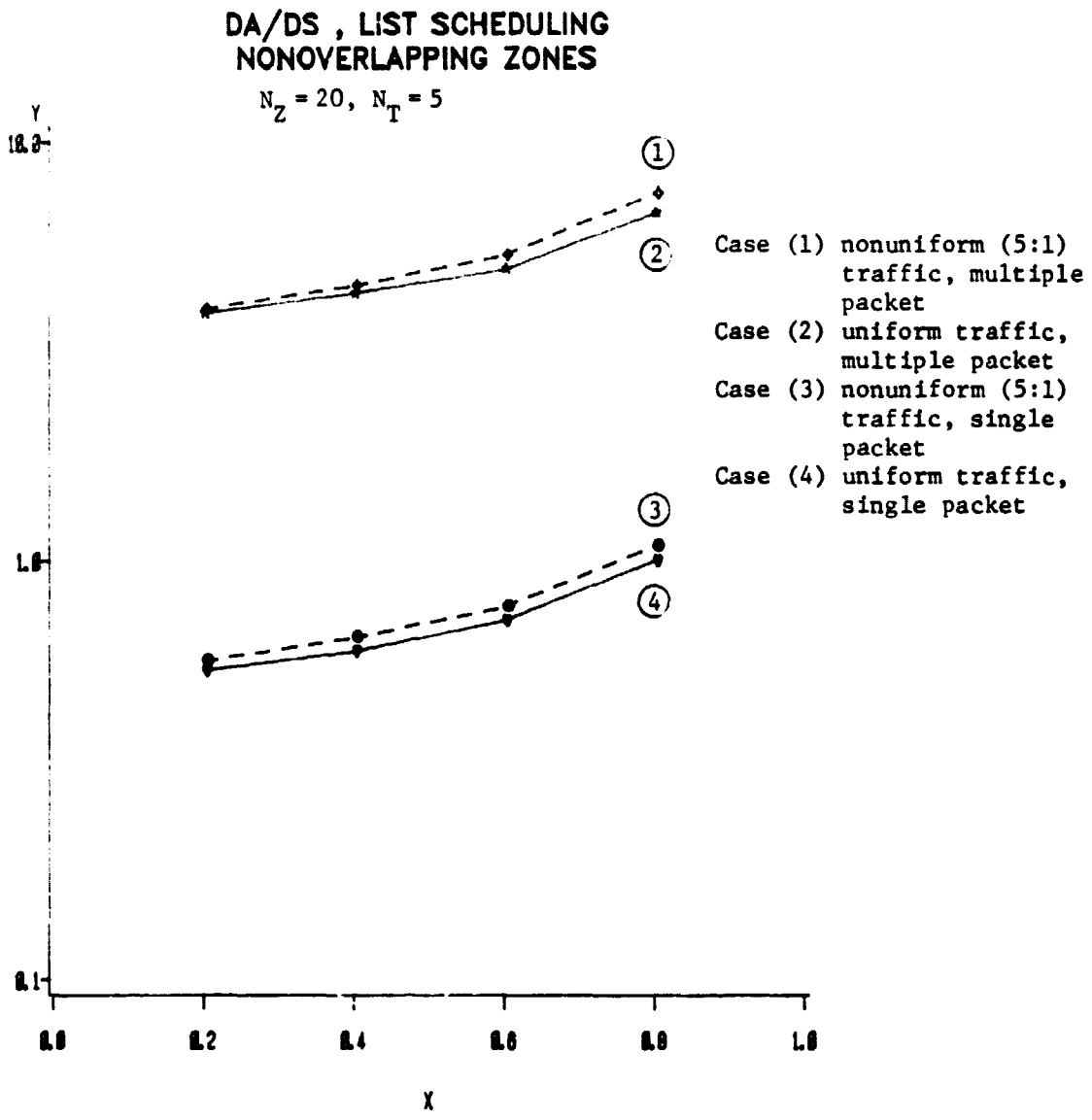


Fig. 2.42 Comparison of waiting time for different traffic distributions and message lengths.

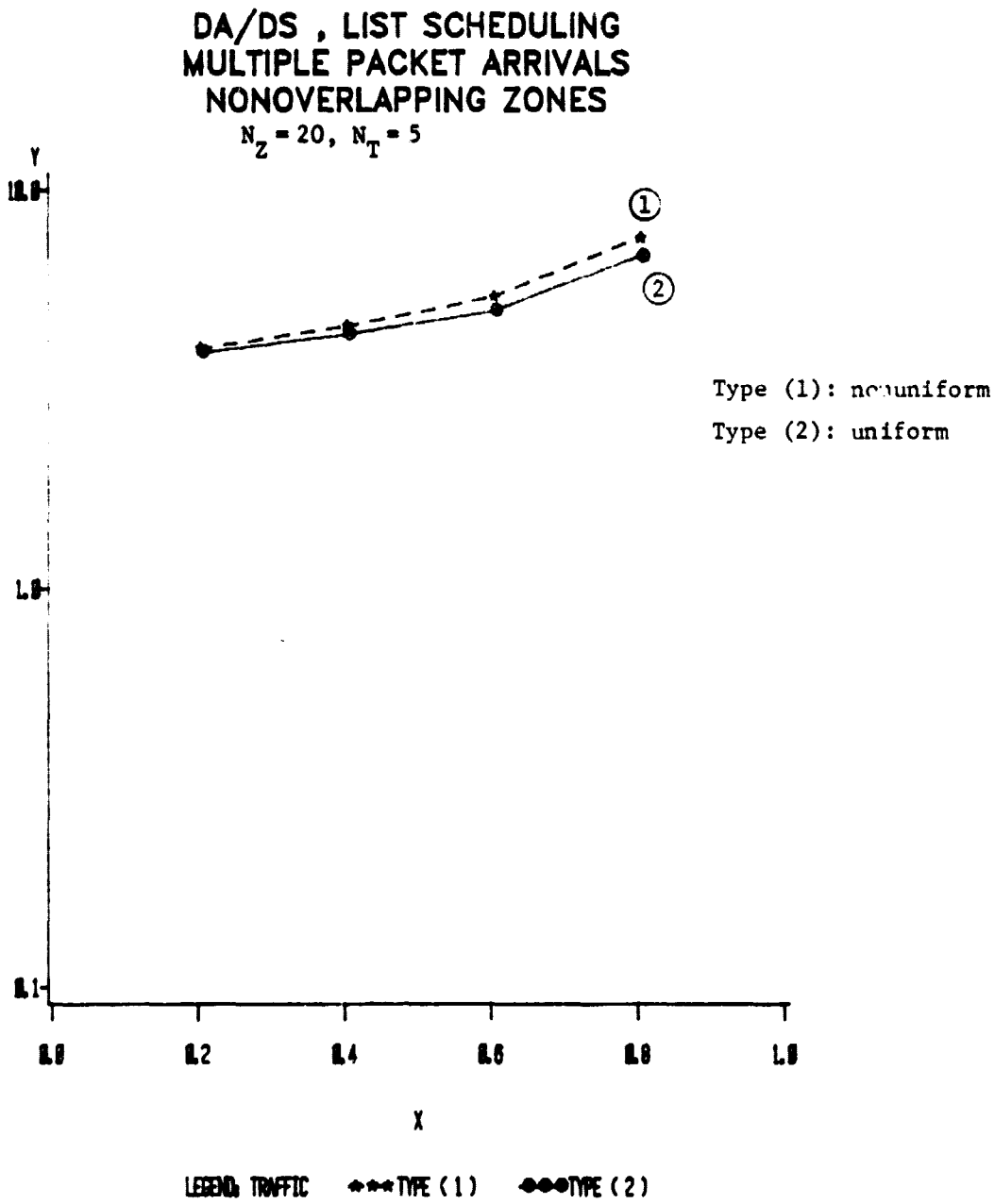


Fig. 2.43 Comparison of waiting time for different traffic distributions

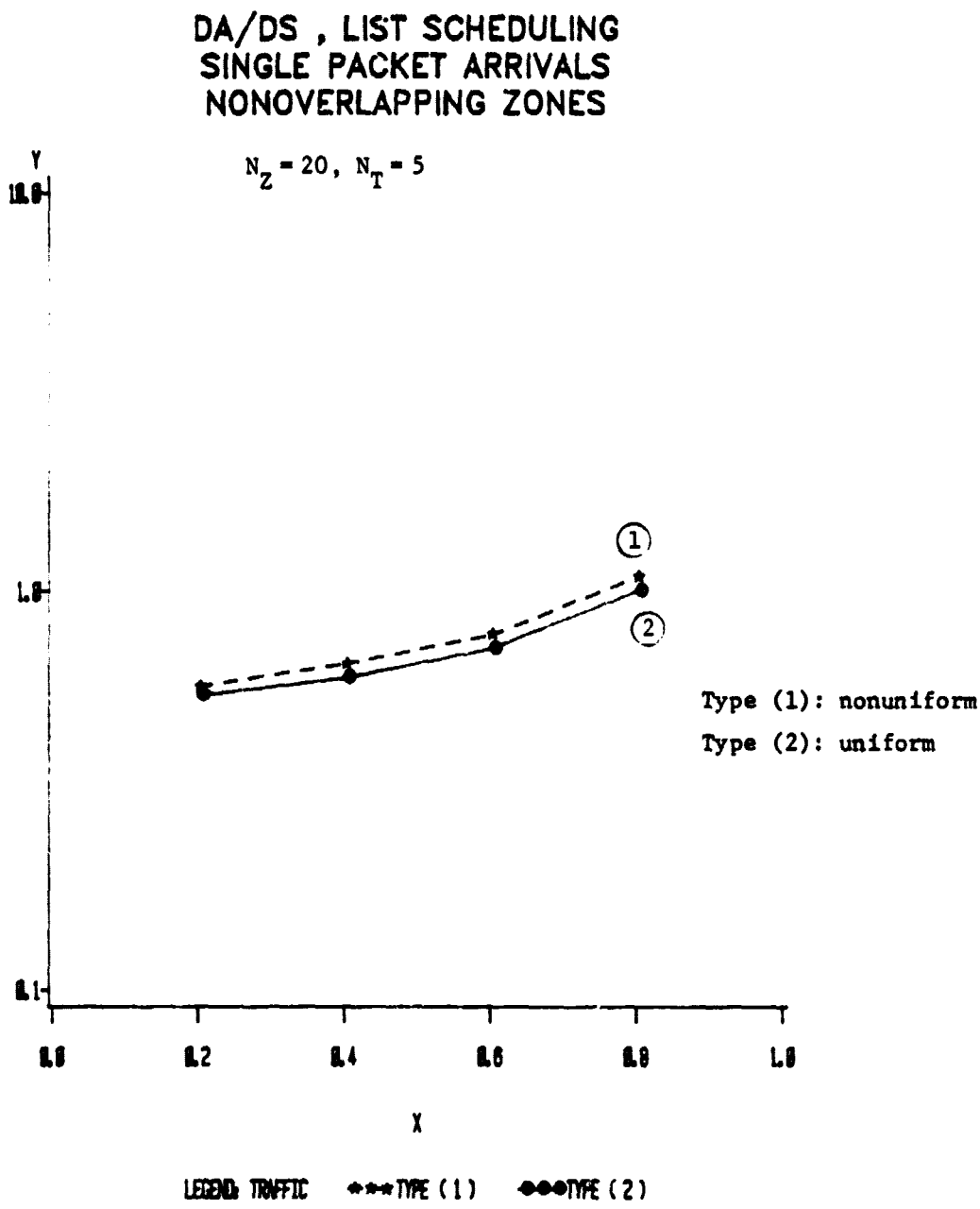
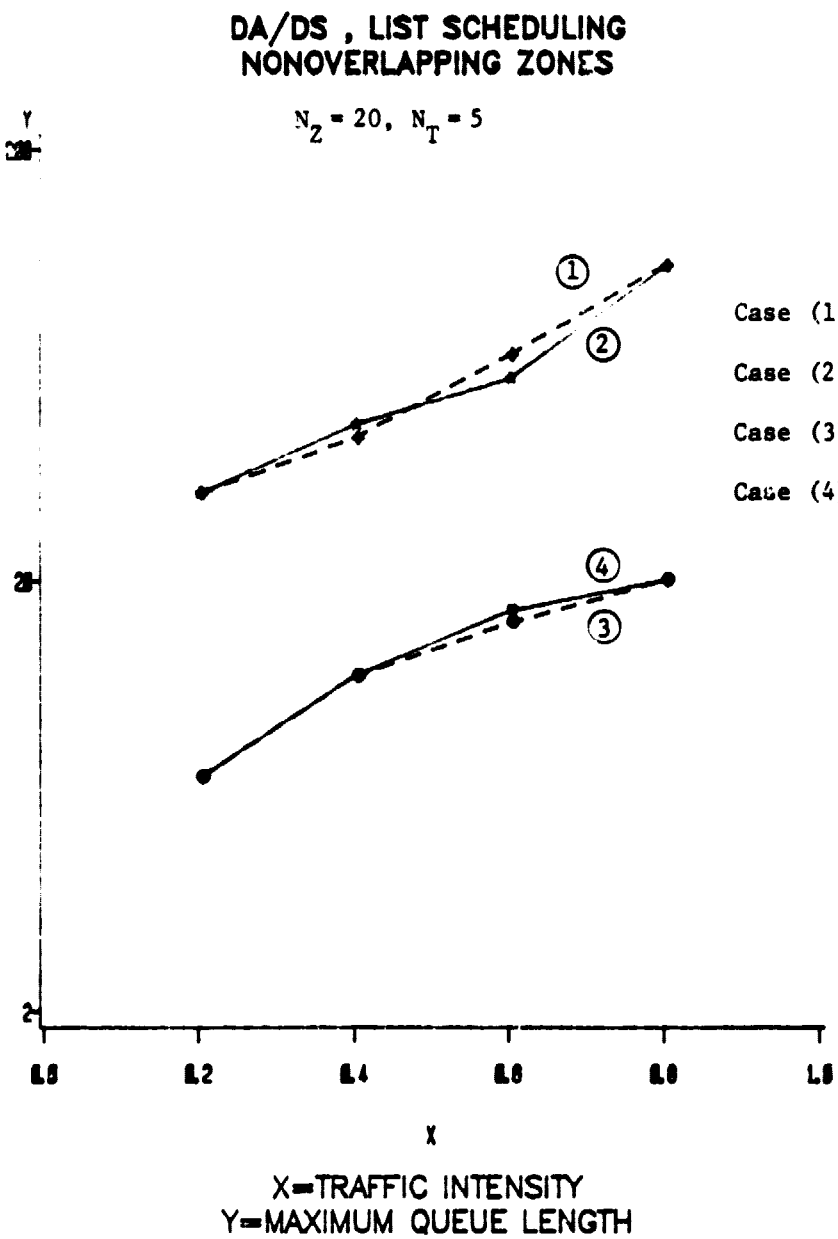


Fig. 2.44 Comparison of waiting time for different traffic distributions.



Case (1) nonuniform (5:1) traffic, multiple packet
Case (2) uniform traffic, multiple packet
Case (3) nonuniform (5:1) traffic single packet
Case (4) uniform traffic, single packet.

Fig. 2.45 Comparison of maximum number of waiting packets for different traffic distributions and message lengths.

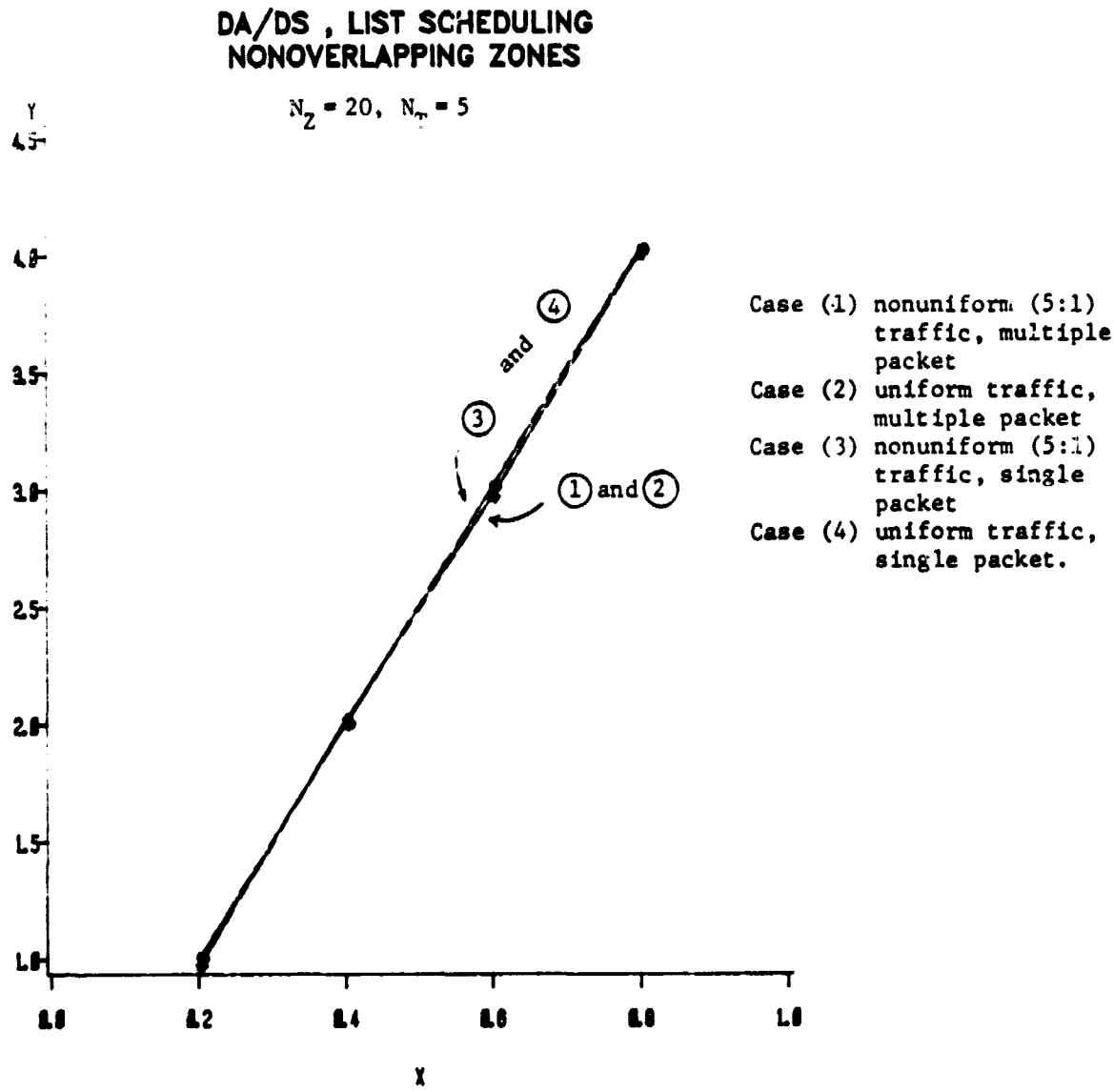


Fig. 2.46 Comparison of average number of transponders used, $\bar{N}$, for different traffic distributions and different message lengths.

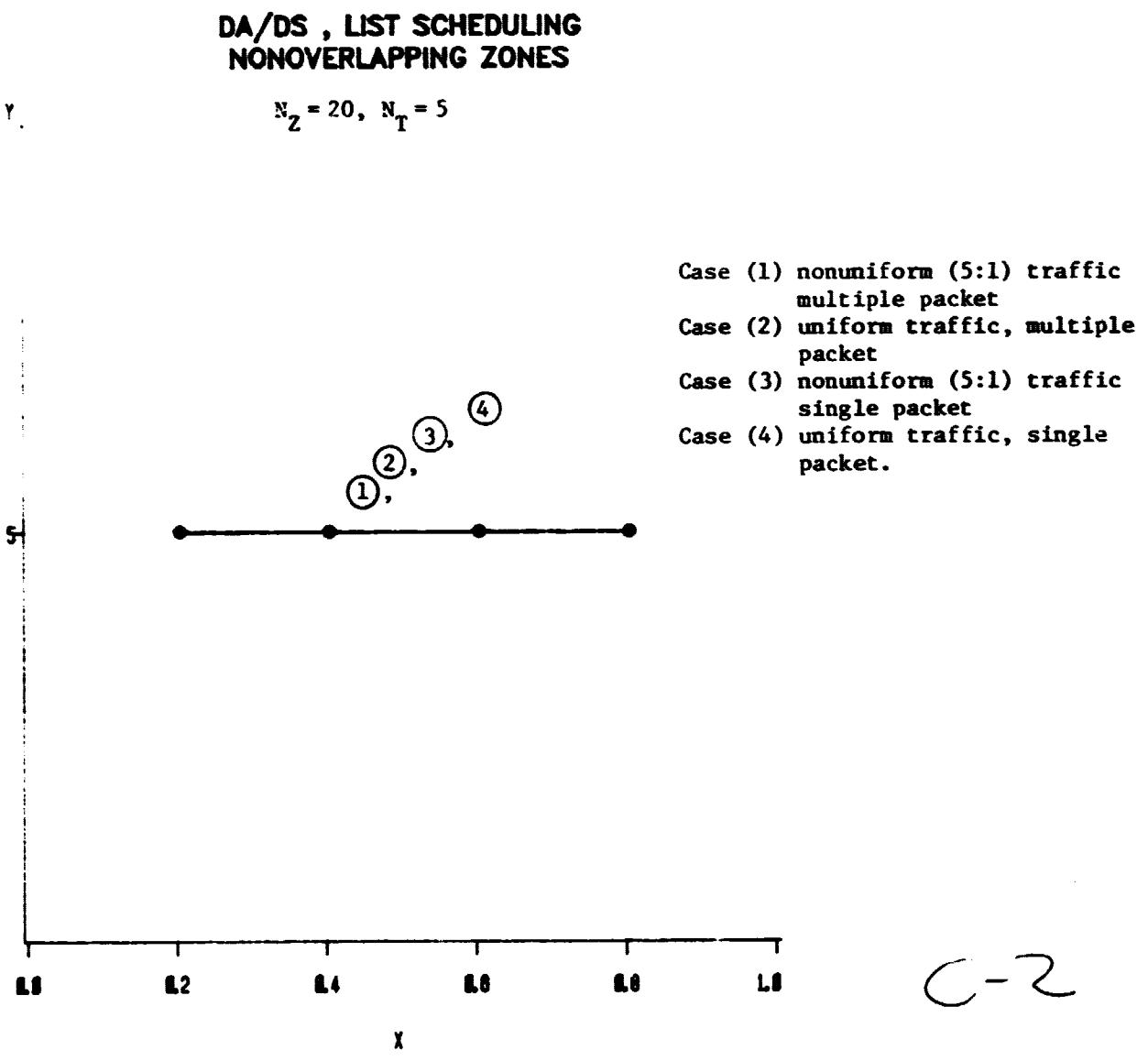


Fig. 2.47 Comparison of maximum number of transponders used for different traffic distributions and different message lengths.

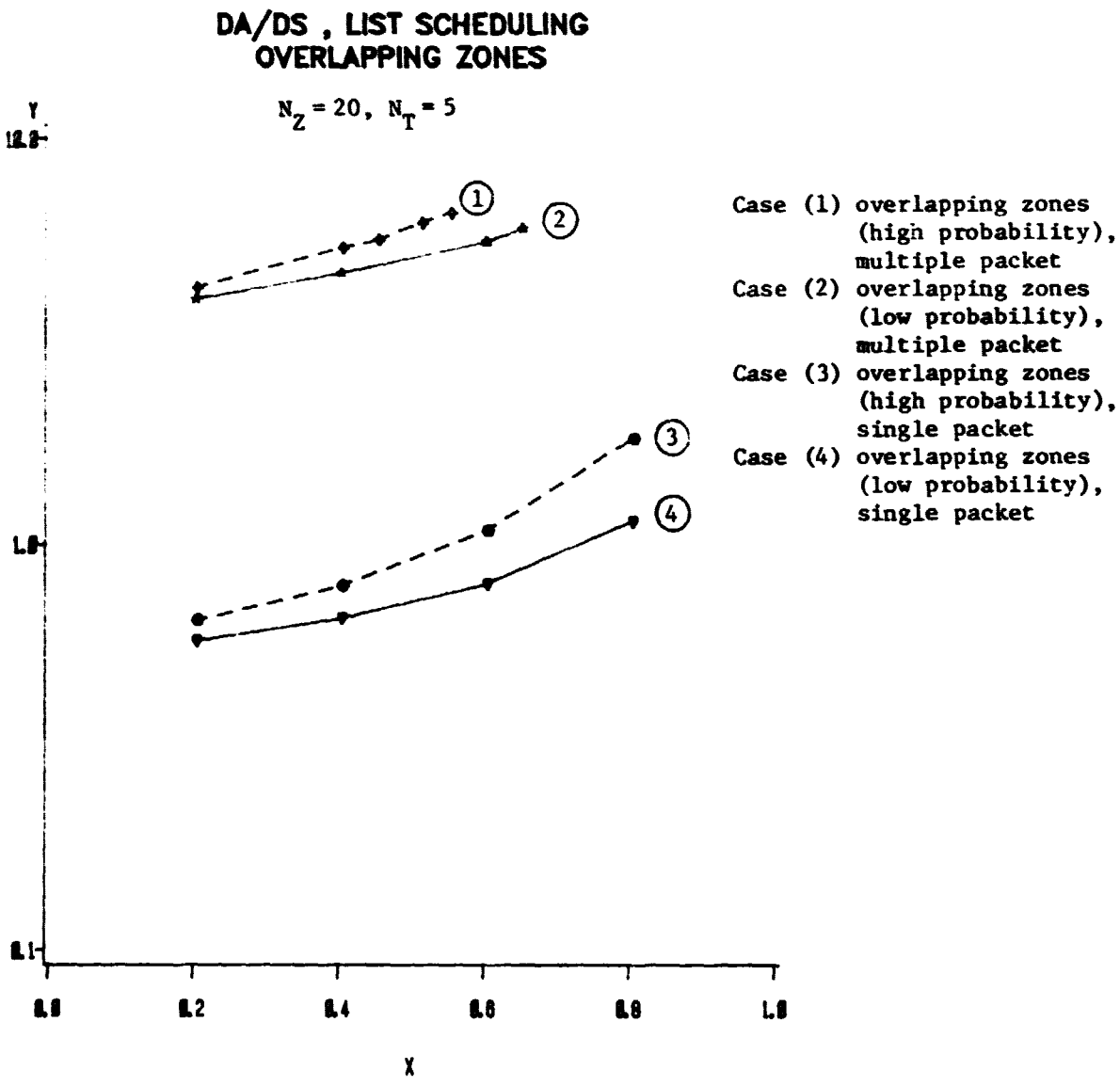


Fig. 2.48 Comparison of waiting time for different traffic distributions and message lengths.

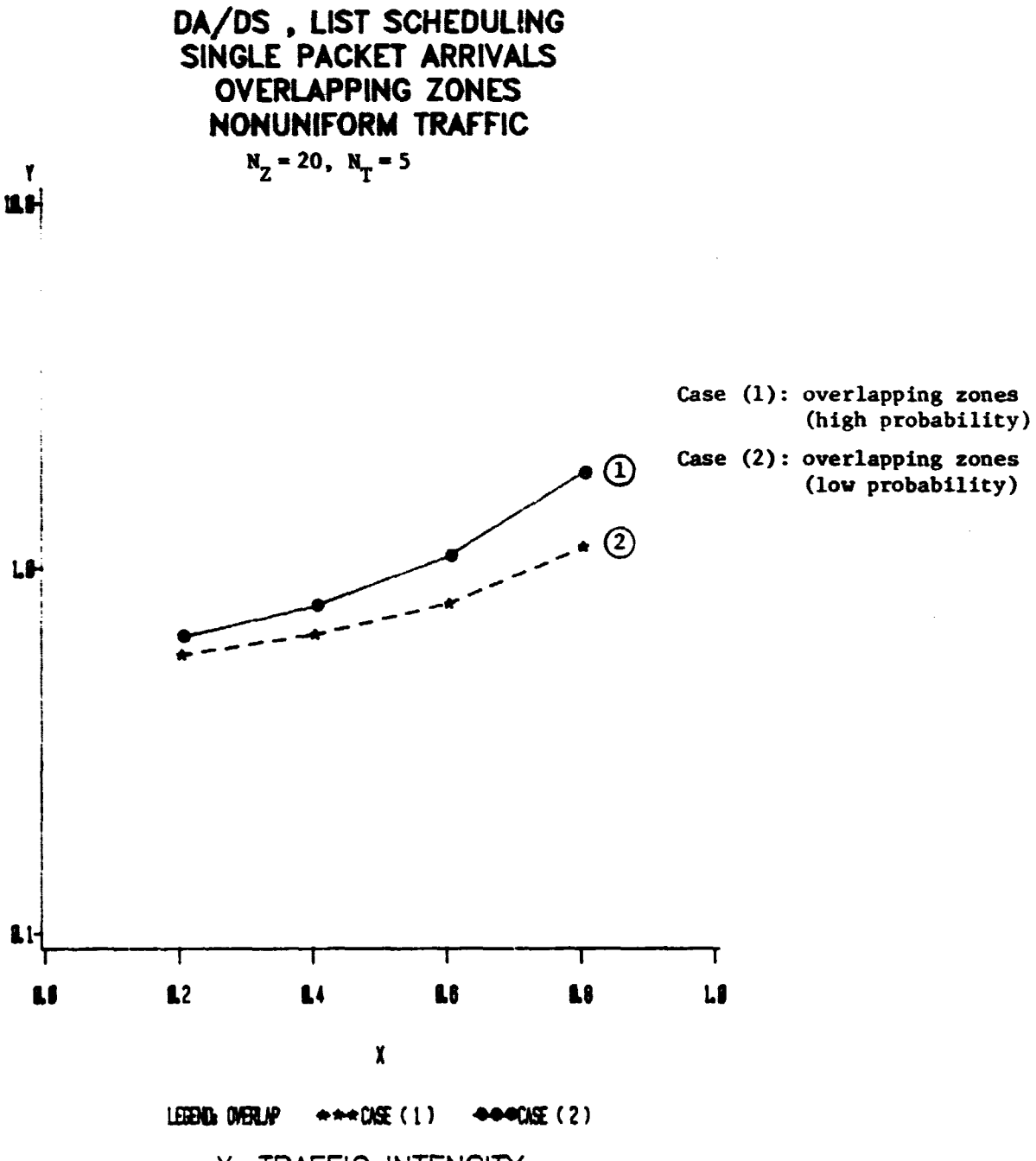


Fig. 2.49 Comparison of waiting time for different overlapping zones.

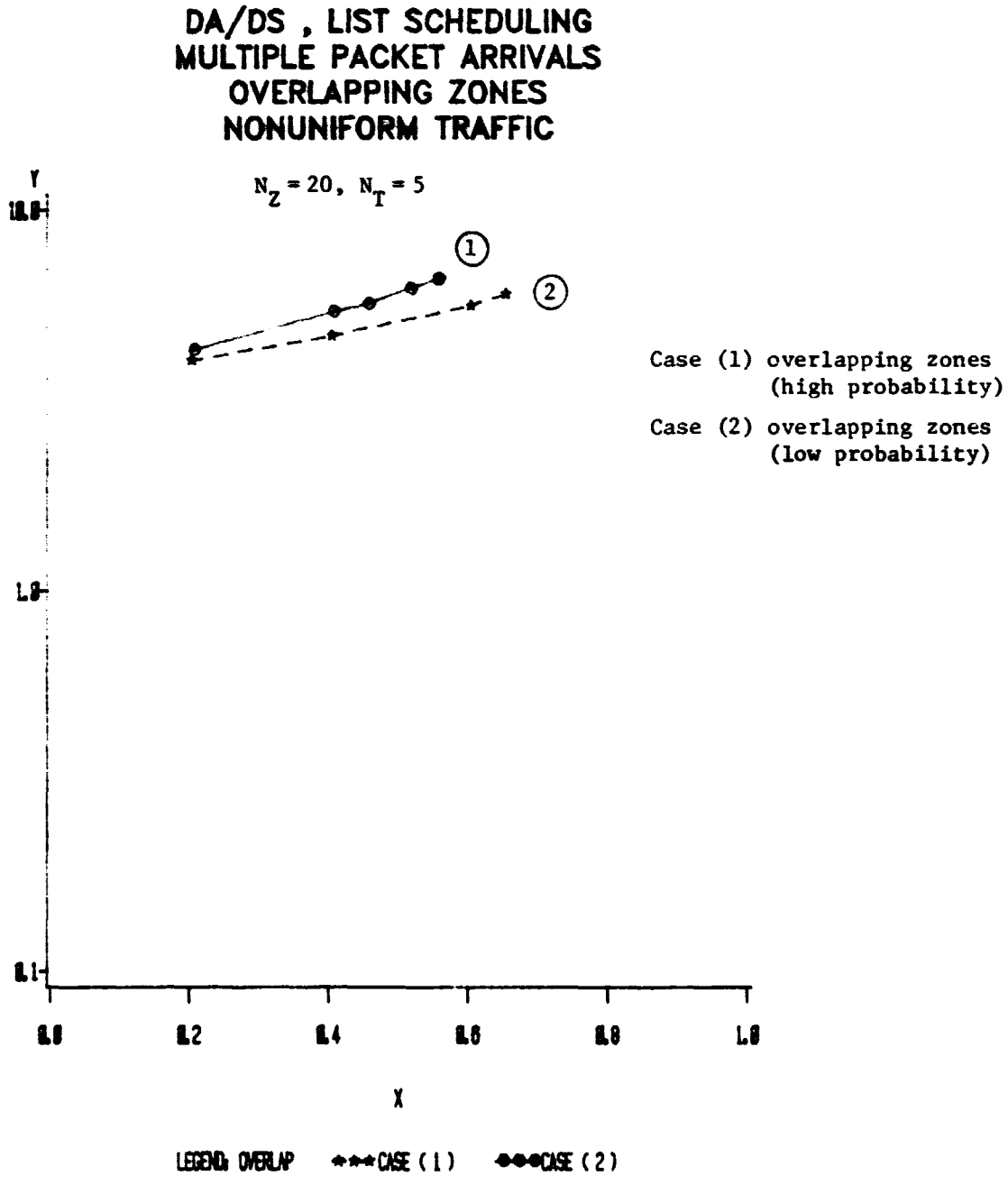


Fig. 2.50 Comparison of waiting time for different overlapping zones.

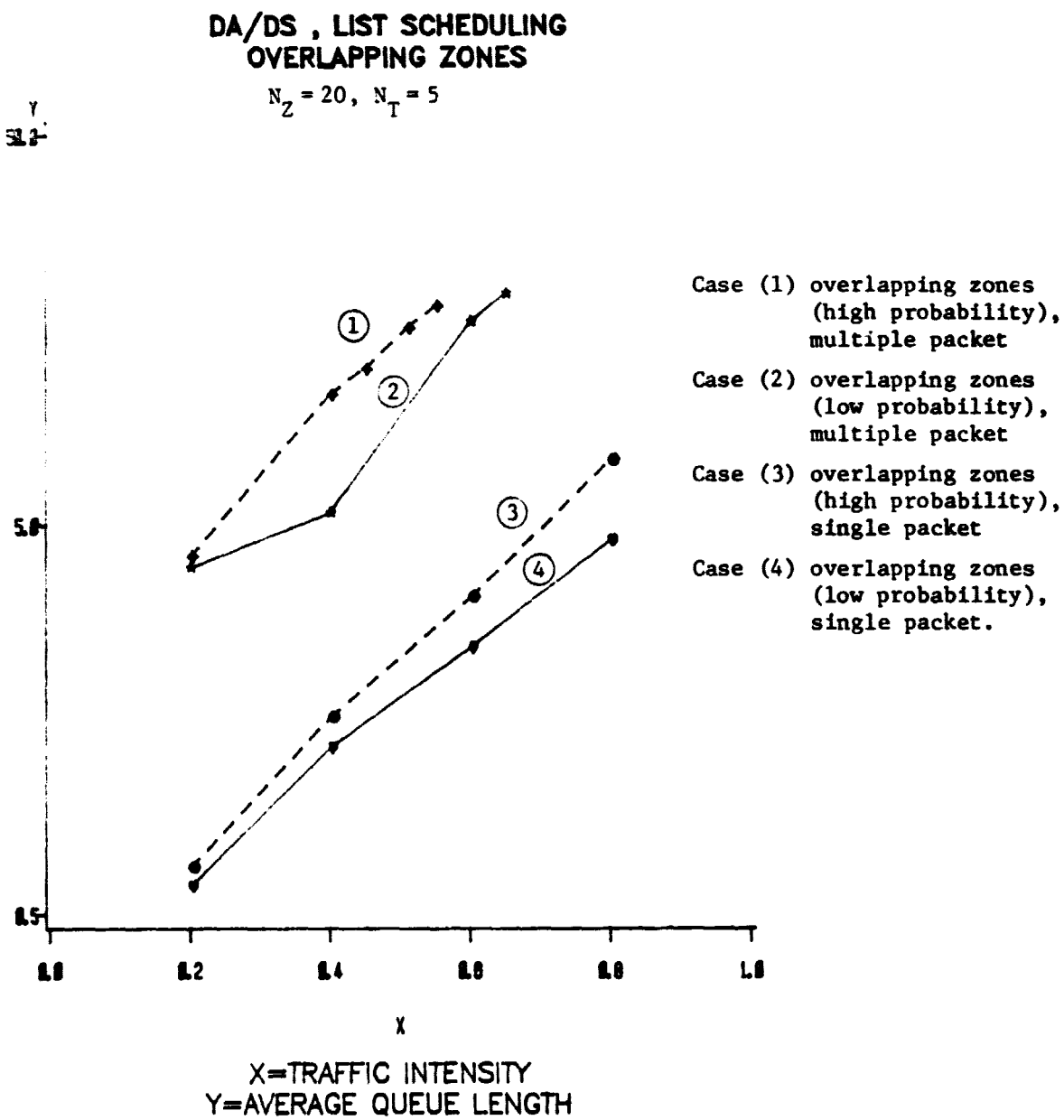


Fig. 2.51 Comparison of average number of waiting packets for different traffic distributions and message lengths.

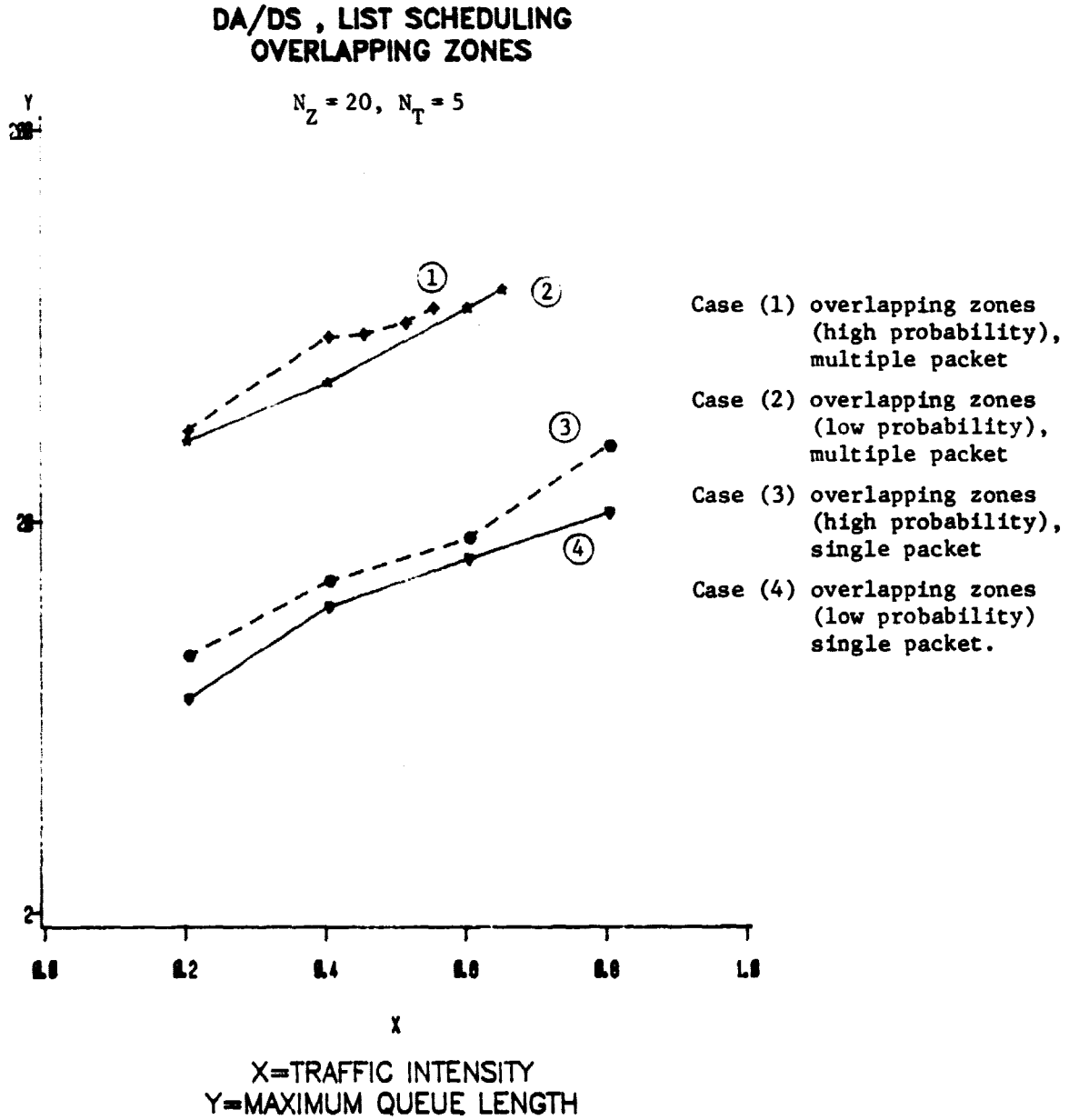


Fig. 2.52 Comparison of maximum number of waiting packets (Q_{max}) for different traffic distributions and message lengths.

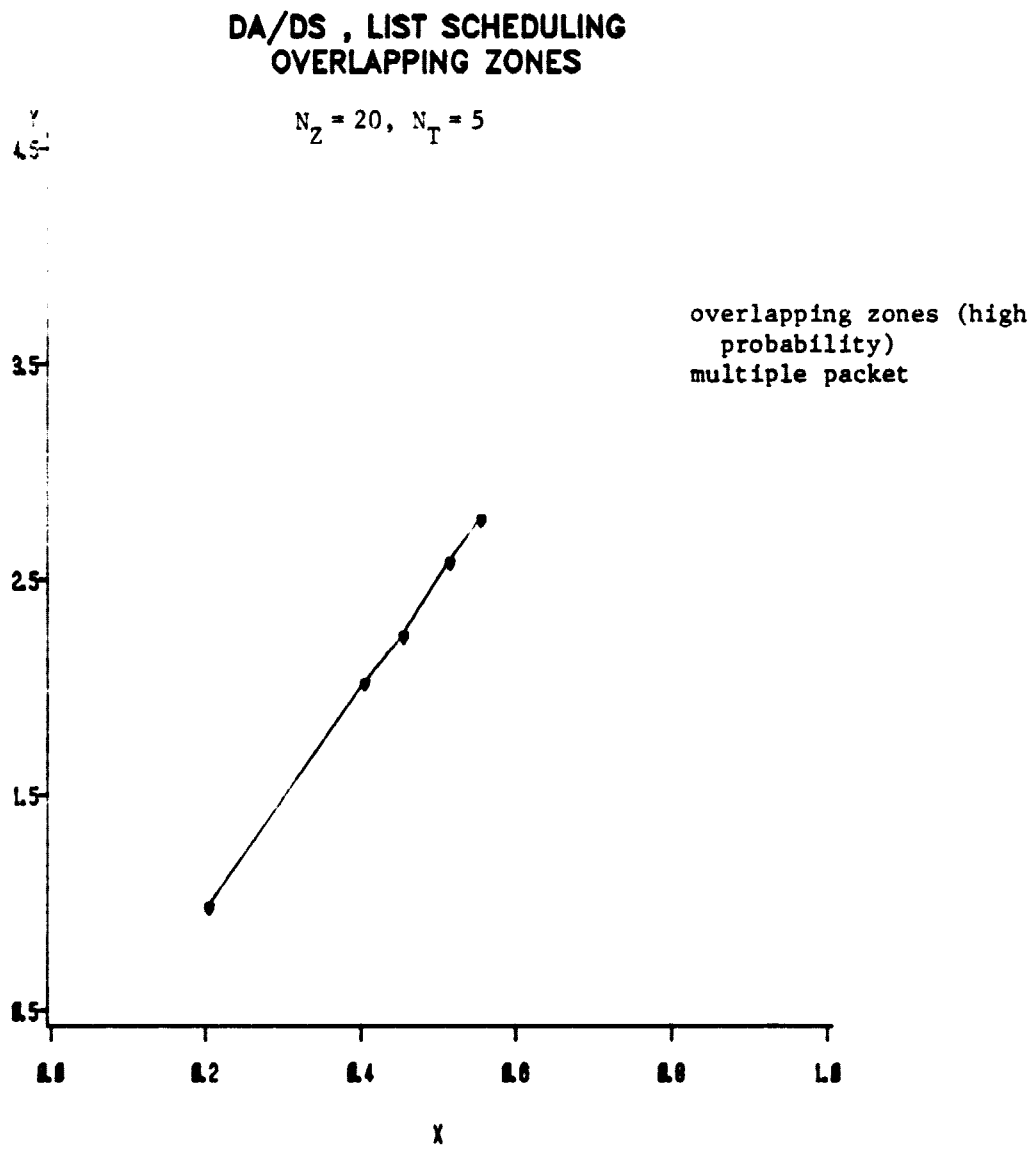


Fig. 2.53 Average number of transponders used for multipacket traffic from overlapping zones (high probability).

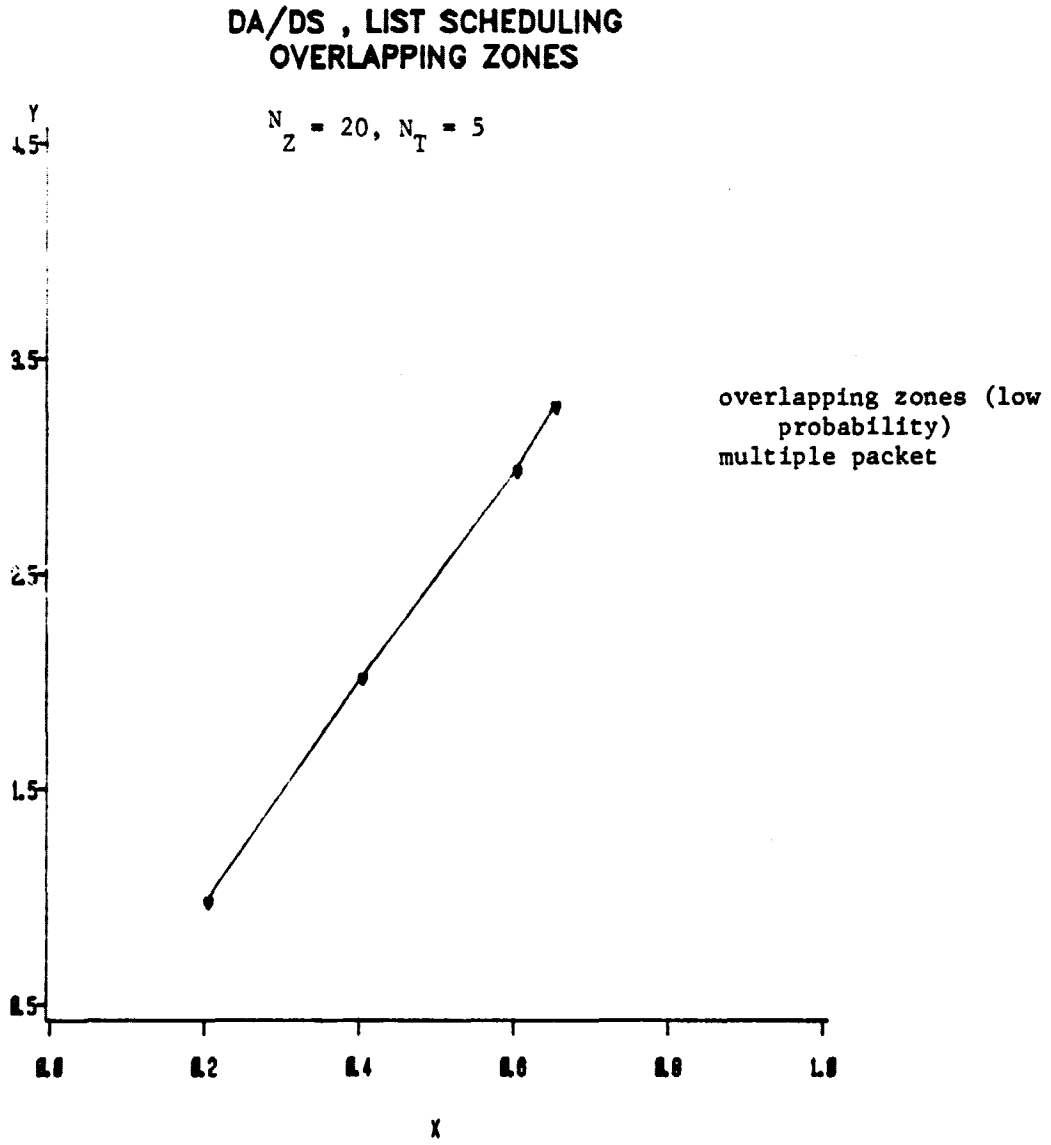


Fig. 2.54 Average number of transponders used for multipacket traffic from overlapping zones (low probability).

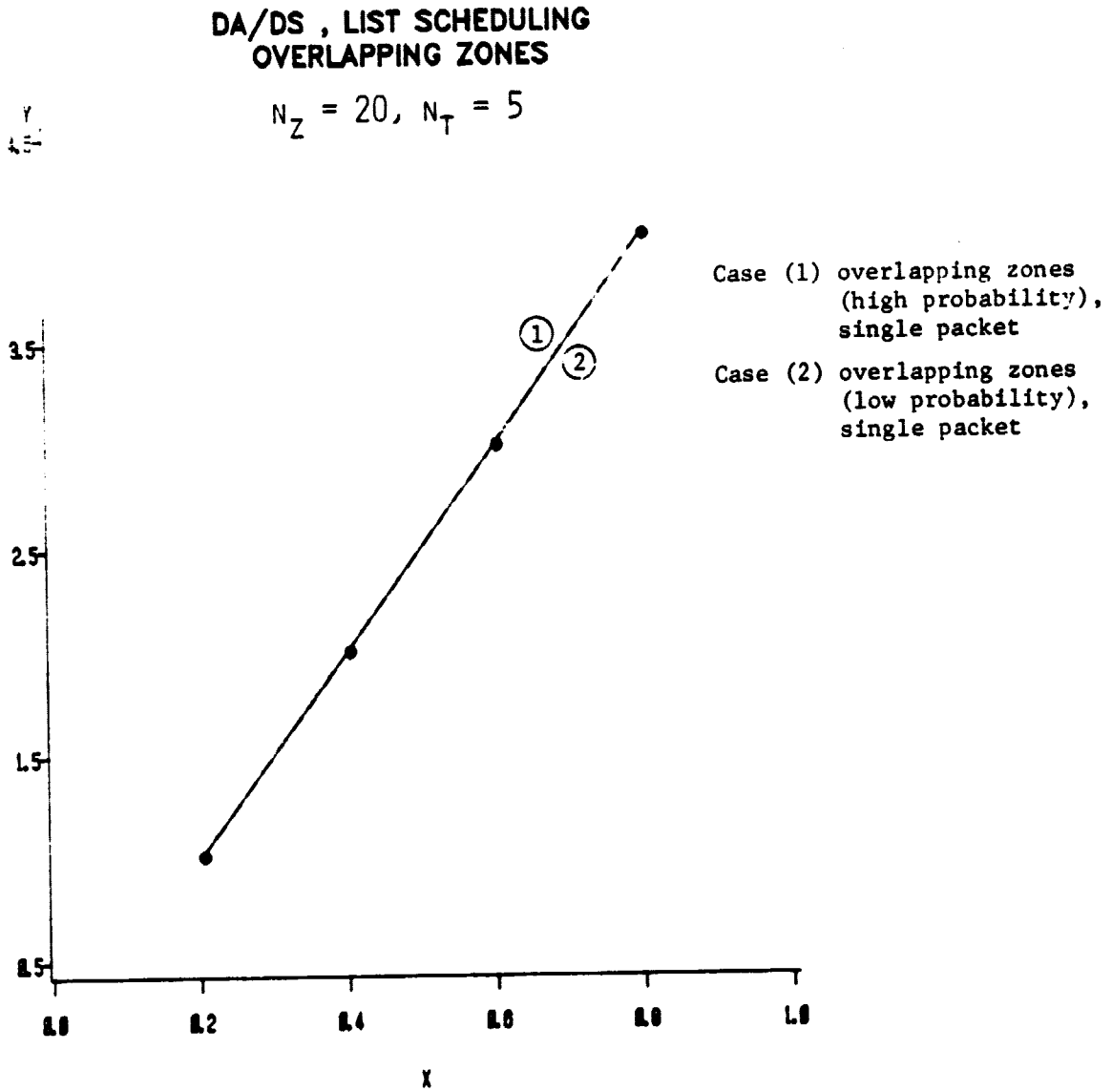


Fig. 2.55 Average number of transponders used for single packet traffic from different overlapping zones.

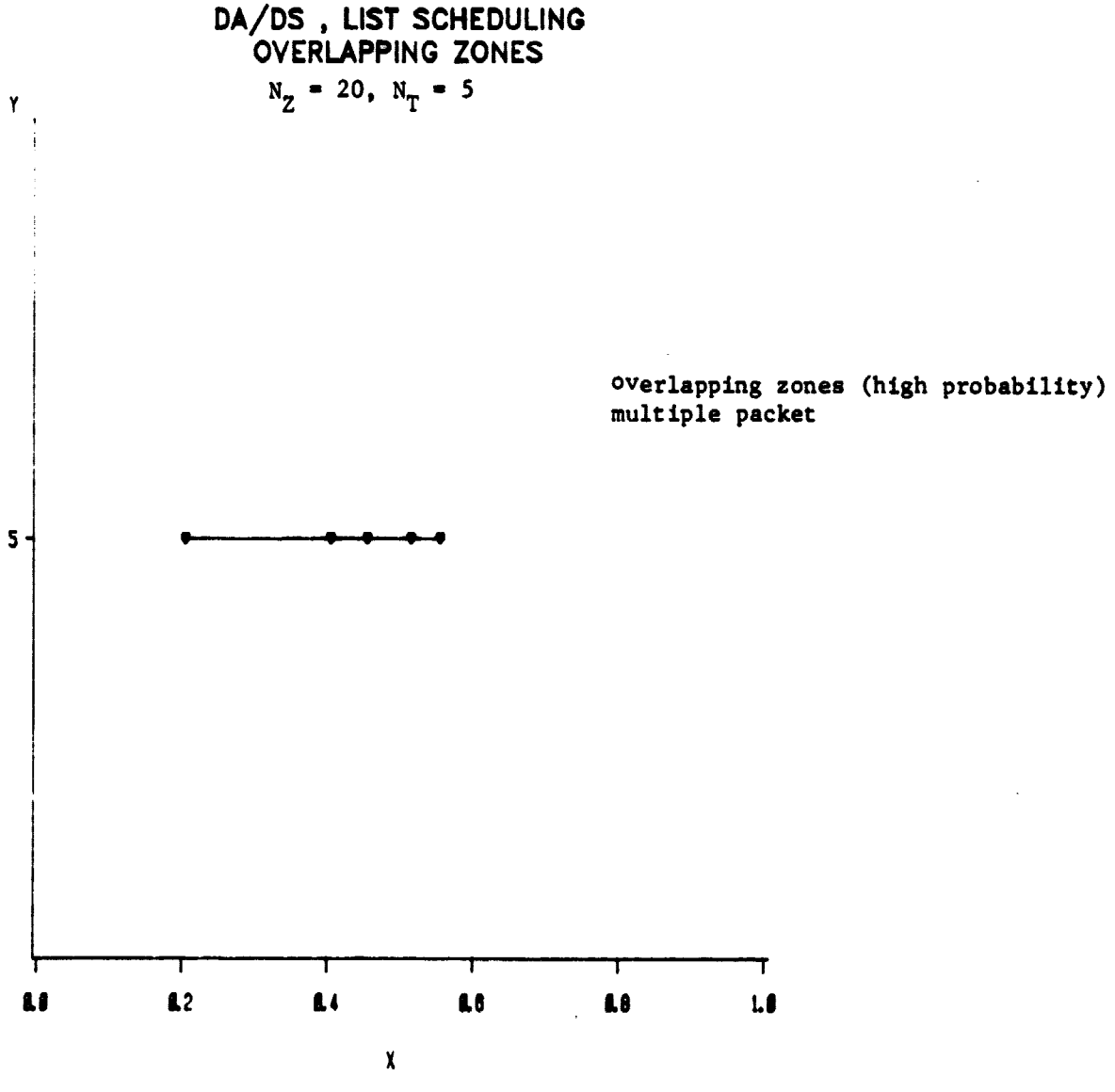
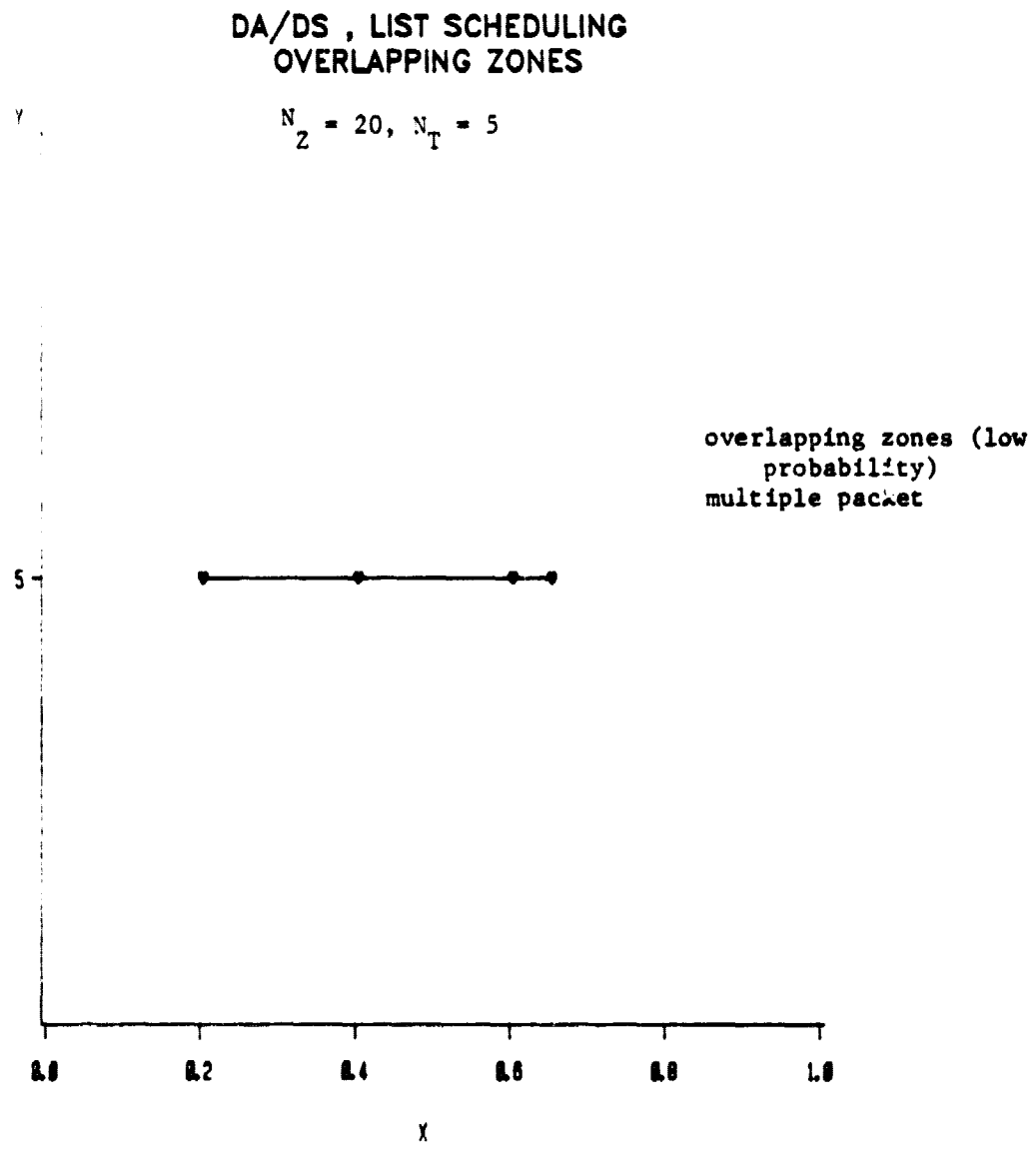


Fig. 2.56 Maximum numbers of transponders used for multipacket traffic from overlapping zones (high probability).



X=TRAFFIC INTENSITY
Y=MAXIMUM NUMBER OF TRANSPONDERS USED, N_{max}

Fig. 2.57 Maximum number of transponders used for multip ket traffic from overlapping zones (low probability).

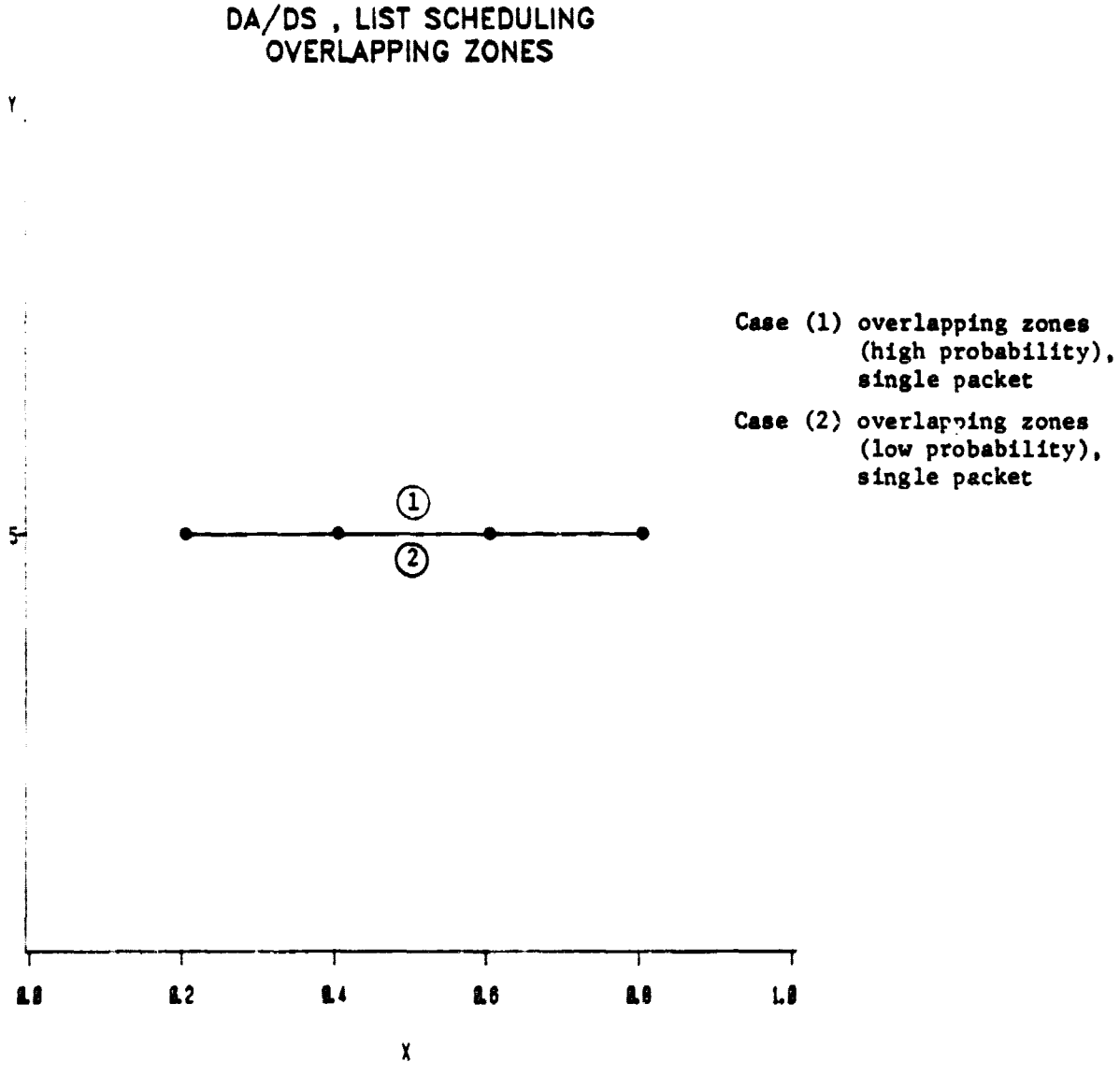
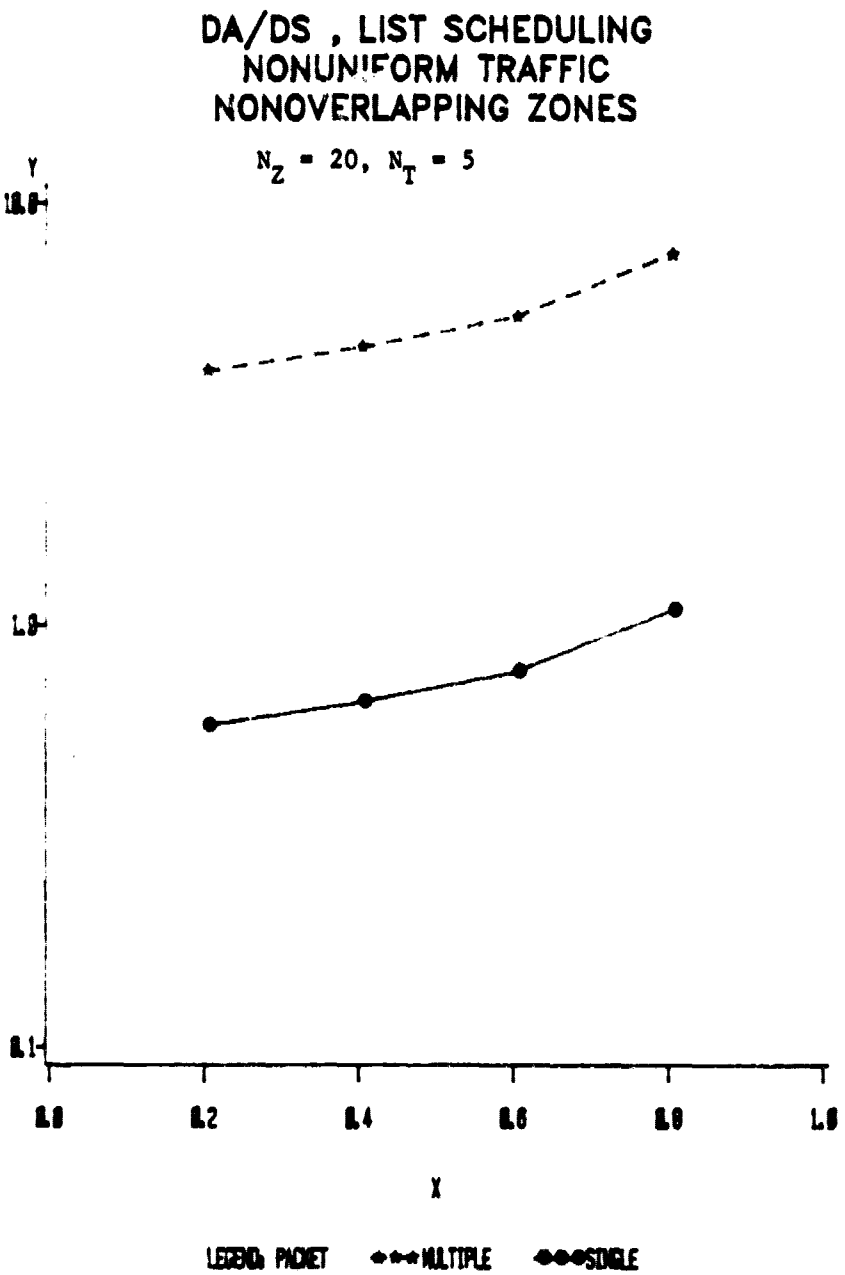
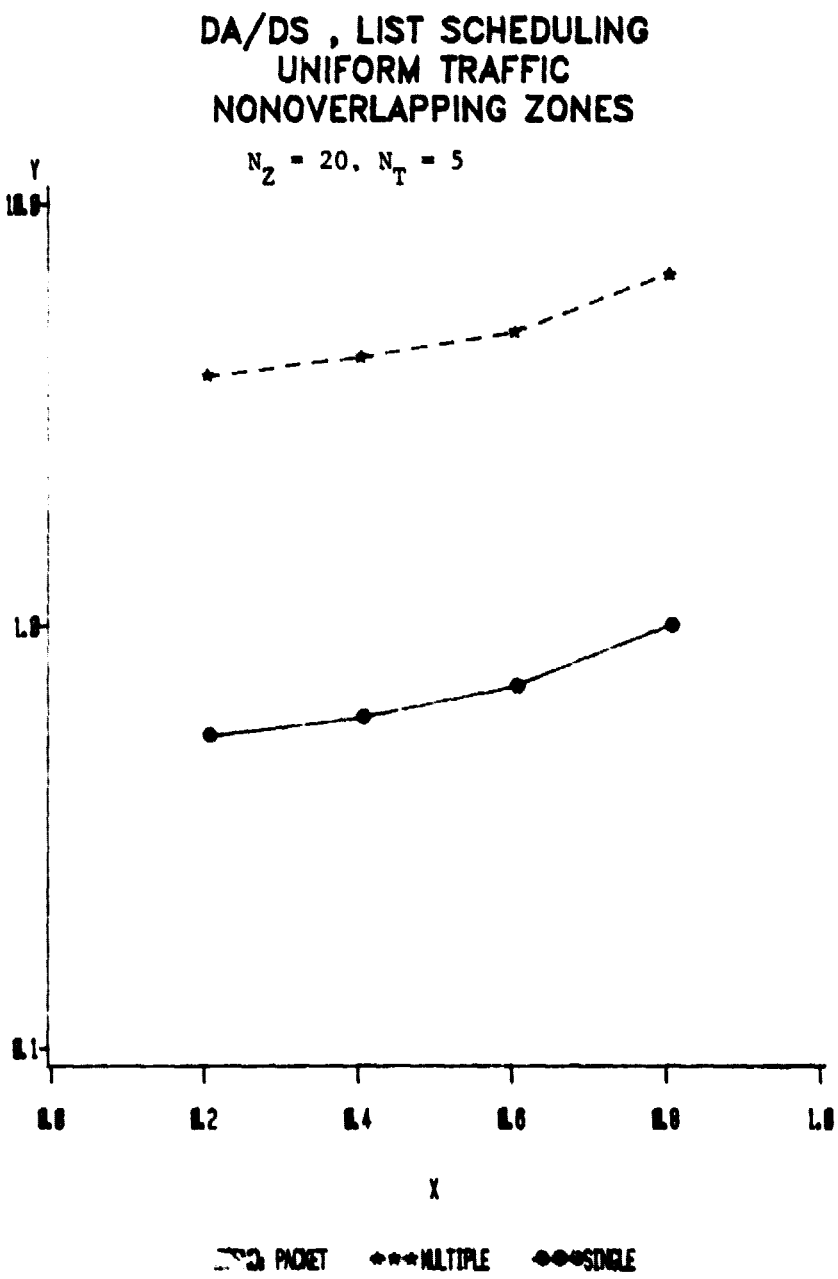


Fig. 2.58 Maximum number of transponders used for single packet traffic from different overlapping zones.



X=TRAFFIC INTENSITY
Y=MEAN PACKET WAITING TIME (SLOTS)

Fig. 2.59 Comparison of waiting time for different message lengths.



X=TRAFFIC INTENSITY
Y=MEAN PACKET WAITING TIME (SLOTS)

Fig. 2.60 Comparison of waiting time for different message lengths.

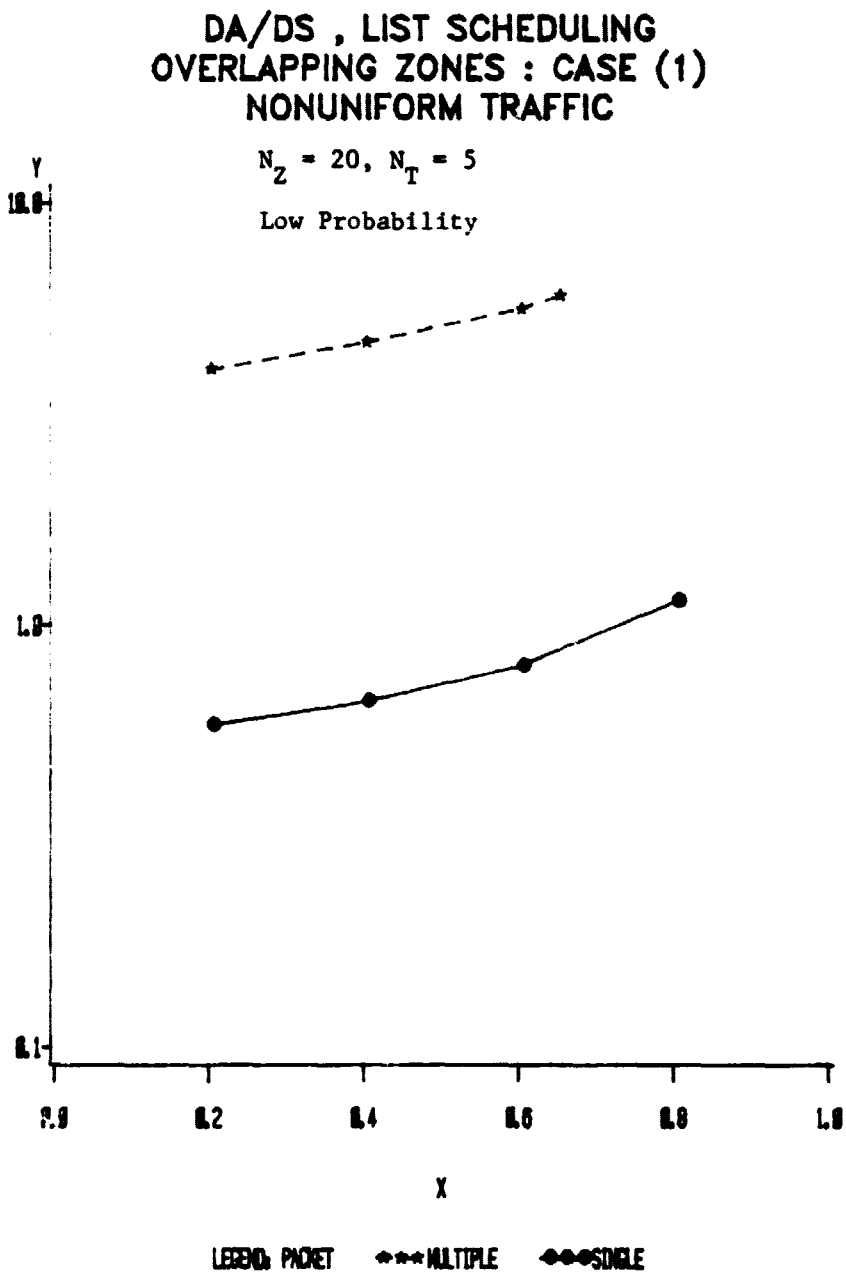
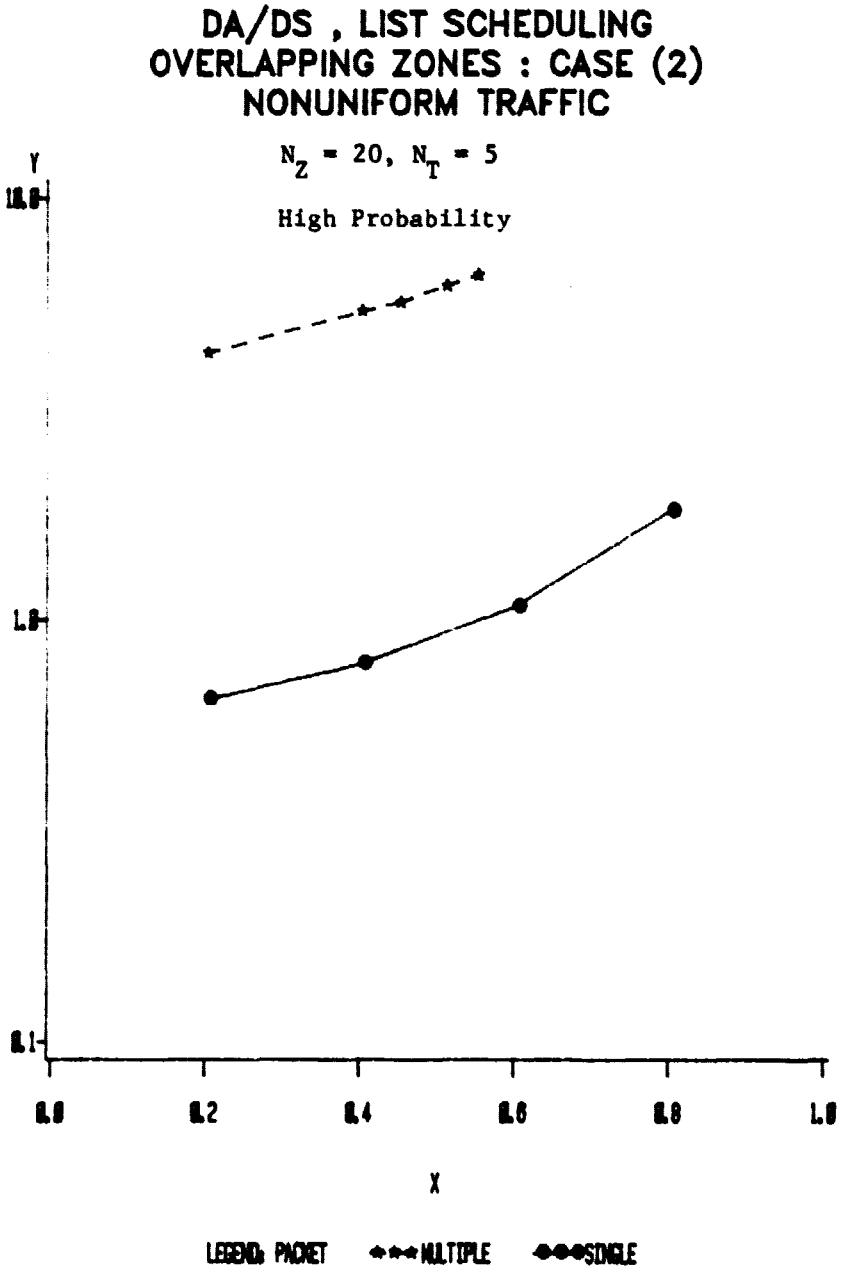


Fig. 2.61 Comparison of waiting time for different message lengths.



X=TRAFFIC INTENSITY
Y=MEAN PACKET WAITING TIME (SLOTS)

Fig. 2.62 Comparison of waiting time for different message lengths.

3. INTEGRATION OF STREAM AND BURSTY TRAFFIC

3.1. INTRODUCTION

3.1.1 Conventional Segregated-Switched Networks:

It has been a tradition in both public and private sectors to have communication networks handling separately voice and data traffic. Initially, the transmission of information was successfully implemented in two modes: Analog for voice, digital for data. Correspondingly, two switching techniques were developed and proven successful up to now:

- o Circuit Switching for Voice
- o Packet Switching for Data

The circuit switching (CS) mode consists of establishing a source-to-destination physical path (circuit) and maintaining it for the entire duration of the transaction (call). The cross network delay is then small and approximately constant. The network manages the access of CS traffic on a loss basis; i.e. if the outgoing trunk is fully busy, then arriving calls (customers) are blocked. The degree of service in such networks is typically measured by the probability of blocking (or loss) of a given call . The packet switching (PS) mode consists of grouping data into packets, each packet being routed independently through the network on a store-and-forward basis. The network communication resources (Buffers, links, etc.) are dynamically shared, hence, better utilized. The degree of service expected from a PS network is commonly evaluated by two interrelated measures: The end-to-end delay suffered by a transmitted packet and the delivery rate of packets to their destinations, referred to as throughput. In this mode of switching, time delay and throughput are subject to severe degradation when the communication resources of the network are overloaded by an uncontrolled offered traffic load. To prevent or alleviate such performance degradation and maintain the network operating point in a

desired range of efficiency (minimum delay, maximum throughput), three issues have been actively studied in recent years, namely:

- o Flow control
- o Routing
- o Buffer management

These issues, although important, are not addressed in this work, with the exception of section (3.4.1.2), where a simple data flow control mechanism is presented.

3.1.2 Evolving Integrated-Switched Networks

As the variety of users is rapidly expanding and the volume of their traffic is expected to increase sharply, future communication networks are planned to have the capabilities to handle heterogeneous users with different characteristics and requirements. Considering typical classes of traffic, and based on the transaction size, traffic rates and delivery time requirements, we may classify this traffic mix into two categories:

- o Stream Traffic: (e.g., voice, video, file transfer, ...)
- o Bursty traffic: (e.g., Interactive computing, information retrieval ...)

Stream traffic messages are mainly long and uninterrupted. Their users have a wide spectrum of characteristics (e.g., the bit rate ranges from tens of Kbps for voice to Mbps for video) and requirements (e.g., the max time delay variation ranges from 200 ms for voice to possibly tens of minutes for file transfer). This traffic may be managed on a blocking basis (e.g., voice) or on a delay basis (e.g., data).

Bursty traffic messages are short and discrete with a high peak-to-average traffic ratio. The transmitted information requires extensive error control with tolerable total network delay. This traffic is commonly managed on a delay basis.

In a communication network, the resource of primary cost is the transmission bandwidth available (or trunk capacity). The switching mode (CS or PS) for these 2 classes of traffic is thus based on the efficiency achieved in bandwidth utilization. Thus, CS is clearly more cost-effective for stream-type traffic because of the long and contiguous nature of information messages. CS leads to bandwidth waste during the inherently

long interruptions of bursty traffic flow. Hence, PS is a more likely candidate for bursty-type traffic. Figure 3.1 shows how CS and PS are compared in terms of time delay with respect to message length [OCC77]. As was mentioned earlier, CS and PS networks evolved traditionally in parallel to support separately voice and data communication needs, respectively. As more and more communication services emerge (FAX, video teleconferencing, word processing etc.), it is expected that the future trend will favor the integration of both CS and PS in a common rather than separate system. This integrated switching approach promises the following benefits [MAG 81]

- o Dynamic sharing of the network facilities leads to much higher utilization of communication resources, yielding significant dollar savings.
- o New technology trends and new applications are more easily adapted to integrated networks.
- o Equipment duplication is minimized because there is more flexibility in interconnections between the various types of user terminals.

However, in contrast to these advantages, drawbacks may be imposed at the implementation level of an integrated system. This is due primarily to the large differences between applications requirements (e.g., bandwidth ranging from tens of Hz for TTY to several Mhz for video, digital transmission for some applications, analog transmission for others...) Implementation costs, in both software and hardware, may be considerably higher and will have to be studied, however.

3.1.3 Overview of the Existing Integrated-switching Approaches

Apparently, the earliest proposal for integrated switching was initiated by IBM in 1974 [JAN 80]. It consisted of multiplexing CS and PS traffic in a slotted frame structure representing the common channel. Figure 3.2 illustrates a time division integrated multiplexor. The time axis is quantized by frames subdivided into time slots. A boundary separates the frame into a CS portion which handles CS traffic (e.g., voice) and a PS portion which handles PS traffic. This is referred to as fixed frame fixed boundary (FFFB). Due to the unpredictable variations in CS traffic, part of the CS region of the frame may be unused, thereby wasting bandwidth. To avoid this waste and improve channel utilization, PS traffic may be allowed to cross the nominal boundary and occupy idle CS slots. This is referred to as fixed frame movable boundary (FFMB). A variable frame which adapts its length to traffic variations has also been considered in the literature.

The first analytical investigation of these multiplexing structures (FFFB, FFMB) was carried out by Kummerle [JAN 80] who derived approximate
expressions for circuit blocking probability (using classical Erlang-B-formula) and average packet delay. His results, supported by simulation, showed higher performance of the MB, as may be expected. In 1976, Fisher & Harris [FIS76] attempted an exact analytical approach to the MB system, using a multiserver model technique, resulting in an M/D/N queue formulation of the PS process (data). This analysis was validated by assumptions which were reported to be erroneous in [JAN80,WEI80] and the model is claimed to give very underestimated values for packet delay when the effective capacity available for PS traffic is smaller than the offered load.

In 1977, an approximate model was developed by Occhiogrosso et al [OCC 77], in which an arbitrary slot within a frame is intermittently available for packet service. Analysis of this model is tractable, and is presented in Section 3.2 for subsequent use. At about the same time, Chang approached the analysis of the FFMB system using an exact formulation of the joint voice/data stochastic process as a two-dimensional Markov chain. The analysis was successfully carried out and results were succinctly obtained, but the computational complexity poses a severe limitation in treating realistic cases. In order for this report to be self-contained, this approach is also presented in Section 3.2 and its complexity is pointed out.

In the work of Maglaris & Schwartz [MAG 79], the variable frame scheme is first analyzed and its superior performance is shown as compared to the fixed frame. However, its implementation needs considerable hardware and software complexity to meet synchronization and time-transparency requirements for CS services. The fixed frame multiplexing structure was modeled as a Markovian decision problem under two optimality criteria involving blocking probability for CS traffic and time delay for PS traffic. Again, according to this model, the MB scheme offers near optimal performance.

Recently, Janakiraman [JAN 80] modelled the fixed and variable frame schemes both with fixed and moving boundaries, with a single server (synchronous switch) and gave a three-dimensional Markov chain formulation. The three dimensions are associated respectively with the CS process, PS process, and position of the server (switch) within a frame. Numerical results of this contribution agree quite with previous work. Other contributions in this area, mostly numerical (simulation), are also known in the literature. Table 3.1 shows a chronological list of the main contributions

Author(s) & Year	Technique	Bandwidth Management	System	Silent periods in voice used?	Distribution of Packet Length	Buffer Size
Kummerle 1975	Simulation, Approx. analysis	FFFB, FFMB	-	No	Deterministic	Finite, Infinite
Fisher and Harris 1976	Multiserver model	,,	SENET	No	"	Infinite
Forgie and Nemeth 1977	Simulation	VF	PVC	Yes	"	Finite
Chang 1977	Single and Multiserver model	FFMB	SENET	No	Exponential	Infinite
Occhiogrosso et al. 1977	Single server model	,,	,,	Yes	Deterministic	"
Miyahara and Hasegawa 1978	Simulation	VF	-	No	"	Finite
Weinstein et al. 1978	Simulation, Single server model	FFMB	SENET	No	Deterministic, Exponential	Finite, Infinite
Bially, T. 1978	Simulation	,,	SVC	Yes	Deterministic	Finite
Watanabe et al. 1978	Simulation, Diff. approx.	FFFB	-	No	"	Finite, Infinite
Anderson et al. 1979	Single server model	,,	Sub Frame Switching	No	"	Infinite
Maglaris and Schwartz 1979	Markovian Decision model	FFFB, FFMB	-	No	"	Finite
Janakiraman 1979	Single server model	FFFB, FFMB VFFB, VFMB	-	No	"	"

Table 3.1 A Chronology of Contributions to the Analysis of a Hybrid Multiplexer

with a summary of their characteristics [JAN 80].

3.1.4 Problem Definition and Outline of the Chapter

The 30/20 GHz market assessment studies predict a large number of services included in three main categories: Voice, Data, Video. This heterogeneous population of users ranging from interactive computer terminals to television channels suggests a high integration of multiplexing, multiple access and switching procedures in order to meet economically the user needs with efficient usage of the available bandwidth. In this chapter we address the isues of integrated swiching and bandwidth allocation in the environment of the 30/20 GHz multiple-beam satellite communication system. The problem of user access to this satellite system can be seen in a three-level control hierarchy.

Level 1. Connection and Concentration of the various types of information sources (classified in our terminology into stream & bursty terminals) to their earth-station.

Level 2. The M stations, competing for the same uplink channel, have to be multiplexed in an integrated structure. The TDMA protocol at this level is considered throughout this chapter.

Level 3. The onboard switching facilities are shared by the N_Z zones (covered by a multibeam antenna) to access the N_Z downlinks via N_T transponders, $N_T \leq N_Z$.

A strategy is needed to sequence and schedule the connections of uplinks to the appropriate downlinks in order to achieve some desired criterion.

This hierarchy is illustrated in Fig. 3.3.

The access problem of terminals to their earth-station (Level 1) is not addressed here. We consider first the integrated multiplexing of stations to access their common uplink channel (Level 2), second, we examine the problem of scheduling the uplink-downlink transactions, with attention paid to the constraint of time requirement inherent to some CS traffic.

In Section 3.2 the analysis problem associated with CS and PS stochastic processes arising in integrated systems is posed; then the queuing problem is analyzed by adapting two models that have appeared in the literature. These models have been selected to point out, on one hand, the computational intractability of an exact analysis, and on the other hand, the simplicity of an approximate method.

In Section 3.3, we approach the problem of allocating satellite capacity to integrated users (stations) communicating via the satellite in a multizone, multibeam system. Focussing on the station access problem (Level 2), different capacity assignment strategies are presented and analyzed for CS traffic. A numerical example is treated and performance evaluation is carried out for both CS and PS traffic. The closed form results obtained from the approximate models of Section 3.2 are used for packet performance evaluation. Finally, the different capacity assignment schemes are compared in terms of both circuit and packet performance.

3.2 TRAFFIC BEHAVIOR AT THE INTEGRATED MULTIPLEXOR LEVEL

We adopt the common assumption of Poissonian statistics for both circuit and packet arrival processes. In addition, with the assumption that both circuit switched call durations (circuit holding times) and packet lengths (packet transmission times) are exponentially distributed random variables,we achieve also Poissonian statistics for circuit disconnection and packet departure processes. Thus, the joint circuit/packet switched (CP/PS) traffic process arising in an integrated multiplexor evolves at steady state as a two-dimensional Markovian birth-death process. The parameters of this model are described in the following section.

3.2.1 Analytic Model with Exponential Message Lengths: Two-dimensional Markov Chain

The integrated system, described earlier, is essentially a queuing system with two types of arrival processes: stream arrival (CS) and bursty arrival (PS). An FFMB frame management is adopted in this analysis because of its higher performance (proven in relevant studies: FIS 76, MAG 79 etc) and its potential feasibility for use in an integrated-network design procedure. The integrated system in this case is succinctly characterized by the following characteristics and parameter assumptions:

. CS stream traffic

Arrival process is assumed Poissonian with rate λ_1.

Holding time is exponentially distributed with mean $\frac{1}{u_1}$.

Managed on a blocking basis.

. PS bursty traffic

Arrival process Poissonian with rate λ_2.

Packet length exponentially distributed with mean $\frac{1}{u_2}$.

Managed on a buffering basis.

Buffer size finite/infinite

. Channel characteristics

Frame duration: F slots

CS portion : N_s slots

PS portion : N_B slots

. Performance measures

CS traffic: Blocking probability P_B.

PS traffic: Average pkt delay and/or probability of overflow.

The time slots (i.e. circuits) are assigned to CS users (customers) at the start of each frame (SOF). It is assumed that CS customers arriving during a frame are queued until the start of the next frame and, if all the N_S slots are busy then these customers are lost. In queuing terminology, the CS process behaves like an $M/M/N_S/N_S$ process. Packets are served in each of the N_B slots plus any idle slot in the CS region. An arriving CS customer which cannot find a free slot preempts a PS customer. Preempted PS customers return to the data buffer and await service. The PS process behaves like an $M/M/N_B+S/M$ process, where S is a random variable representing the number of idle circuit slots ($0 \leq S \leq N_S$) and B is the buffer size. This system, illustrated in Fig. A.1, lends itself to a two-dimensional Markov chain formulation, as in Fig. A.2 and Appendix A. The variables Q_S and Q_B represent the number of CS and PS customers in the system, respectively. The state diagram of this joint process and the corresponding state equations for the case B finite appear in Appendix A. In principle, the state equations could be solved to determine the steady-state joint probabilities for various numbers of occupied stream traffic (CS) slots and bursty (PS) buffer occupants. However, it is not practical to do so. (See Appendix A).

Case $B = \infty$ (Infinite buffer size)

If the storage space available (B buffer units) for data is assumed to be very large compared to the date queues that may build up, we can consider the buffer size (B) to be infinite. In this case,

the system of difference equations appearing in Appendix A, is solved by the use of the standard technique of moment generating functions (mgf). Given that there are $Q_S = i$ CS customers in the system, we define the conditional m.g.f. as:

$$Pi(z) = \sum_{j=0}^{\infty} P_{ij} z^j \quad \text{for } i=0,1,\ldots N_S \qquad (3.2.1)$$

Applying this m.g.f. to the equations (A.1) to (A.6) in Appendix A, we get the linear system

$$Aij \begin{vmatrix} P_0(z) \\ P_1(z) \\ \cdot \\ \cdot \\ \cdot \\ P_{NS}(z) \end{vmatrix} = \begin{vmatrix} b_0(z) \\ b_1(z) \\ \cdot \\ \cdot \\ \cdot \\ b_{NS}(z) \end{vmatrix} \qquad (3.2.2)$$

Where Aij, the (i,j)the element of the $N_S x N_S$ system matrix, is a function solely of the parameters (λ_1, λ_2, μ_1, μ_2, N_S, N_B), and bi(z) is a function of the unknowns Pij, $i=0,1,\ldots N_S$ $j=0,1,\ldots F-i-1$ that we need to determine in order to solve the linear system (3.2.2) for the Pi(z). Using Eqs. (A.1) through (A.3) of Appendix A, the Pij may be expressed in terms of Poj, j=0,...F. To find these Poj, we need (F+1) equations, obtained as follows:

- Equation (A.4) solved recursively yields N_B equations
- One equation is obtained by Σ_{ij} Pij = 1
- Considering the Cramer solution $Pi(z) = \frac{\det Ai(z)}{\det A(z)}$ where Ai(z) is the matrix A with i-th Column replaced by vector b(z).

It can be proved that det A(z) has N_S unique roots z_j in (0,1). The existence of the Pi(z) requires

$$\det A_i(z_j) = 0 \quad \text{for } j=1,2,\ldots N_B \tag{3.2.3}$$

which leads to a set of N_B equations.

At the costof great computational effort in this intermediary step, one can manage to solve for Pi(z). Related details can be found in [FIS 77] and related work by Chang.

a) PS traffic performace

Case B finite: In addition to the average packet time delay, an appropriate measure, useful for buffer size design, is the probability of overflow.

$$P_o = \Pr[i+j = B+F] = \sum_{i=0}^{N_S} P_{i,\, B+F-i} \tag{3.2.4}$$

Case B infinite: Pkt performance is measured by the average time delay

$$E(T) = \frac{1}{\lambda_2} \sum_{i=j}^{N_S} \frac{d\, P_i\,(z)}{dz} \Bigg|_{z=1} \tag{3.2.5}$$

Only simple examples (N_S=1, N_B=0) have been treated in the literature [WEI 80] using this method because realistic cases involve unmanageable computations inherently encountered in recursive solutions and finding the roots of det A(z) = 0 in the disk (0,1).

b) CS traffic performance

The probability of steady state occupance of the $M/M/N_S/N_S$ system governing the CS process is

$$P\left[\text{k busy slots}\right] = \frac{\rho_1^k/k!}{\sum_{m=D}^{N_S} \rho_1^m/m!}, \text{ where } \rho_1 = \frac{\lambda_1}{\mu_1} \qquad (3.2.6)$$

The blocking probability, the CS performance measure, is then a special case:

$$P_B = \left[N_S \text{ slots are busy}\right] = \frac{\rho_1^{N_S}/N_s!}{\sum_{m=0}^{N_s} \rho_1^m/m!} \qquad (3.2.7)$$

It should be noted that the above expression for P_B, known in telephony as the Erland-B formula, is a approximation in this case because CS customers access the queuing system only at the start of the frame. Therefore, the Erlang-B formula is valid only for short frame durations. Fig. 3.4 shows the dependence on P_B and the validity of the Erlang-B formula with frame duration.

In conclusion, the model, introduced here for completeness, highlights the computational complexity, mentioned earlier that is involved in analyzing integrated switching. Consequently, this model is abandoned for the rest of this report. Instead, approximate methods to analyze PS traffic performance are considered in the next section.

3.2.2 Approximate analytic models:

a) Formulation as M/D/1 queue with ON-OFF server

In the frame configuration that we have been describing so far, we seperate the CS region (N_S slots) from the PS region N_B slots). However, as far as implementation is concerned, the CS slots are not confined to occupy only the CS portion of the frame,

but can be spread out or interleaved along the frame in a deterministic or quasi-random manner. Thus, a packet arriving at any instant within the frame may attempt service at the start of the next slot if this latter is available (not occupied by a circuit). So, a time slot selected randomly in the frame is intermittently available for packet transmission. Clearly, the probability of the slot availability at some time, denoted by σ depends on the frame management policy (fixed or moving boundary) and the number of circuits in progress at that time. With this frame implementation, the system lends itself to be modelled as in Fig. 3.5 by an M/D/1 queue where the server is available for PS customers with probability σ given by:

* Fixed boundary

$$\sigma = \frac{N_B}{N_S+N_B} \qquad (3.2.8)$$

* Movable boundary

$$\sigma = \frac{F-\varepsilon}{F} \qquad (3.2.9)$$

where ε is the average number of slots occupied by CS customers within a frame.

An expression for the queue length of the randomized service M/D/1 system may be obtained as in [OCC 77].

Let P_n = Pr [n pkts are present in the system at the start of the current data slot.]

γ_k = Pr [k new arrivals during a data slot.]

Then, as in [OCC 77],

$$P_n = \sum_{k=0}^{n-1} \gamma_k \left[(1-\sigma)P_{n-k} + \sigma P_{n+1-k}\right] + \gamma_n(P_0+\sigma P_1) \qquad (3.2.10)$$

and the moment generating function is

$$P(z) = \frac{\sigma P_0 (z-1)}{z + \frac{1}{\Omega(z)} - \sigma} \qquad (3.2.11)$$

. P_0 = Pr [empty system]

$$= 1-\lambda_2 \Delta^* = 1-\lambda_2 \frac{\Delta}{\sigma} \qquad (3.2.12)$$

Δ^* is the effective constant service time (slot duration)

. $\Omega(z) = \sum_{j=0}^{\infty} \gamma_j z^j$ is the mgf of the pkt arrival process. (3.2.13)

Using this expression, the average number of pkts in the integrated switching system is then

$$L = \left. \frac{dP(z)}{dz} \right|_{z = 1} \qquad (3.2.14)$$

Assuming that packets arrive according to a Poisson process, and using Little's result, we get the average packet time delay

$$T = \frac{L}{\lambda_2} = \frac{\Delta (1-\rho_2/2)}{\sigma-\rho_2} \qquad (3.2.15)$$

or normalized

$$\frac{T}{\Delta} = \frac{1-\rho_2/2}{\sigma-\rho_2} \qquad (3.2.16)$$

It can be noticed that this result is similar to the one of a normal M/D/1 queue (given by $\frac{\Delta(1-\rho_2/2)}{1-\rho_2}$). Clearly, the effect of service interruptions, which occur with relative frequency σ, increases the delay suffered by an average packet and limits the PS

traffic intensity ρ_2 to be smaller than σ, otherwise the system gets unstable.

This model is used in the next chapter to evaluate PS traffic performance in an integrated multiplexor transmitting over a satellite channel.

b) Formulation as a single server subject to "Breakdowns"

Contrary to the previous implementation model, the CS slots are now assigned adjacently in the circuit region of the frame (N_S slots). Thus, as far as the transmission of packets is concerned, the FFFB or FFMB frame structures can be modelled as a single-service system whose service is interrupted during the portion of the frame occupied by CS traffic (see Fig. 3.6). A packet arriving during the PS portion of the frame starts service at the start of the next slot, whereas a packet arriving during the CS portion of the frame "sees" the server in "breakdown" state and thus must wait until the start of the PS portion to be served. (See Fig. 3.6). Packets are assumed of constant length.

As before, the CS slot occupancy is characterized by the probability

$$P_k = \Pr\left[k \text{ occupied CS slots}\right] = \frac{\rho_1^k/k!}{\sum_{m=0}^{NS} \rho_1^m/m!} \qquad (3.2.17)$$

and

$$\Pr\left[\text{blocking}\right] = P\left[k=N_S\right] \qquad (3.2.18)$$

Here, σ, as defined previously, will represent the fraction of the frame reserved for PS traffic (and $1-\sigma$ that reserved for CS traffic).

We focus on PS traffic performance analysis by evaluating the mean time delay using a "work" concept and taking into account the

"repair time of service breakdowns". [YU 81]. A typical realization of the "work" function is portrayed in Fig. 3.7.

Let $\tau = \Delta F$: duration of the frame

Δ = pkt service time (or slot duration)

T = mean system time of a pkt from arrival until departure.

A new pkt may arrive during the CS or PS portion of the frame with probability $(1-\sigma)$ or σ, respectively.

Case of fixed boundary

Assume first that a packet arrives during the PS portion of the frame, and let T_1 be its system time. The overall time spent by a packet in the system consists of three components, namely:

- $\frac{\sigma\tau}{2}$: expected time before pkt service will be interrupted.
- $(1+\lambda_2 T_1)\Delta$: Residual work to be completed on average by the packet before leaving the system. The term $\lambda_2 T_1$ represents the average number of packets that would arrive during the time (T_1) the packet is in the system. This is equivalent to the number of packets that the arriving customer will find on average in the queue, and which have to be served before the considered packet can go to service.
- $\frac{(1+\lambda_2 T_1)\Delta - \sigma\tau/2}{\sigma\tau}(1-\sigma)\tau$: Service repair time which is equal to the number of breakdowns times the average interruption period.

where $\lfloor x \rfloor$, $\lceil x \rceil$ denote the smallest integer $\geq x$ and the largest integer $\leq x$ respectively. We can approximate $\lceil x \rceil$ by $x+1$, and therefore the mean system time of a pkt arriving during the PS portion of the frame is

$$T_1 = \frac{\sigma\tau}{2} + (1+\lambda_2 T_1)\Delta + \left\lfloor \frac{(1+\lambda_2 T_1)\Delta + \sigma\,\tau/2}{\sigma\tau} \right\rfloor (1-\sigma)\tau \qquad (3.2.19)$$

If the packet arrives during the CS portion of the frame, it will be delayed, similarily, by

- $(1-\sigma)\frac{\tau}{2}$: expected time until start of PS portion (start of service)
- $(1+\lambda_2 T_2)\Delta$: residual work (T_2 is Average System time)
- $\left\lfloor \frac{(1+\lambda_2 T_2)\Delta}{\sigma\tau} \right\rfloor (1-\sigma)\tau$: service repair time

therefore, we have

$$T_2 = (1-\sigma)\frac{\tau}{2} + (1+\lambda_2 T_2)\Delta + \left\lfloor \frac{1+\lambda_2 T_2)\Delta}{\sigma\tau} \right\rfloor (1-\sigma)\tau \qquad (3.2.20)$$

The total pkt system time is the weighted sum of T_1 and T_2,

$$T = \sigma T_1 + (1-\sigma)T_2 \qquad (3.2.21)$$

If the number of service interruptions (breakdowns) is large, we can approximate $\lfloor x \rfloor \approx x$ in Eqs. (3.2.19) and (3.2.20).

Thus, Eq. (3.2.21) becomes

$$T = (1-\sigma)^2\frac{\tau}{2} + (1+\lambda_2 T_1)\Delta + \left\lfloor \frac{(1+\lambda_2 T_1)+\Delta\frac{\tau}{2}}{\sigma\tau} \right\rfloor (1-\sigma)\tau + \left\lfloor \frac{(1+\lambda_2 T_2)\Delta}{\sigma\tau} \right\rfloor (1-\sigma)^2\tau \qquad (3.2.22)$$

Using Eqs. (3.2.19) and (3.2.20), we get the normalized system time [YU 81].

$$\frac{T}{\Delta} = \frac{(1-\sigma)\,\tau/2 + \frac{1}{\sigma}}{1 - \frac{\lambda_2\Delta}{\sigma}} \qquad (3.2.23)$$

Case of movable boundary

The effective boundary is represented by the normalized average number of busy slots in the CS region, that is

$$E(\gamma) = \sum_{k=0}^{N_S} \frac{kP_k}{N_s} \quad , \text{ where } P_k = \frac{\rho_1^k/k!}{\sum_{m=0}^{N_S} P_1^m/m!} \tag{3.2.25}$$

and its second moment is

$$E(\gamma^2) = \sum_{k=0}^{N_S} k^2 \frac{P_k}{N_S^2} \tag{3.2.26}$$

The average system time delay encountered by a packet is estimated in the same manner as in the previous case, noting that a packet that arrives during the CS or PS portion of the frame is expected to wait $E(\gamma^2)\tau/2E(\gamma)$ or $E(1-\gamma)^2\ \tau/2\ E(1-\gamma)$ before the CS or PS portion ends, respectively [YU 81] .

The partial system times, conditioned on the arrival time within CS or PS portion, are respectively

$$T_2 = \frac{E(\gamma^2)}{2E(\gamma)}\tau + (1+\lambda_2 T_2)\Delta + \left\lfloor \frac{(1+\lambda_2 T_2)\Delta}{E(1-\gamma)} \right\rfloor E(\gamma)\tau \tag{3.2.27}$$

$$T_1 = \frac{E(1-\gamma)^2}{2E(1-\gamma)}\ \tau + (1+\lambda_2 T_1)\Delta + \left\lceil \frac{(1+\lambda_2 T_1)\Delta - E(1-\lambda)^2\ \tau/2E(1-\gamma)}{E(1-\gamma)\tau} \right\rceil E(\gamma)\tau \tag{3.2.28}$$

The total time is

$$T = \left[1-E(\gamma)\right]\ T_1 + E(\gamma)\ T_2$$

After some calculation, we get [YU 81].

$$\frac{T}{\Delta} = \frac{\frac{\alpha\tau}{2\Delta} + \frac{1}{1-E(\gamma)}}{1 - \frac{\lambda_2 \Delta}{1-E(\gamma)}} \tag{3.2.30}$$

where

$$\alpha = \frac{E(\gamma^2)+E(\gamma)-2E(\gamma)E(\gamma^2)-2E^2(\gamma)\ +2E^3(\gamma)}{1-E(\gamma)}$$

3.3 INTEGRATED SWITCHING IN A MULTIPLE BEAM SATELLITE SYSTEM

In the context of the NASA 30/20 GHz studies, the problem of multiple access and resource allocation has been previously considered in a packet switched traffic environment [STE80]. The same problem is adressed here in an integrated circuit-switched (CS) and packet-switched (PS) traffic environment arising in a heterogeneous population of users to be accommodated by a multibeam satellite system. We first concern ourselves with the integrated station access problem. Several schemes are then presented and their performance compared.

3.3.1 Uplink Capacity Allocation

The satellite system is assumed to serve N_Z zones via uplinks and downlinks of the same capacity (e.g., 2 Mbps). The same average traffic load flows between zones. M earth stations located in each zone collect traffic from stream (CS) and bursty (PS) terminals in that zone. The access problem of these terminals to their earth station is not considered here. Therefore, earth stations within a zone are viewed as integrated users, sharing the same uplink channel in a TDMA mode. The capacity (C bps) of the satellite channel is represented on the time axis by a fixed frame structure managed on a fixed-frame, fixed-boundary (FFFB) or fixed-frame, moveable-boundary (FFMB) policy. The frame is divided into F time slots of duration Δ . A nominal boundary separates the CS portion (N_S slots) from the PS portion (N_B slots). The multiple access of stations within a zone as illustrated in Fig. 3.8 leads to the following resource allocation problem.

Capacity allocation at the station access level:

Given a satellite uplink channel of capacity C bps, shared by M CS/PS integrated users in a TDMA mode, we seek strategies to assign this capacity to the integrated users in order to achieve the following objectives:

- Efficient utilization of the bandwidth (maximum utlization of the frame).
- A desirable level of blocking for CS traffic (e.g., Voice Calls).
- The best possible delay-throughput characteristic for PS traffic. (i.e., minimize delay, maximize throughput).

3.3.1.1 Circuit Capacity Assignment

Focusing first on the CS capacity (N_s slots) assignment problem, several schemes are now presented (see Fig. 3.9).

a) Fixed assignment.

The same number of slots, $N = \frac{N_s}{M}$ is assigned to each station during each frame in a fixed manner as in Fig. 3.9a. The circuit arrival process from each station is assumed Poissonian with the same rate λ_1, and the circuit holding time is exponential with mean $\frac{1}{\mu_1}$. Assuming that the frame duration is short enough (compared to circuit holding time), the acces of CS customers from a given station to the channel is governed by an M/M/N/N system. Again, the state probability of this system is

$$P_k = \frac{\rho_1^k/k!}{\sum_{m=0}^{N} \rho_1^m/m!} \qquad (3.3.1)$$

where $\rho_i = \frac{\lambda_1}{\mu_1}$

These CS customers are blocked when all N slots are busy, that is with probability.

$$P_r[k = N] = P_B = \frac{\rho_1^N/k!}{\sum_{m=0}^{N} \rho_1^m/m!} \qquad (3.3.2)$$

A quantity of crucial interest, used later in packet performance evaluation, is the total average number of slots occupied by the CS customers. It is given by

$$\varepsilon = M \sum_{k=0}^{M} k P_k$$

$$= M\rho_1(1-P_B), \qquad (3.3.3)$$

after some calculation.

It is clear that this fixed assignment scheme leads to bandwidth waste during low traffic periods because many slots remain unoccupied.

b) Shared-pool assignment

The corresponding frame configuration is illustrated in Fig. 3.9b.

A pool of U unassigned slots is made available in the CS portion. The remaining (N_S - U) slots are assigned to the M users in a fixed manner. The free pool U is shared by all CS customers on a demand basis, that is if the N slots are all busy then the overflowing traffic has access to the pool. The probability of overflow is just

$$P_{B_1} = \frac{\rho_1^N / N!}{\sum_{m=0}^{N} \rho_1^m / m!} \qquad (3.3.4)$$

The intensity of the total overflowing traffic competing for the pool is given by

$$\rho_u = M \rho_1 P_{B_1} \qquad (3.3.5)$$

The overflow arrival process is no longer Poisson, but under heavy traffic the Poisson property remains approximately valid. With this assumption, an overflow CS customer which finds the pool full is blocked with probability

$$P_{B_2} = \frac{(\rho_u)^u / u!}{\sum_{m=0}^{U} (\rho_u)^m / m!} \qquad (3.3.6)$$

Finally, the total blocking probability for a particular CS customer is given approximately by the product

$$P_B = P_{B_1} P_{B_2} \qquad (3.3.7)$$

The average number of busy slots can be derived as follows.

$$\epsilon = M \rho_1 (1 - P_{B_1}) + M \rho_1 P_{B_1} (1 - P_{B_2})$$

$$\epsilon = M \rho_1 (1 - P_B) \qquad (3.3.8)$$

c) Demand assignment.

In the scheme of Fig. 3.9c, any arriving CS customer is assigned a time slot (if available) in the CS portion of the frame, on a demand basis. The

intensity of arrivals is $M\rho_1$ and P_B is, similarily, given by

$$P_B = (M\rho_1)^{N_s}/N_s! \; / \sum_{m=0}^{N_s} (M\rho_1)^m/m! \qquad (3.3.9)$$

The average slot occupany is again

$$\epsilon = M\rho_1(1 - P_B). \qquad (3.3.10)$$

d) Shared Channels Assignment

In the previous model, the pool of U slots is of arbitrary size and its access is unrestricted. Because of traffic variations, a particular station may occupy most of the pool, creating congestion among other incoming CS channels. To avoid this, the same model is refined in the following way: The pool is divided into R shared channels consisting of N_A slots each. Any station, on demand, may take one of the R channels (if any) and use up to N_A slots. The station must release this channel when it is not needed to make it available for the other stations. The corresponding frame configuration is illustrated in Fig. 3.9d. It should be noticed, as opposed to the previous model, that N,R,N_A are variable parameters and are related by

$$N_S = MN + RN_A \qquad (3.3.11)$$

It is of interest to find optimal values of R, N_A and N (or equivalently, optimal pool size $U = RN_A$) which yield minimum blocking probability for an arriving CS customer.

The queuing system handling the M CS stations with their associated shared channels can be viewed as a set of interconnected M/M/m/m queues (where m is either N or $N + N_A$). In Appendix B, we analyze this system to find the state equation from which we derive the blocking probabilty P_B for CS arrivals as well as the average state occupancy.

Choice of R and N_A:

By examining the expression for P_B, one would like to choose values for R and N (or equivalently N_A), say R* and N* (N_A*), that reduce the blocking probability as much as possible. To be able to choose these values, one needs to know how P_B varies with R and N(N_A). The search for (R*,N*) was carried out numerically for several examples. It was observed from these examples that P_B decreases when N increases for R fixed ($0 \leq N \leq \frac{N_s}{M}$), and that the smallest P_B is attained for $N^* = \frac{N_s}{M} - 1$. On the other hand, among the minima of P_B, the smallest value tends to correspond to the value of R given empirically by $R^* = \frac{M}{2}$.

From the equation $N_s = N^*M + R^*N_A$ we deduce $N^*_A = 2$ using the general empirical results $N^* = \frac{N_s}{M} - 1$ and $R^* = \frac{M}{2}$ stated above.

(e) Impact of digitization rate on circuit capacity assignment.

It was mentioned earlier that the services to be integrated in the 30/20 GHz system occupy a large spectrum, from narrow to wideband signals. The CS portion of the integrated multiplexor TDMA frame must thus handle transmissions with a variety of digitization rates (bits/sec). This suggests the idea of assigning channel capacity (time slots) in proportion to these rates. Let us assume that, classified by decreasing order of rate, we have K clases of CS customers. Frame slots may be assgined as in Table 3, where

$$\begin{cases} R_1 > R_2 \ldots > R_k \\ \alpha_1 > \alpha_2 \ldots > \alpha_k \end{cases} \qquad (3.3.12)$$

and

$$\alpha_i = \frac{R_i}{\sum_i R_i} \quad . \qquad (3.3.13)$$

Table 3.2 Traffic Classes by Rate

Class	Traffic Intensity	Rate	Slot Allocation/frame
1	$\rho_1 = \frac{\lambda_1}{\mu_1}$	R_1	α_1 slots (bytes)
2	ρ_2	R_2	α_2 "
'	'	'	'
'	'	'	'
'	'	'	'
'	'	'	'
k	ρ_k	R_k	α_k

Assuming that the frame duration is much shorter than the average circuit holding time (exponential), this system can be viewed as a K-dimensional Markov Chain. The state equation is given by the closed form expression.

$$P(n_1,n_2,\ldots n_k) = P(\underline{0}) \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{n_i!} \qquad (3.3.14)$$

The constant $P(\underline{0})$ is determined by the normalization condition

$$\sum_{S} P(\underline{n}) = 1 \qquad (3.3.15)$$

where S is the space of all allowable states, defined by the resource constraint

$$\underline{\alpha}\,\underline{n}^T = \alpha_1 n_1 + \alpha_2 n_2 + \ldots + \alpha_k n_k \leq N_s \qquad (3.3.16)$$

From Eq. (3.3.15), we have

$$P^{-1}(\underline{0}) = \sum_{n_1=0}^{I_1} \sum_{n_2=0}^{I_2} \cdots \sum_{n_k=0}^{I_k} \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{n_i!} \qquad (3.3.17)$$

where I_1, is the maximum number of class 1 customers in the system, I_2 in the maximum number of class 2 customer given that n_1 class 1 customers are in the system, etc. That is,

$$\begin{cases} I_1 = \left\lfloor \frac{N_s}{\alpha_1} \right\rfloor \\ I_2 = \left\lfloor \frac{N_s - n_1 \alpha_i}{\alpha_2} \right\rfloor \\ \cdot \\ \cdot \\ \cdot \\ I_k = \left\lfloor \frac{N_s - \sum_{i=1}^{k-1} n_i \alpha_i}{\alpha_k} \right\rfloor \end{cases} \qquad (3.3.18)$$

where $\lfloor x \rfloor$ denotes the largest integer $\leq x$,

A customer of class i is blocked with probability

$$P_{B_i} = \sum_{\underline{n}} P(\underline{n}) \qquad (3.3.19)$$

where the sum is over states such that $\underline{n} \in S$ and $\underline{n} + \underline{e}_i \notin S$, where e_i is the i^{th}call entry vector (all zeros except 1 in the i^{th} position).

Exponential Assignment:

Assume, without loss of generality that

$$N_S = \alpha^{k-1}m \qquad (3.3.20)$$

where α is an integer constant greater than 1, and m is an integer. Time slots are assigned to a class i arriving customer according to the rule

$$\alpha_i = \alpha^{k-i} \text{ slots} \qquad (3.3.21)$$

In this case, we have

$$P^{-1}(\underline{0}) = \sum_{n_1=0}^{I_1} \sum_{n_2=0} \cdots \sum_{n_k=0}^{I_k} \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{n_i!} \qquad (3.3.22)$$

where

$$\begin{cases} I_1 = \dfrac{N_s}{\alpha_1} = \dfrac{\alpha^{k-1}\, m}{\alpha^{k-1}} = m \\ I_2 = \dfrac{\alpha^{k-1}\, m - \alpha^{k-1} n_1}{\alpha^{k-2}} = \alpha(m-n_1) = \alpha(I_1-n_1) \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ I_i = \dfrac{\alpha^{k-1}m - (\alpha^{k-1} n_1 + \alpha^{k-2} n_2 + \ldots + \alpha^{k-i+1} n_{i-1})}{\alpha^{k-i}} \\ \qquad = \alpha\left[\alpha^{i-2}\, m - (\alpha^{i-2} n_1 + \alpha^{i-3} n_2 + \ldots + n_{i-1})\right] = \alpha\left[I_{i-1} - n_{i-1}\right] \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ I_k = \alpha\left[\alpha^{k-2} m - \sum_{j=1}^{k-1} \alpha^{k-j-1} n_j\right] = \alpha\left[I_{k-1} - n_{k-1}\right] \end{cases} \qquad (3.3.23)$$

This problem is known in classic telephone trunking as the multi-channel problem because of the N_s slots (or channel) within a frame. It arises typically in a T1 transmission system in which, for example, 64 Kps and 32 Kbps rates are to be handled.

For illustration, let us consider the case of $K = 4$, $\alpha = 2$, $N_s = 2^3 m$. The N_s slots are allocated to arriving customers (calls) as in Table 3.3.

Table 3.3 Slot Allocation

$k = 4, \alpha = 2, N_s = 8m$

Class	Slots allocated
1	$2^3 = 8$
2	$2^2 = 4$
3	$2^1 = 2$
4	$2^0 = 1$

We are interested in the blocking probability for each traffic class. For instance, this probability can be computed for class 3 in the following way:

$$P_{B_3} = \sum_{n_1} \sum_{n_2} \sum_{n_3} \left\{ P(n_1,n_2,n_3,8m-8n_1-4n_2-2n_2) + P(n_1,n_2,n_3,8m-8n_1-4n_2-2n_2-1) \right\}$$

$$+ P(0,0,4m,0) \qquad (3.3.24)$$

The total blocking probability is given by the weighted sum

$$P_B = \sum_{i=1}^{k=4} \frac{\rho_i}{\sum_{j=1}^{k} \rho_j} P_{B_i} \qquad (3.3.25)$$

The average number of busy slots (channels) in the system is given by

$E(8n_1+4n_2 + 2n_3 + n_4) = 8E(n_1) + 4E(n_2) + 2E(n_3) + E(n_4)$.

And it is known from before that:

$$E(n_i) = \rho_i(1 - P_{B_i}), \qquad (3.3.26)$$
$$i = 1, \ldots, K = 4$$

which is just the average number of class i customers in the system.

Consider the simple numerical example of $K = 2$, $\alpha = 2$, $N_s = 32 = 2 \times 16 = 2\ m$. From the previous expressions we have

$$P(n_1,n_2) = \left\{ \sum_{i=0}^{m} \sum_{j=0}^{2(m-i)} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \right\}^{-1} \frac{\rho_1^{n_1}}{n_1!} \frac{\rho_2^{n_2}}{n_2!} \qquad (3.3.27)$$

$$P_{B_1} = \sum_{n_1=0}^{m-1} \left\{ P(n_1,\ 2m-2n_1-1) + P(n_1, 2m-2n_1) \right\} + P(m,0) \qquad (3.3.28)$$

$$P_{B_2} = \sum_{n_1=0}^{m} P(n_1,\ 2m - 2n_1) \qquad (3.3.29)$$

$$E(2n_1+n_2) = 2\ \rho_1(1-P_{B1}) + \rho_2(1-P_{B2}) \qquad (3.3.30)$$

Considering the traffic mix

$$\beta = \frac{\rho_1}{\rho_1 + \rho_2} \qquad (3.3.31)$$

as a parameter, the blocking probabilities for classes 1 and 2 are plotted versus the offered load in Figs. 3.10 and 3.11 for two values of β, 0.5 and

0.75. Similar parametric curves can be obtained for $k > 2$ classes of circuit-switched services with various rates. These curves may form an important tool for the system designer so as to evaluate a desired degree of service (here P_B) under a given traffic mix (β) condition. In a sense, ranking the users' traffic according to the rates is analogous to assigning partial priorities. Figs. 3.10 and 3.11 show the gap in the degree of service (blocking) between the two classes of users. This gap, as expected, is relatively more accentuated for light traffic.

3.3.1.2 Packet Capacity Allocation

The frame may be shared by PS traffic under both fixed and moving boundary policies. The average time delay, the packet performance measure, is evaluated using the approximate models described in section 3.2.2, namely:

*M/D/1 with random server

*Single server subject to "breakdowns."

The expressions for packet time delay are given by Eqs. (3.2.16),(3.2.23) and (3.2.30) Because of its superiority in performance, the FFMB scheme is given more emphasis here and is considered in the following numerical example. It is important to stress the fact that the two models above correspond to two different frame implementations. Consequently, these two models cannot be compared on a numerical basis. In the example treated in the next section, we used only the first model to evaluate packet performance . Clearly, the evaluation of this performance is straightforward using the second model by referring to Eqs.(3.2.23) and (3.2.30).

3.3.2 Performance Comparison

The uplink satellite channel (assumed of capacity 2 Mbps) is synchronously clocked into time slots of duration Δ msec, grouped in frames of F slots. Data packets, assumed of constant length P = 1000 bits, are transmitted one per slot, so that $\Delta = \frac{P}{} = \frac{1000}{2\times10^6} = 5$ msec. The circuit and packet performances are now evaluated on the following example: F = 50 slots, N_S = 24 slots, N_B = 26 slots.

Fig. 3.12 shows the variation of the blocking probabilities with the CS offered load per user, ρ_1, for each of the four slot allocation schemes. We note, as expected, that the performance of the fixed assignment scheme is much worse than the two demand assigned schemes for small blocking probabilities. The shared channel (where the pool is accessed two slots at a time) scheme is somewhat better than fixed assignment for this uniform traffic environment, but still considerably worse than the demand assigned schemes in the same range of traffic.

The comparative performance of the schemes reverses in the case of packet delay, however, since in demand-assigned schemes, fewer slots remain unused. Fig. 3.13 showing packet delay as a function of packet-switched traffic intensity ρ_2, with ρ_1 = 2 Erlangs circuit-switched traffic per user, indicates that the fixed assignment scheme is now the best, although not by very much. Note that the curves here are shown dashed beyond $\rho_2 = \frac{N_B}{F} =$ 0.52. This indicates that Eq.(3.2.16) on which these curves are based, is presumably no longer valid beyond that point. This is because in the original model [OCC 77], it was assumed that the CS and PS processes are uncoupled which is only valid approximately if one does not attempt to use the added slots made available by the movable boundary to drive the system with a PS traffic intensity $\rho_2 > 0.52$. The added slots may be used

to reduce packet time delay, not to increase throughput.

All four schemes utilizing a movable boundary strategy show considerable reduction in packet time delay over a fixed boundary strategy. For this range of parameters, then, it is the movable boundary that provides the basic improvement in packet performance. The added effect of the specific circuit-switched assignment strategy represents a relatively minor perturbation. In Fig. 3.14 we observe that the performance of the demand assignment strategy deteriorates with increased circuit-switched load more sensitively than the other strategies. However, this effect must be interpreted with care since blocking probabilities become unacceptable for high range of ρ_1.

The comparative study carried out here may be extended to the case of a more heterogeneous-traffic user environment, where results may differ from those obtained in this work.

3.4 ON BOARD PROCESSING AND STORAGE

Many advantages are associated with equipping the satellite with processing repeaters instead of conventional translating repeaters. The former provide on-board demodulation and remodulation which permits switching to be performed at base band frequencies, resulting in less hardware complexity. The effects of noise and interference can be reduced by appropriate control and regeneration of the detected data.

More importantly, the availability of detected data (bit streams, packets, messages...) at the satellite, with the addition of some intelligence and buffering capabilities, suggest the use of the satellite as an integrated store and forward node performing the sequencing and switching of the CS/PS traffic from uplinks to downlinks (see Fig. 3.17).

3.4.1 Integrated Downlink Channel

The outgoing pipe consists of N_z downlink channels connecting independently the satellite to the N_z zones as in Fig. 3.18. A buffer is associated with each downlink channel and the collected traffic is routed to the appropriate buffer corresponding to its destination. The channel is assumed accessible to CS traffic with preemptive priority over the PS traffic. CS traffic may be managed on a blocking basis.

In this section we study the buffer behavior and the resultant queuing delay at the satellite level. We first assume that both CS and PS traffic have the same statistics, namely Poisson arrivals and exponential service time. A frame structure is used for the entire satellite-earth pipe and each downlink channel is assigned a sub-frame of 2 slots. The first slot forms a CS subchannel and can be used, if free, for packet transmission. The two slots are not necessarily of the same duration. Thus, the integrated channel can be modelled by a single-server queuing system in which the capacity of the server alternates between

two levels in a Poissonian manner. This is an application of a general queuing model presented in [YEC 69].

Next, we introduce priority traffic classes associated with ranked services. Following the work by Shantikumar [SHA 81], his model is extended below to our application of an integrated-switched channel, assuming constant packet length.

3.4.1.1 Two-Dimensional M/M/1 System

As far as PS traffic is concerned, the server (channel) oscillates between two states denoted 1 (channel occupied by a circuit call) and 0 (channel free). The time that the server remains at a given state is exponentially distributed just as interarrival and service times of the CS process have an exponential distribution. The queuing model is shown in Fig. 3.15 and the corresponding state diagram in Fig. 3.16.

The joint steady-state probabilities $\{P_{ij},\ i=0,1;\ j=0,1,\ldots\}$ are related by the following balance equations:

$$\left[\lambda_2(1)+\mu_1\right] P_{10} = \mu_2(1)P_{11} + \lambda_1 P_{00} \qquad (3.4.1)$$

$$\left[\lambda_2(0)+\lambda_1\right] P_{00} = \mu_2(0)P_{01} + \mu_1 P_{10} \qquad (3.4.2)$$

$$\left[\lambda_2(1)+\mu_2(1)+\mu_1\right] P_{ij} = \mu_2(1)P_{1,j+1} + \lambda_2(1)P_{1,j-1} + \lambda_1 P_{oj} \qquad (3.4.3)$$

$$\left[\lambda_2(0)+\mu_2(0)+\lambda_1\right] P_{oj} = \mu_2(0)P_{o,j+1} + \lambda_2(0)P_{o,j-1} + \mu_1 P_{1j} \qquad (3.4.4)$$

In the case of finite buffer size (say B), the system of equations above, together with the condition $\Sigma_{ij} P_{ij} = 1$, can be solved recursively for the P_{ij}s. This involves (2B+3) equations, which may be manageable if B is not very large. Quantities of interest, such as average buffer content or probability of overflow are readily derived when the P_{ij}'s are known. This case fits very well with the application to the on-board buffer for a permissible probability of overflow.

If a large memory capacity is provided on the satellite so that queue lengths cannot exhaust the storage space, we can assume infinite buffer size. This case is treated in the following.

Since the average persistence time of the server (channel) in state 0 and 1 is $1/\lambda_1$ and $1/\mu_1$ respectively, we have the probabilities:

$$\Pi_1 = \Pr\left[\text{server in state } 1\right] = \frac{1/\mu_1}{1/\mu_1+1/\lambda_1} = \frac{\rho_1}{1+\rho_1} \tag{3.4.5}$$

$$\Pi_0 = \Pr\left[\text{server in state } 0\right] = \frac{1/\lambda_1}{1/\mu_1+1/\lambda_1} = \frac{1}{1+\rho_1} \tag{3.4.6}$$

The average arrival and service rates for the packets in this queuing system are defined as follows:

$$\bar{\lambda}_2 = \Pi_0\ \lambda_2(0) + \Pi_1\ \lambda_2(1) \tag{3.4.8}$$

$$\bar{\mu}_2 = \Pi_0\ \mu_2(0) + \Pi_1\ \mu_2(1) \tag{3.4.8}$$

It should be noticed that the packet arrival process is totally independent of the server state, i.e., $\lambda_2(1) = \lambda_2(0) = \lambda_2$, but the notation above is kept for convenience. It will appear later that a proportionality relation between the traffic parameters $\frac{\lambda_2(1)}{\mu_2(1)} = \frac{\lambda_2(0)}{\mu_2(0)}$ leads to extremely simple solvability of the problem. This relation can be generalized and it is discussed later as a Flow Control mechanism.

The set of difference equations (3.4.1) through (3.4.4) can be reduced to the single equation

$$\lambda_2(1)P_{1j} + \lambda_2(0)P_{oj} = \mu_2(1)P_{1,j+1} + \mu_2(0)P_{0,j+1} \tag{3.4.10}$$

Put in vector form, this becomes

$$\underline{\lambda_2}\underline{P}_j^T = \underline{\mu_2}\ \underline{P}_{j+1}^T \tag{3.4.11}$$

with the vector definition

$$\underline{P}_j = \left[P_{1j},\ P_{oj}\right] \quad \text{and} \quad \underline{\lambda}_2 = \left[\lambda_2(1),\ \lambda_2(0)\right] \tag{3.4.12}$$

Eq. (3.4.10) recalls the M/M/1 result $\lambda P_j = \mu P_{j+1}$. Thus, the queuing system under consideration can be viewed as a 2-dimensional M/M/1 system. The average residual service capacity obtained from (3.4.10) by summing over j, is given by

$$\bar{\mu}_2 - \bar{\lambda}_2 = P_{10}\mu_2(1) + P_{00}\mu_2(0) \tag{3.4.13}$$

For the system to be stable, this residual capacity must be positive, i.e.,

$$\bar{\mu}_2 - \bar{\lambda}_2 > 0$$

Recall that $\bar{\lambda}_2$ and $\bar{\mu}_2$ are given by (3.4.8) and (3.4.9). Now, defining the conditional m.g.f. of the system as

$$G_i(z) = \Sigma_{j=0}^{\infty} P_{ij} z^j \ , \quad i = 0,1 \tag{3.4.14}$$

We can transform Eqs. (3.4.1) through (3.4.4) into the following linear system in $G_1(z)$ and $G_0(z)$:

$$\begin{bmatrix} A_1(z) & -\lambda_1 \\ -\mu_1 & A_0(z) \end{bmatrix} \begin{bmatrix} G_1(z) \\ G_0(z) \end{bmatrix} = \begin{bmatrix} B_1(z) \\ B_0(z) \end{bmatrix} \tag{3.4.15}$$

where

$$A_1(z) = \left[\lambda_2(1) + \mu_1 + \mu_2(1)\right] - \left[\lambda_2(1)z + \mu_2(1)/z\right] \tag{3.4.16}$$

$$A_0(z) = \left[\lambda_2(0) + \lambda_1 + \mu_2(0)\right] - \left[\lambda_2(0)z + \mu_2(0)/z\right] \tag{3.4.17}$$

$$B_i(z) = \mu_2(i)\ (\tfrac{z-1}{z})\ P_{io} \ , \text{ for } i = 1,0 \tag{3.4.18}$$

To solve this system, we need to determine P_{io}, i = 1,0 i.e., the probabilities that the system is empty of packets in one state or the other. For this, we need two equation in P_{10} and P_{00}. One equation is given by Eq. (3.4.13) and the second is obtained as follows:

The Cramer solution of the system of (3.4.15) is given by

$$G_i(z) = \frac{\det M_i(z)}{\det M(z)}, \tag{3.4.20}$$

where $M = \begin{bmatrix} A_1(z) - \lambda_1 \\ -\mu_1 \; A_0(z) \end{bmatrix}$

and M_i is M with the i^{th} column replaced by vector $\underline{B}(z)$

It is known that det M(z) has a root at some z_o, $|z_0| < 1$. Therefore, det $M_i(z)$ must have the same root for a solution to exist.

$$\det M(z_0) = \det M_i(z_0) = 0, \quad \text{for } |z_o| \leq 1 \tag{3.4.21}$$

This condition leads to finding the root of the third-degree polynomial [YEC 69].

$$g(z) = Az^3 + Bz^2 + Cz + D \tag{3.4.22}$$

where

$$A = \lambda_2(1)\ \lambda_2(0) \tag{3.4.23}$$

$$B = -\left[\mu_1\lambda_2(0)+\lambda_1\lambda_2(1)+\lambda_2(1)\lambda_2(0)+\lambda_2(1)\mu_2(0)+\lambda_2(0)\mu_2(1)\right] \tag{3.4.24}$$

$$C = \left[\mu_1\mu_2(0)+\lambda_1\mu_2(1)+\mu_2(0)\mu_2(1)+\lambda_2(1)\mu_2(0)+\lambda_2(0)\mu_2(1)\right] \tag{3.4.25}$$

$$D = -\mu_2(1)\mu_2(0) \tag{3.4.26}$$

Combining equations (3.4.13) and (3.4.21), we can get the quantities

P_{10}, P_{20} [YEC 69], which enable us to obtain $G_1(z)$ and $G_0(z)$ from (3.4.15).

The quantity of interest for us is the average number of packets in the buffer, given by

$$EL = \frac{d}{dz}\left[G_1(z) + G_0(z)\right]\Big|_{z=1} = \sum_{j=0}^{\infty} j(P_{1j} + P_{0j}) \qquad (3.4.27)$$

The final expression [YEC 69] is

$$EL = \frac{\bar{\rho}_2}{1-\bar{\rho}_2} + \left[\mu_2(1)(\mu_2(0)-\lambda_2(0))\, P_{10} + \mu_2(0)(\mu_2(1)-\lambda_2(1))P_{00} - (\mu_2(1)-\lambda_2(1))(\mu_2(0)-\lambda_2(0))\right] \frac{1}{(\lambda_1+\mu_1)(\bar{\mu}_2-\bar{\lambda}_2)} \qquad (3.4.28)$$

It is important to observe that the first term is the average queue length of an M/M/1 queue with utilization factor $\bar{\rho}_2 = \frac{\bar{\lambda}_2}{\bar{\mu}_2}$. In the case of an integrated channel, and during the long holding time of a CS call ($\frac{1}{\mu_1} \sim 3$ min.) the service capacity for packets is reduced ($\mu_2(1) < \mu_2(0)$) and the queue builds up to arbitrary large sizes. A sensitivity analysis would show that the second term in (3.4.28) grows very rapidly with ratios such as $\frac{\mu_2(0)}{\mu_1}$, $\frac{\bar{\lambda}_2}{\mu_1}$. The first ratio compares a call holding time to packet service time, the second ratio indicates the average number of packets arriving when the channel is (partially or totally) occupied by a circuit.

This point is investigated in the following special case: For example, consider a single slot channel, i.e., $\mu_2(1) = 0$, with

$$\lambda_2(1) = \lambda_2(0) = \lambda_2$$

$$\bar{\mu}_2(0) = \mu_2$$

the parameters become

$$\bar{\lambda}_2 = \lambda_2$$

$$\bar{\mu}_2 = \Pi_0 \mu_2(0) = \Pi_0 \mu_2 \qquad (3.4.29)$$

From Eq. (1.10), we have

$$P_{00} = \Pi_0 - \frac{\lambda_2}{\mu_2} = \Pi_0 - \rho_2 \tag{3.4.30}$$

Substituting these parameters in Eq. (3.4.28), we get the average number of packets in the queue:

$$EL = \frac{\bar{\rho}_2}{1 - \bar{\rho}_2} + \frac{\bar{\rho}_2}{1 - \bar{\rho}_2} \cdot \frac{\mu_2}{\mu_1} \cdot \Pi_0(1 - \Pi_0), \tag{3.4.31}$$

where

$$\bar{\rho}_2 = \frac{\lambda_2}{\Pi_0 \mu_2} \tag{3.4.32}$$

We readily observe that the ratio $\frac{\mu_2}{\mu_1}$, which is in the order of 10^3 to 10^4 (e.g., typically $\mu_2 = 100$, $\mu_1 = 0.01$ in Voice Data traffic), makes the second term much larger than the first term of Eq. (3.4.31).

Using Little's result, the average packet queuing time is $D = \frac{EL}{\lambda_2}$. (3.4.33) This delay, which is the packet performance measure, is plotted versus the PS load in Fig. 3.22 for the same channel availability Π_0 but with two different ratios $\frac{\mu_2}{\mu_1}$ of 10^4 and 10^3. The gap between the two curves is very significant and increases drastically with λ_2. The packet buffer behavior in terms of time delay is depicted in Fig. 3.24 versus CS traffic load for a fixed PS traffic intensity ($\rho_2 = 0.3$).

3.4.1.2 Flow Control

From the simple numerical case treated above, it appears from the curves of Figs. 3.22 and 3.23 that the queue of packets builds up during periods of high utilization of the channel by CS traffic during periods of peak circuit holding times, the accumulation of packets causes very long queuing delays if the stability condition is not violated. Thus, the need for a flow contral (FC) and/or buffer management mechanism is apparent. This need is more urgent for our underlying

application in the satellite system because of long distance propagation and on-board weight requirements. Next, a simple feedback control mechanism is presented to regulate λ_2, the arrival rate of packets, according to the state of the channel (server). If a circuit is being transmitted (served), the feedback controller reduces λ_2 to $\lambda_2(1)$, and if no circuit is on the channel the controller does nothing, i.e., $\lambda_2 = \lambda_2(0)$ (see Fig. 3.19a). Interestingly enough, if this FC mechanism adjusts the packet rate λ_2 such that

$$\frac{\lambda_2(1)}{\mu_2(1)} = \frac{\lambda_2(0)}{\mu_2(0)} = \rho \quad \text{(constant)} \tag{3.4.33}$$

then, it can be readily proved by induction [YEC 69] that the general state of the system is described by the joint probabilities $P_{i,j} = \Pi_i(1-\rho)\rho^j$ for $i = 0,1$.

Therefore the packet queue length is governed by $P_j = (\Pi_0+\Pi_1)(1-\rho)\rho^j = (1-\rho)\rho^j$ which is nothing but the steady-state equation of an M/M/1 of factor ρ. It is very important to observe that the PS process is now completely uncorrelated from the CS process, i.e., the state of the channel (server), and the number of packets in the system is simply

$$EL = \frac{\rho}{1-\rho}, \tag{3.4.34}$$

where

$$\rho = \frac{\lambda_2(0)}{\mu_2(0)} = \frac{\text{pkt arrival rate}}{\text{Total channel capa.}} \tag{3.4.35}$$

The dynamic behavior of this oscillating queuing system is governed by the parameters λ_1, μ_1 characterizing the stastistics of the CS traffic emanating from users of various types. This traffic may be very dense (large λ_1) or very infrequent (small λ_1) with variable call duration or circuit holding time depending on the service (voice, file transfer, etc.). It is useful to quantify

the behavior of the system under study when such extreme traffic situations occur. Two cases of particular interest, namely:

- Very heavy CS traffic ($\lambda \to \infty$) and very short circuit holding time ($\mu_1 \to \infty$) but with constant traffic intensity, i.e., $\frac{\lambda_1}{\mu_1} = \alpha$ (constant). The system oscillates very rapidly between the two states and the probabilities of being in these states are

$$\Pi_1 = \frac{\rho_1}{1+\rho_1} = \frac{\alpha}{1+\alpha} \tag{3.4.36}$$

$$\Pi_0 = \frac{1}{1+\rho_1} = \frac{1}{1+\alpha} \tag{3.4.37}$$

Equations (3.4.1) - (3.4.4) then yield

$$P_{ij} = P_{oj}, \quad \text{for } j = 0,1,\ldots \tag{3.4.38}$$

From Eq. (3.4.13), we have

$$P_{00} = \left[\mu_2(0) + \alpha\, \mu_1(0)\right] = \bar{\mu}_2 - \bar{\lambda}_2 \tag{3.4.39}$$

or

$$P_{00} = \Pi_0\,(1 - \bar{\rho}_2), \tag{3.4.40}$$

where

$$\bar{\rho}_2 = \frac{\bar{\lambda}_2}{\bar{\mu}_2} \tag{3.4.41}$$

Similarily,

$$P_{10} = \Pi_1(1 - \bar{\rho}_2) \tag{3.4.42}$$

By inductive reasoning, we can show that

$$P_{ij} = \Pi_i(1 - \bar{\rho}_2)\bar{\rho}_2^j, \; j = 0.1,\ldots, \; i = 0.1 \tag{3.4.43}$$

and the total state probability is

$$P_j = (1 - \bar{\rho}_2)\bar{\rho}_2^j. \tag{3.4.44}$$

We observe that the system converges to an M/M/1 system with average parameters $\bar{\lambda}_2$ and $\bar{\mu}_2$.

- Very light CS traffic ($\lambda_1 \to 0$) and very long circuit holding time ($\mu_1 \to 0$), (e.g., transfer of a very long file in a single message) with $\frac{\lambda_1}{\mu_1} = \alpha$.

The set of equations (3.4.1) through (3.4.4) becomes

$$\lambda_2(i)\, P_{i0} = \mu_2(i)\, P_{i1}$$

$$\left[\lambda_2(i)+\mu_2(i)\right] P_{ij} = \lambda_2(i)\, P_{i,j-1} + \mu_2(i) P_{i,j+1} \qquad \text{for } i = 0,1 \qquad (3.4.45)$$

which are the standard difference equation of an M/M/I system with $\rho_i = \frac{\lambda_2(i)}{\mu_2(i)}$. In other words, if $\lambda_2(i) < \mu_2(i)$, $i = 0,1$, the system operates in two stable M/M/1 modes with probabilities Π_0 and Π_1.

Therefore, the total state probability is:

$$P_j = \Sigma_{i=0,1}\, \Pi_i \left[1 - \rho_2(i)\right] \rho_2^j(i) \qquad (3.4.46)$$

Quantities of interest are readily derived from these probabilities.

3.4.1.3 Generalization to Multislot Downlink Channels

The control feedback mechanism described previously, which regulates the packet flow and uncouples the two traffic processes (CS and PS), can be extended to the general case of a multislot channel where the frame is divided into F slots (N_1 slots for circuit subchannel, N_2 for packet subchannel). The packet arrival rate is maintained proportional to the effective channel capacity $\mu_2(i)$ where i is the number of circuits in the system; that is

$$\frac{\lambda_2(i)}{\mu_2(i)} = \frac{\lambda_2(i-1)}{\mu_2(i-1)} \doteq \frac{\lambda_2(i)}{F-i} = \frac{\lambda_2(i-1)}{F-(i-1)} \qquad \text{for } i = 1,\ldots,N_1 \qquad (3.4.47)$$

$\lambda_2(0)$ represents the uncontrolled packet arrival rate (i.e. Max) and $\mu_2(0)$ the maximum capacity of the channel.

We discussed this case in section 3.2.1., where the packet subchannel was modelled as an M/M/S/∞ system with S = F - i. Recalling the two-dimensional Markov formulation, it was not easy to find the state probabilities P_{ij} of the joint CS/PS stochastic process (i,j) and we saw how computationally complex the problem is. Fortunately, the FC scheme described above (see Fig. 3.3.19b) transforms the problem to an extremely simple one because it uncorrelates the two processes (CS and PS) arising in the integrated channel, that is:

$$P_{ij} = P_i \, P_j \tag{3.4.48}$$

This fact can be shown by inductive reasoning on the set of balance equations [YEC 69] just as it was shown previously for the case of $N_1 = 1$, $N_2 = 1$. Furthermore, the marginal state probabilities are given by

$$P_i = \frac{\rho_1^i/i!}{\sum_{m=0}^{N_1} \frac{\rho_1^m}{m!}} \quad (M/M/N_1/N_1 \text{ system}) \tag{3.4.49}$$

$$P_j = (1-\rho)\rho^j \quad (M/M/1 \text{ with } \rho = \frac{\lambda_2(0)}{\mu_2(0)} = \frac{\lambda_2}{(N_1+N_2)\mu_2}) \tag{3.4.50}$$

The performance of the system is evaluated by the circuit blocking probability P_B and average packet queue length EL. These are given by:

$$P_B = \frac{\rho_1^{N_1}/N_1!}{\sum_{m=0}^{N_1} \rho_1^m/m!} \tag{3.4.51}$$

$$EL = \frac{\rho}{1-\rho} \tag{3.4.52}$$

where

$$\rho = \frac{\lambda_2}{(N_1+N_2)\mu_2} \qquad (3.4.53)$$

A simulation study would be needed to compare the performance of the system with and without the FC mechanism. It is mentioned in [KWO 80] that a simulation experiment led to a performance improvement by a factor of 1000.

3.4.2 Priority Strategies in On-Board Switching

In the heterogenous environment of the satellite users that we described previously, an interesting issue is the assignment of priority classes among the data services. This priority ordering may be based on packet length, conversation duration, signalling messages, etc. We consider a set of K packet priority classes arriving at a given downlink buffer modelled as a single server (channel) queuing system. The availability of the server is randomly interrupted by CS traffic transmission. The system parameters are as follows:

- Priorities 1,2,...K
- Poisson arrivals with rates $\lambda_1,\ldots,\lambda_K$ packets/time unit.
- Packet service interval = one time slot = one time unit.
- Pr [server available] = σ
- FIFO service for same priority customers.

Following the model in [SHA 81], we define the following quantities:

- waiting time: time from a customer's arrival to its first service attempt.
- service time: time from first attempt to completion of service.
- Queuing time: waiting time + service time

The average queuing time EW_r for priority-r customers is derived using level crossing analysis [SHA 81] on the virtual waiting time V_r^t process defined as the waiting time of a customer r if it were to arrive at t. If an r customer arrives

at t = 0 and the buffer is found empty then it may attempt service; otherwise it has to wait until all higher priority customers who may have arrived are served. Let G_r be the random time needed to clear the buffer from these higher priority customers. So, just after t = 0, $V_r = G_r$ and then drops linearily until it reaches zero or an r-customer arrives. In the first situation, an r-customer arrival increases V_r^t by G_r because he either attempts service or has to wait for the higher priority customers to be cleared. In the second situation, the arriving r-customer waits until the preceeding r customer is served (call B_r this service time) and the buffer is emptied of higher priority customers (call G_r' this time). So V_r^t jumps by the amount $H_r = B_r + G_r'$ (see sample service realization and V_r^t sample path [SHA 81] in Fig. 3.20 and 3.21).

When the buffer gets empty, a new cycle is initiated. We assume that the process V_r^t converges to a steady state, i.e.,

$$\lim_{t\to\infty} V_r^t = V_r \tag{3.4.54}$$

and let $V_r(x)$ be the PDF of V_r, i.e.,

$$V_r(x) = P_r[V_r \leq x] \tag{3.4.55}$$

The level crossing analysis consists of evaluating the expected numbers of times the process V_r^t crosses the level x upward (U(x)) and downward (D(x)), then

$$E\ U(x) = E\ D(x) \tag{3.4.56}$$

The analysis is carried out in [SHA 81] and the details are omitted here. The average queuing time in slots, for the r-priority class is given by

$$EW_r = 2\frac{\sigma(2-\sigma)}{[\sigma-\theta_r][\sigma-\theta_{r-1}]} + 1 \tag{3.4.57}$$

where

$$\theta_r = \Sigma_{i=0}^{r} \lambda_i \tag{3.4.58}$$

and λ_i is in packets/slot.

Using Little's result, the total average number of packets in the buffer is

$$EL = \Sigma_{r=1}^{K} \lambda_r E W_r \tag{3.4.59}$$

Some observations need to be made:

- The first term in the expression EW_r is the total waiting time (from arrival until successful attempt of service) which has the same form as the well known priority waiting time [KLE 75] .

$$EW_r = \frac{E T_0}{(1-\theta_r)(1-\theta_{r-1})} \tag{3.4.60}$$

with

$$\theta_r = \Sigma_{i=1}^{r} \frac{\lambda_i}{\mu_i} = \Sigma_{i=1}^{r} \rho_i \tag{3.4.61}$$

Here $E T_0$ is the residual service time given by

$$E T_0 = \frac{\sigma(2-\sigma)}{2} \tag{3.4.62}$$

in this case.

- In the case of one traffic class (K = 1), we have

$$EW_1 = \frac{\sigma + 2(1-\lambda_1)}{2(\sigma-\lambda_1)} = \frac{(1 - \lambda_1/2)}{\sigma - \lambda_1} + \frac{1}{2} \tag{3.4.63}$$

and interestingly, the first term is just the result derived by Occhiogrosso et al. in [OCC 77] , using a completely different analytic tool. However, the addition of the term $\frac{1}{2}$ makes this result more accurate because it takes into account the fact that a packet arriving within a slot can attempt service only at the start of the next slot.

- More importantly, EL, the total average number of packets in the buffer, is the same under priority service and FIFO service. This is

readily shown by computing

$$EL = \Sigma_{i=1}^{K} \lambda_i \, EW_i \tag{3.4.64}$$

which is

$$EL = \Sigma \, \lambda_i \left[\frac{\sigma(2-\sigma)}{(1-\theta_i)(1-\theta_i-1)} + 1 \right] = \lambda \left[\frac{\sigma + 2(1-\lambda)}{2(\sigma-\lambda)} \right] \tag{3.4.65}$$

where

$$\lambda = \Sigma_{j=0}^{K} \lambda_j \tag{3.4.66}$$

Now, considering a FIFO service discipline, i.e., one class of traffic with intensity $\lambda_1 = \lambda$, we see that Eq. (3.4.63) is exactly the same as (3.4.65) after using Little's result.

This result is quite important if buffer space is a major constraint because we are accommodating priority services with the same buffering resources.

REFERENCES

[KWO 80] R.H. KWONG, A. LEON GARCIA, A.N. VENETSANOPOULOS, A Study of a joint stochastic process for an integrated network, Research Report prepared for the dept of Comunications under Contract OSU79-00041. Univ. of Toronto, Toronto, Canada

[SCH 77] M. SCHWARTZ, M.A. RICH, Buffer Sharing in Computer Communication Network Nodes, IEEE Trans. on Comm., Sept. 77.

[FIS 76] M.J. FISHER, T.C. HARRIS, A Model for Evaluating the Performance of Integrated Circuit and Packet Switched Multiplex Structure, IEEE Trans. on Comm., Feb. 76.

[OCC 77] B. OCCHIOGROSSO, I. GITMAN, H. HSIEH, H. FRANK, Performance Analysis of Integrated Switching Communication Systems, Proc. Nat. Telecom. Conf., 1977.

[YEC 69] U. YECHIALI, P. NAOR, Queuing Problems with Heterogeneous Arrivals and Service, Oper. Res., vol. 19, May-June 71.

[FIS 77] M.J. FISHER, A Queuing Analysis of An Integrated Telecommunications System with Priorities, INFOR, vol. 15, no.3, Oct. 77.

[NAR 76] U. NARAYAN BHAT, M.J. FISHER, Multichannel Queuing Systems with Heterogeneous Classes of Arrivals, Office of Naval Research, Arlington, Va., June 76.

[GRU 81] J.G. GRUBER, Delay related issues in integrated voice and data networks, IEEE Trans. on Comm., June 81.

[MAG 81] B. MAGLARIS, D. WATTERS, State of the art in packet and circuit switching Comm. Tech. Seminars - Princeton, March 81.

[MAG 81] B. MAGLARIS, K.S. NATARAJAN, Network architectures for integrated voice and data Communications, IEEE Seminar on Data Comm., New York, May 81.

[AHM 80] H. AHMADI, T.E. STERN, A new satellite multiple access technique for switching using combined fixed and demand assignment, NTC 80.

[MAG 79] B. MAGLARIS, Studies in integrated line and packet switched computer communication networks, Ph.D. Dissertation, E.E. Dept., Columbia University, 79.

[JAN 80] N. JANAKIRAMAN, Performance analysis of multiplexors in circuit, packet and integrated switching environment, Ph.D. dissertation, Carelton Univ., Ottawa, Aug. 80.

[STE 80] T.E. STERN, M. SCHWARTZ, H.E. MEADOWS et al., Next generation communications satellites: Multiple access and network studies, First year Final Report prepared for NASA, under Contract NAS-5-25759

[SHA 81] J.G. Shanthikumar, "On the Buffer Behavior with Poisson Arrivals Priority Service and Random Server Interruptions," Inter. Conf. on Comm., Denver, 1981.

[WEI 80] C.J. WEINSTEIN, M.L. MALPASS, M.J. FISHER, Data traffic performance of an integrated circuit and packet switch multiplexing structure, IEEE Trans. on Comm., June 80.

[YU 81] W. YU, J.W. WONG, J.C. MAJITHIA, A trunk organization for circuit/packet switching networks, Computer Communications Network Group, University of Waterloo, Ontario, Canada.

[KLE 75] L. KLEINROCK, Queuing systems, vol. I, J. Wiley 1975.

[FIS 76] M.J. FISHER and T.C. HARRIS, A Model for evaluating the performance of an integrated circuit- and packet-switched multiplex structure, IEEE Trans. on Communications, vol. COM-24, no. 2, pp. 195-202, February 1976.

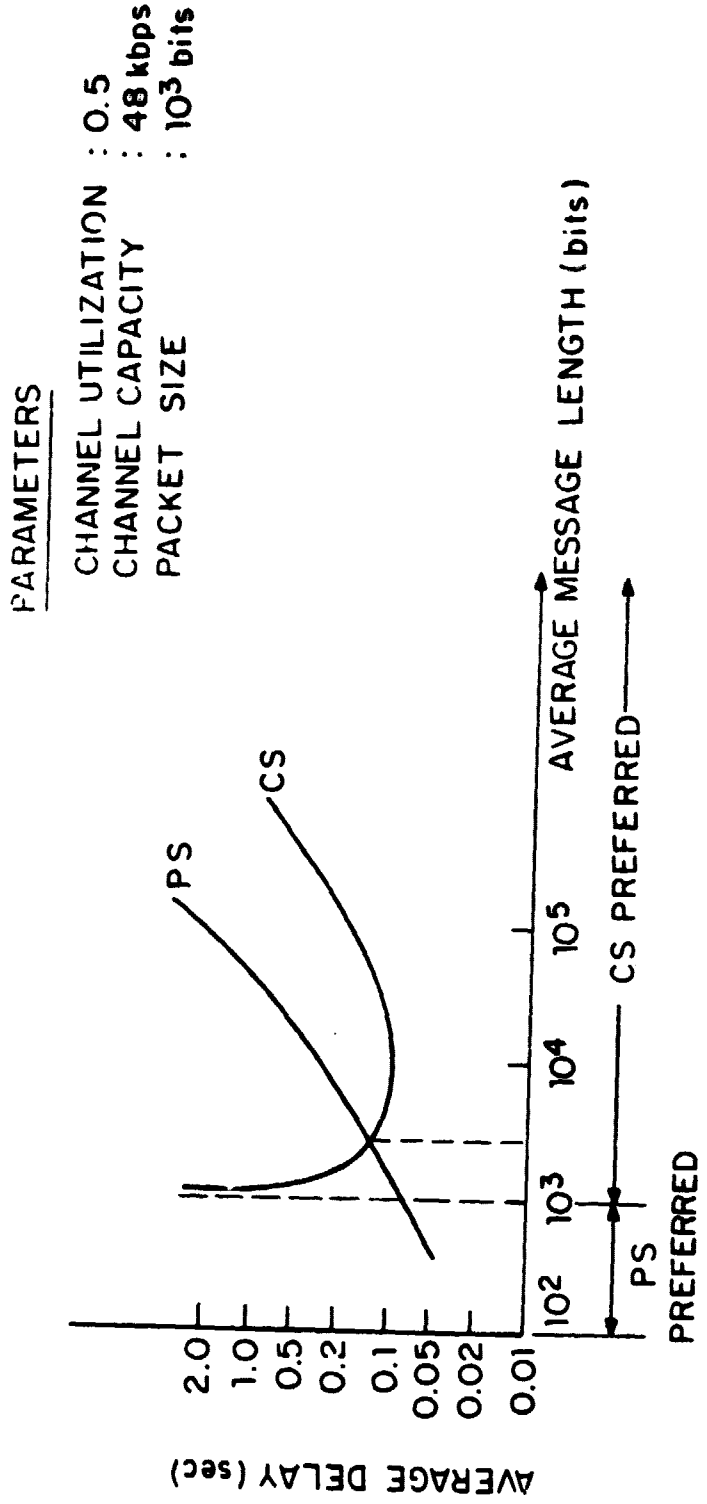


FIG. 3.1 COMPARISON OF PACKET & CIRCUIT SWITCHING

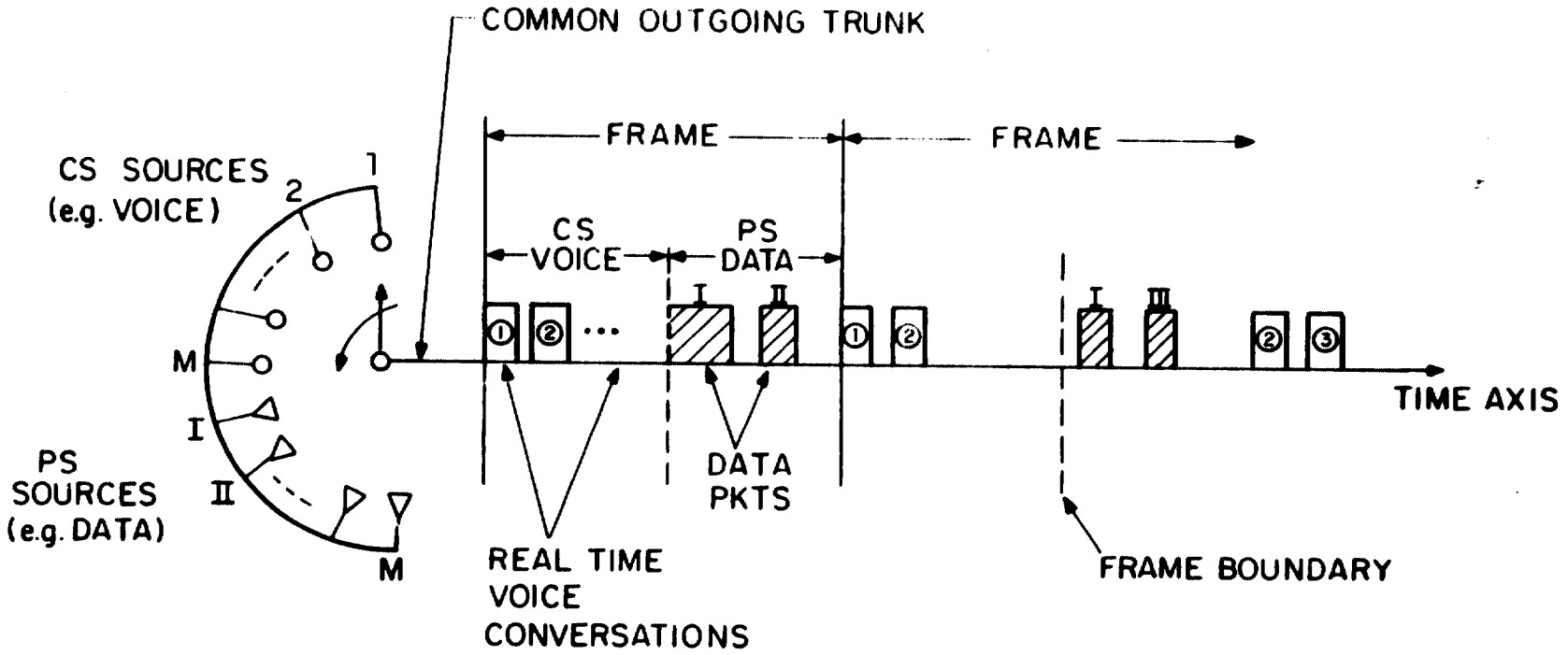


FIG. 3.2 INTEGRATED TDMA FRAME STRUCTURE

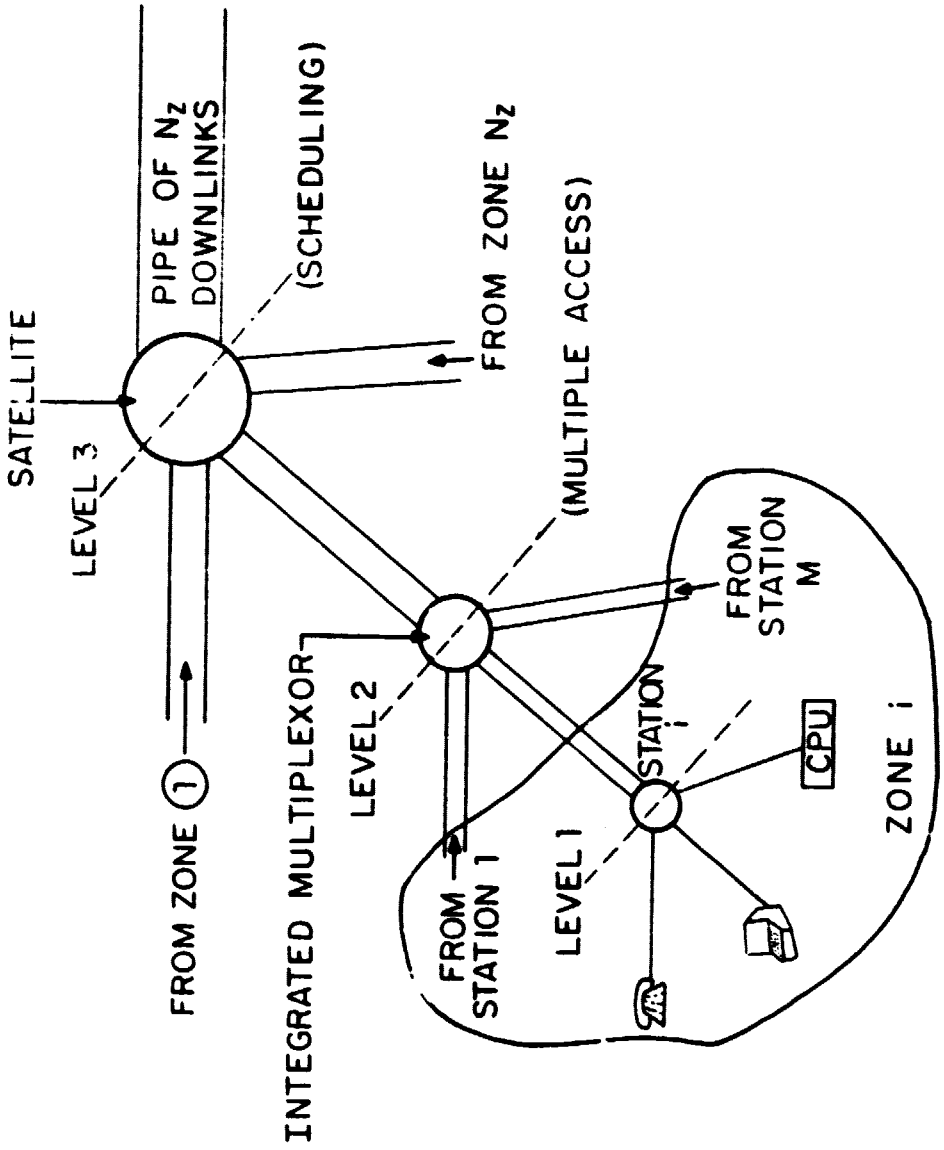


FIG. 3.3 ACESS CONTROL HIERARCHY OF THE SATELLITE SYSTEM

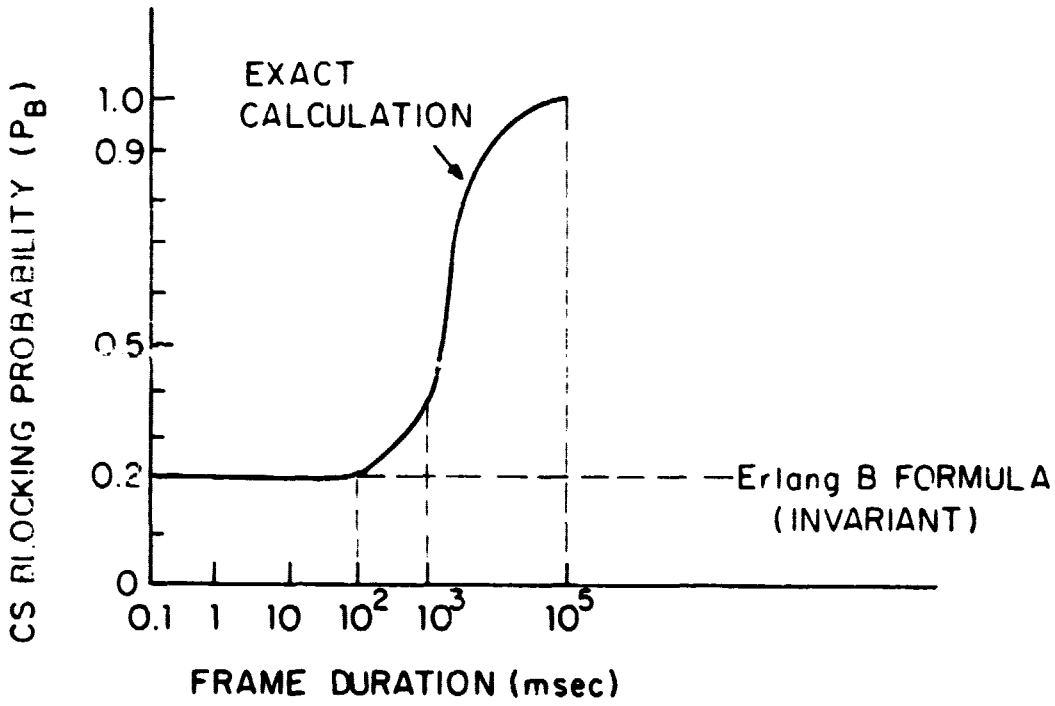


FIG. 3.4 VALIDITY OF THE ERLANG-B-FORMULA WITH FRAME DURATION

$P_r[ON] = \sigma$

$P_r[OFF] = 1 - \sigma$

PKT LENGTH= CONSTANT

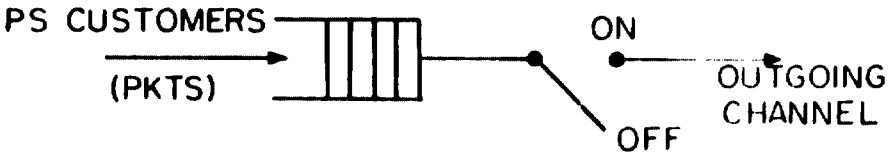
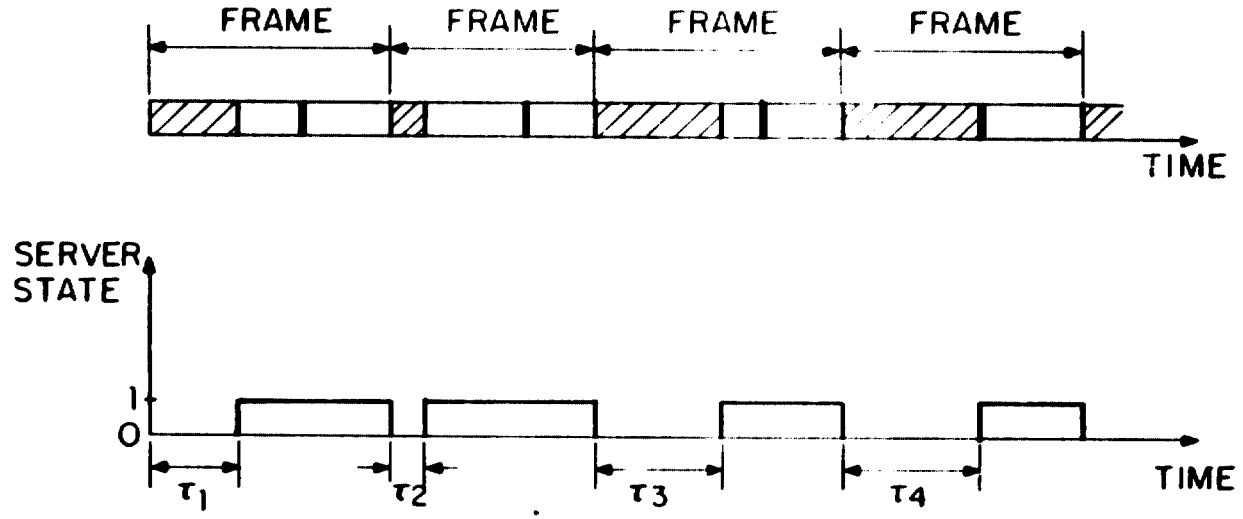


FIG. 3.5 PS TRANSMISSION MODEL; M/D/I WITH ON-OFF SERVER

A-EE-2366 JD

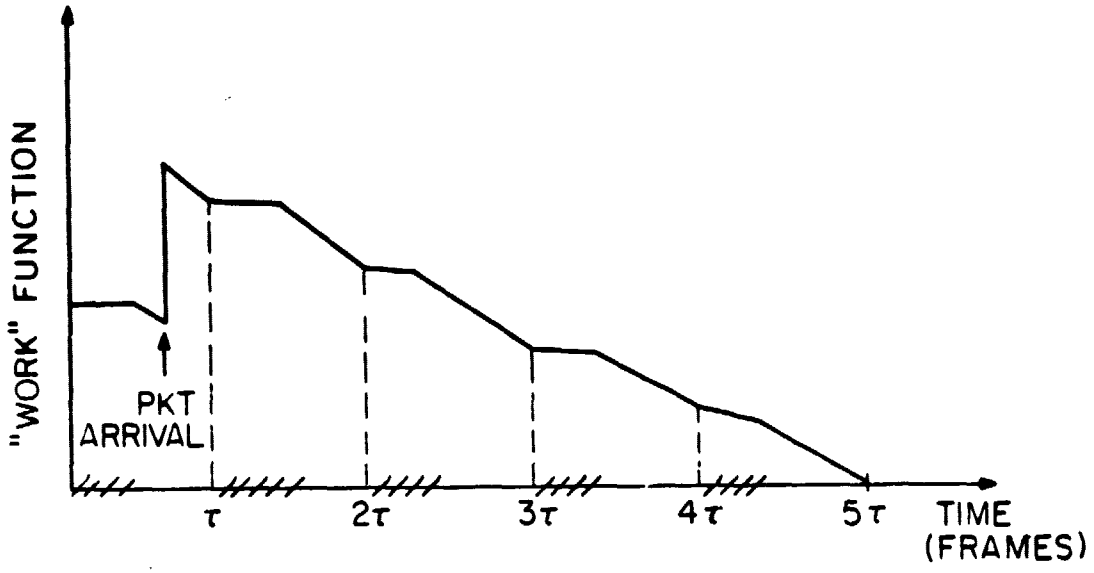


$(\tau_1, \tau_2, \tau_3, \tau_4; \cdots)$ SEQUENCE OF "BREAKDOWN" PERIODS
(SERVICE INTERRUPTIONS)

STATE 0: SERVICE INTERRUPTED

STATE 1 : SERVICE RESUMED

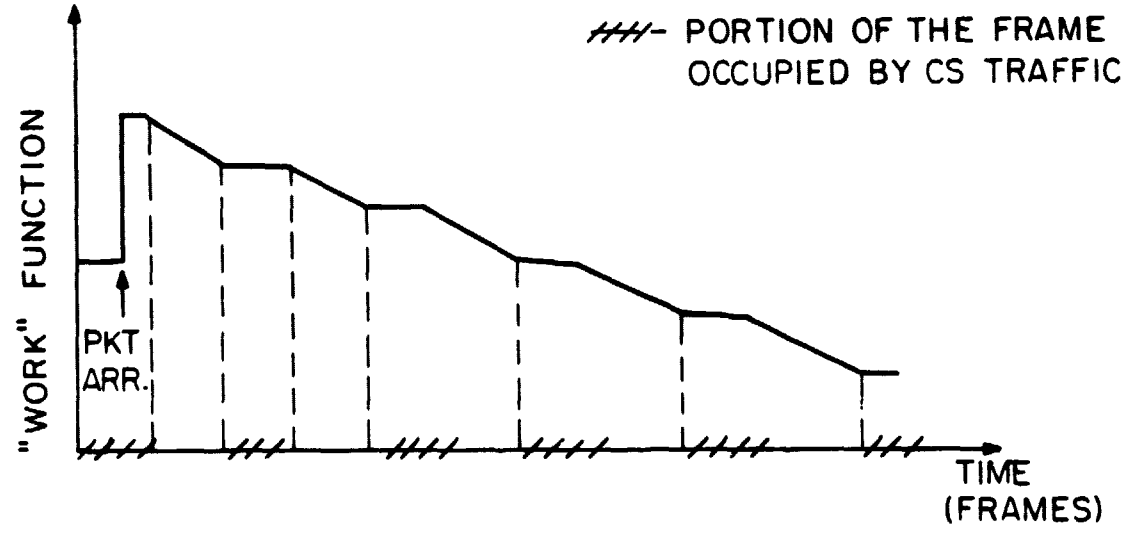
FIG. 3.6 PS TRANSMISSION MODEL; SINGLE SERVER SUBJECT TO "BREAKDOWNS"



(a) PACKET ARRIVES DURING PS PORTION OF THE FRAME

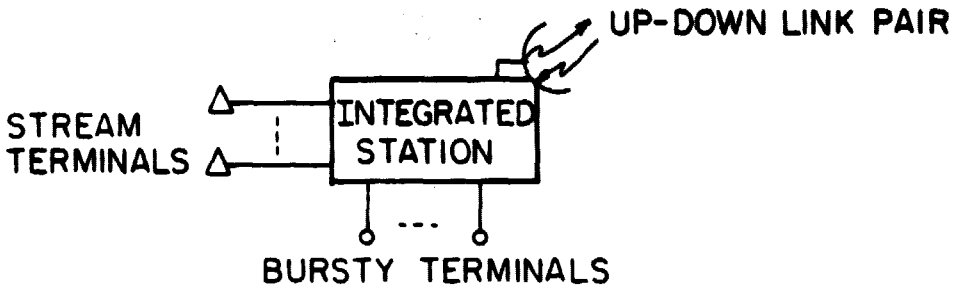
τ - FRAME DURATION IN SEC

//// - PORTION OF THE FRAME OCCUPIED BY CS TRAFFIC



(b) PACKET ARRIVES DURING CS PORTION OF THE FRAME

FIG. 3.7 SAMPLE REALISATIONS OF THE "WORK" FUNCTION



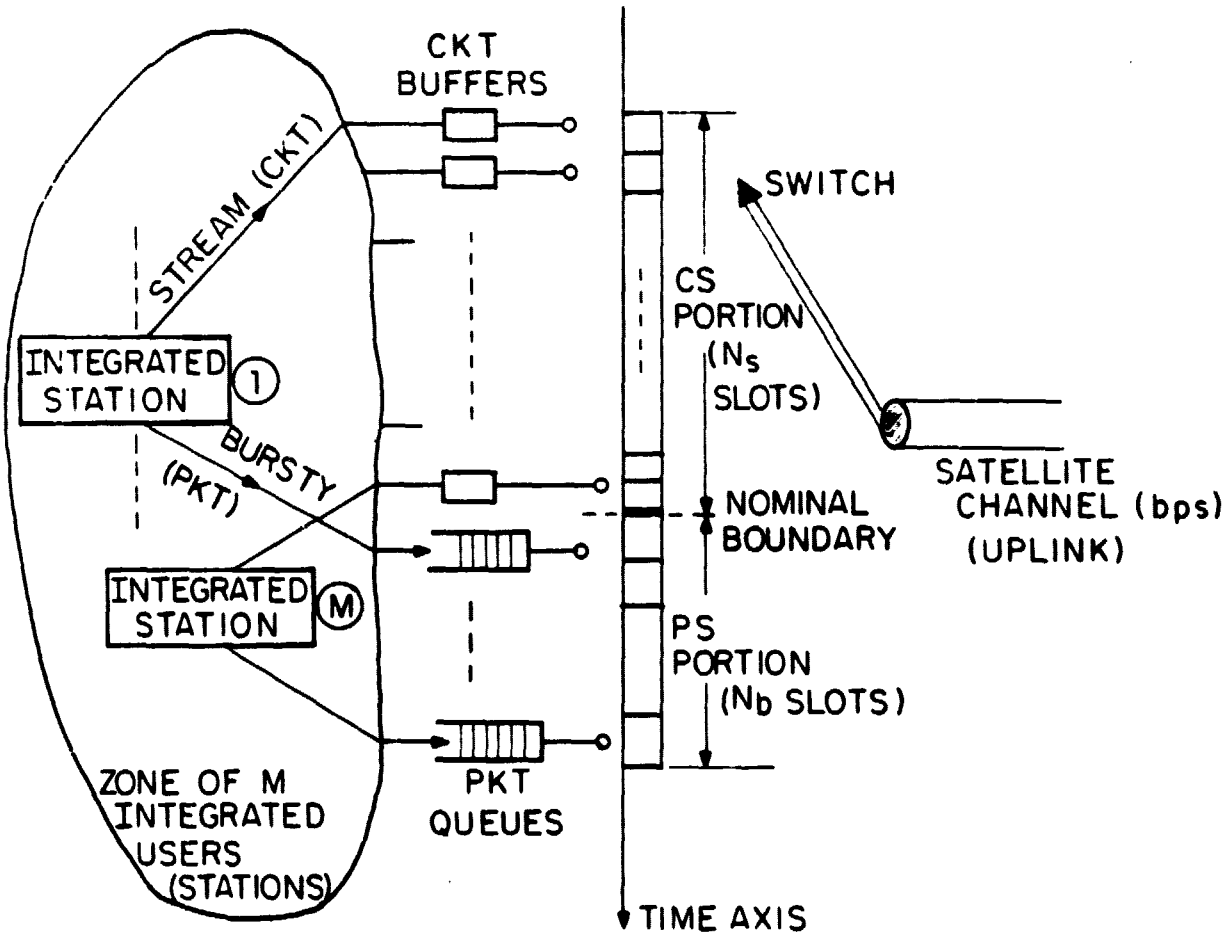


FIG. 3. 8 TDM ACCESS OF THE M ZONE STATIONS TO THEIR UPLINK CHANNEL

A-EE-2364 JD

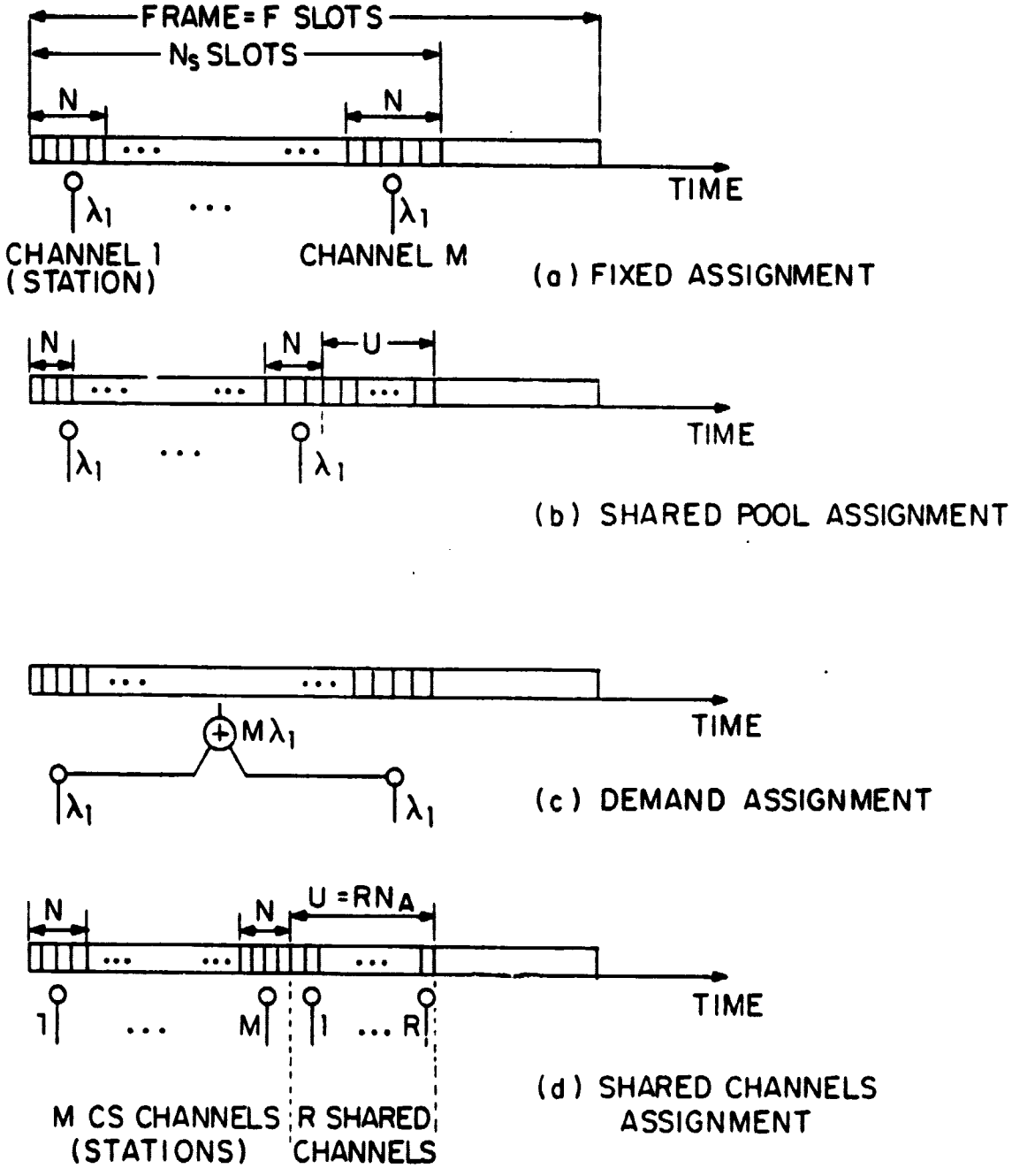


FIG. 3.9 FRAME CONFIGURATIONS OF THE 4 SLOTS ASSIGNMENT STRATEGIES

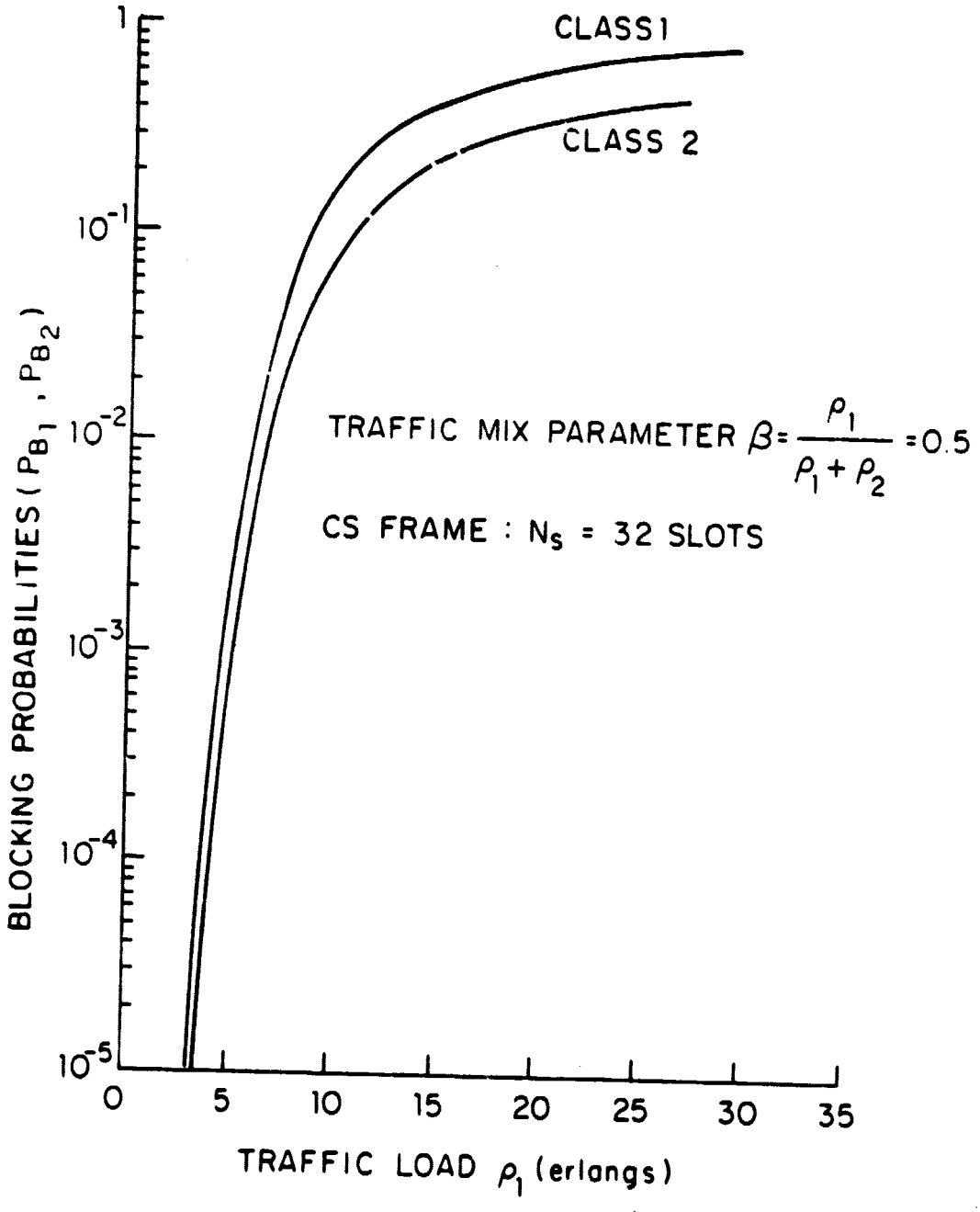


FIG. 3.10 BLOCKING PROBABILITIES VERSUS TRAFFIC LOAD

A-EE-2370 JD

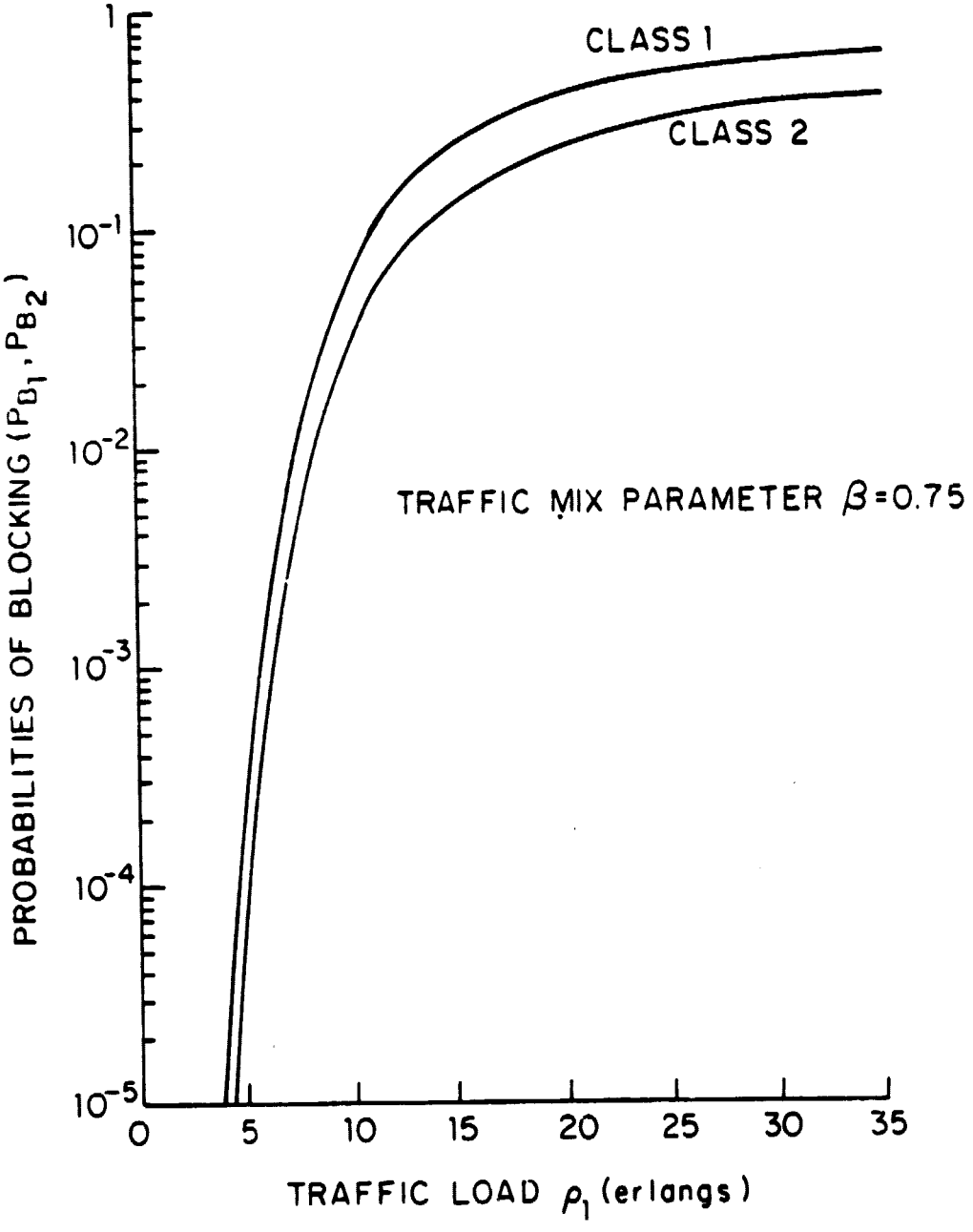


FIG. 3.11 BLOCKING PROBABILITIES VERSUS TRAFFIC LOAD

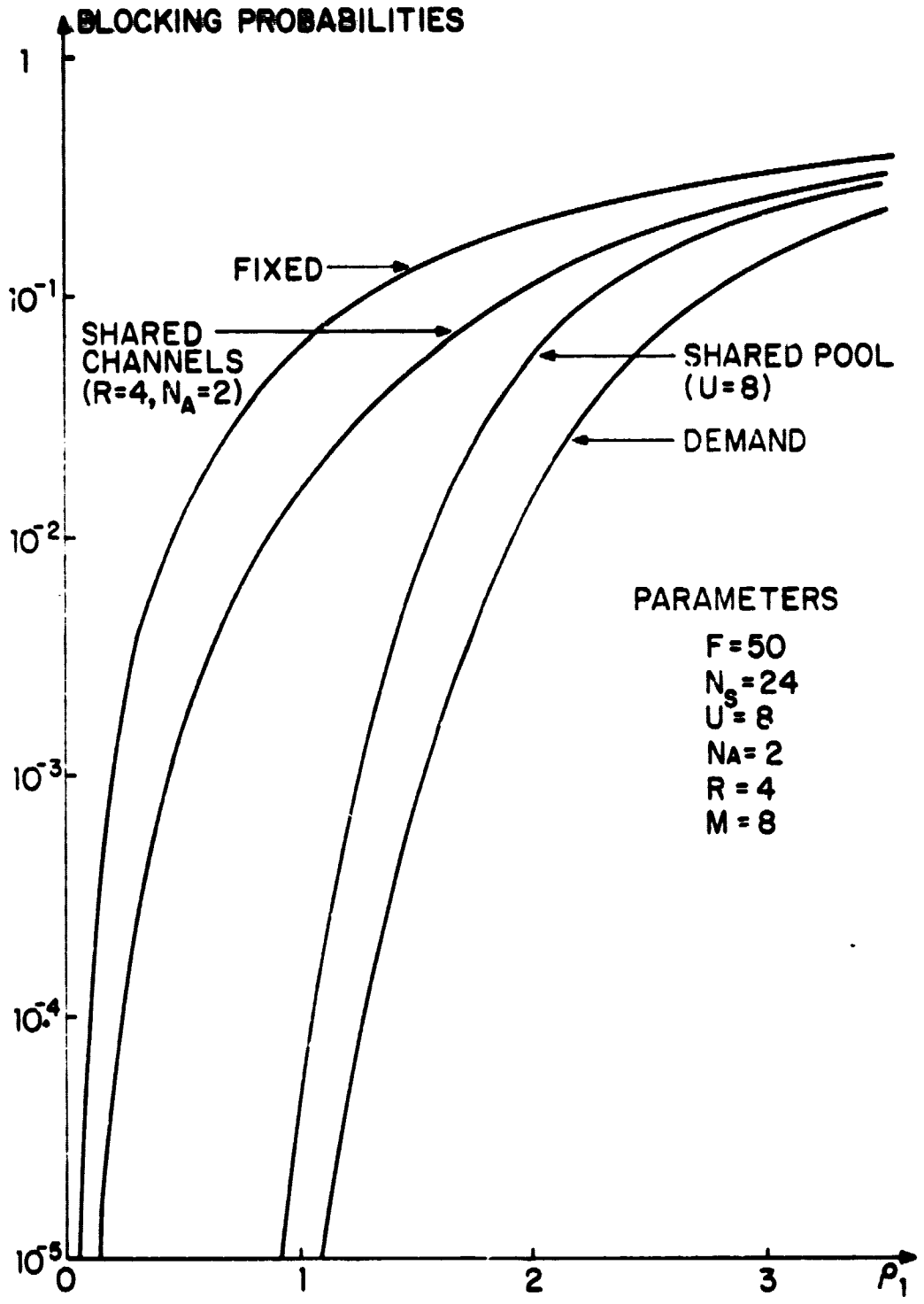


FIG. 3.12 CIRCUIT PERFORMANCE COMPARISON FOR FOUR SCHEMES.

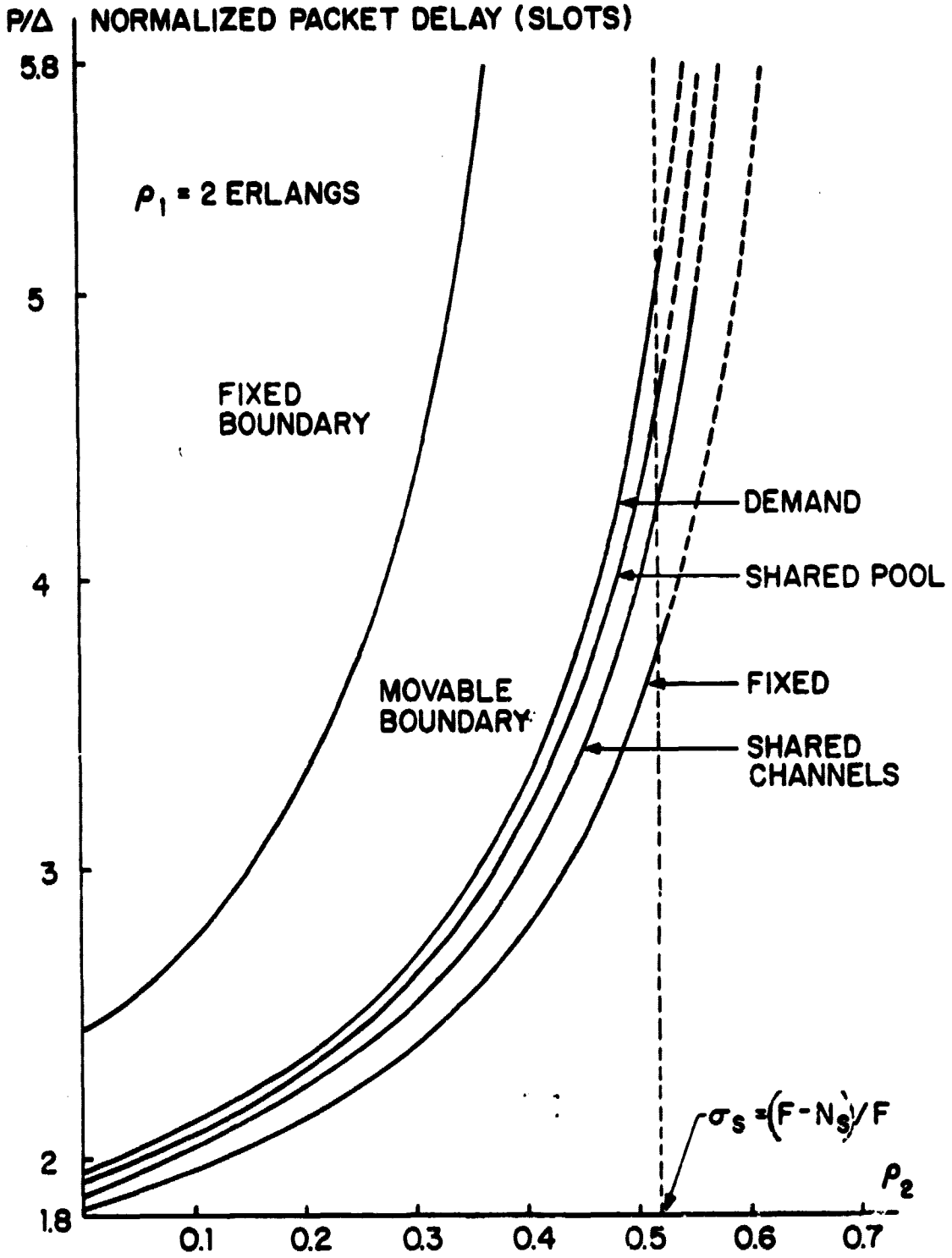


Fig. 3.13 PACKET PERFORMANCE COMPARISON FOR FOUR ASSIGNMENT SCHEMES.

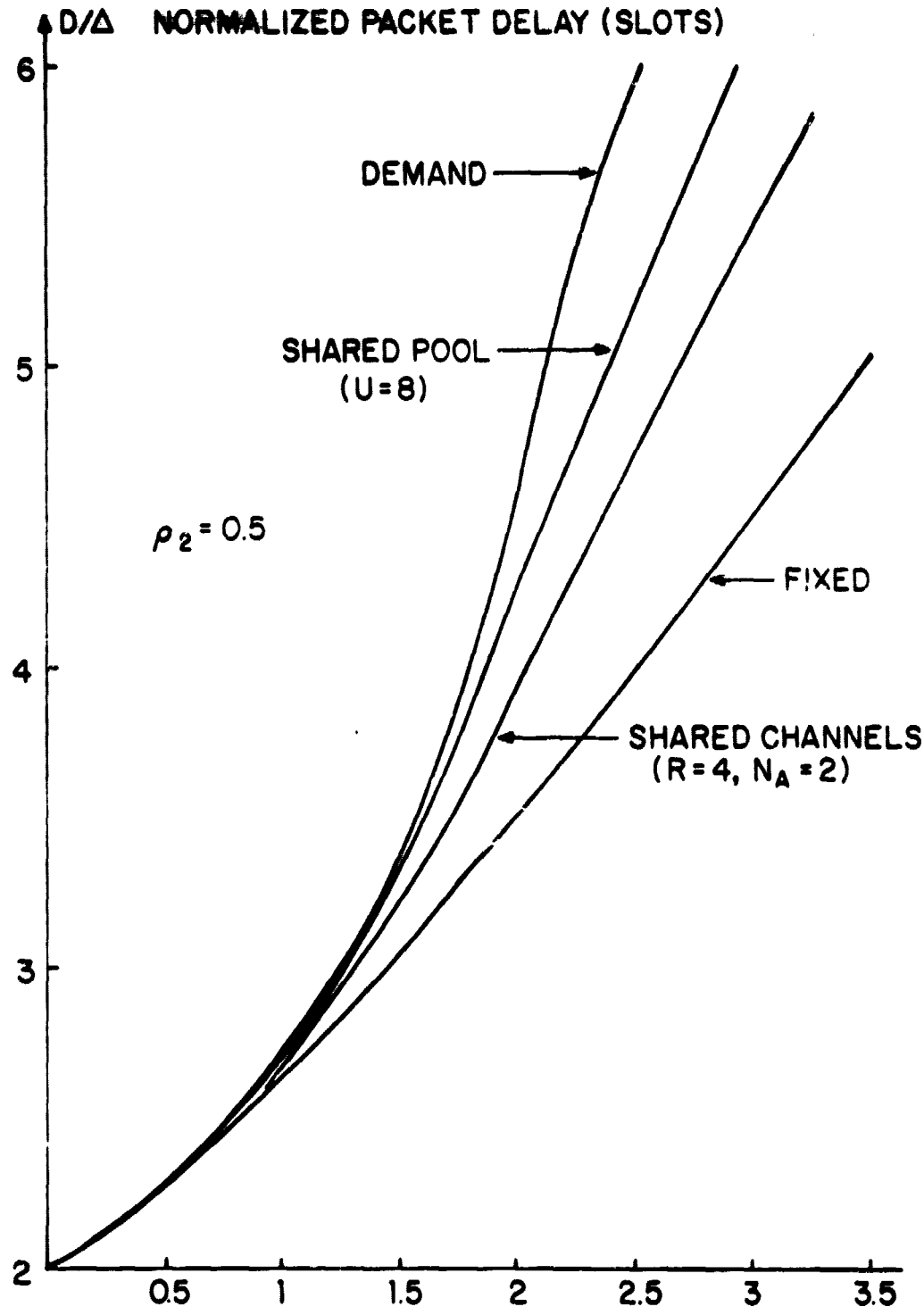
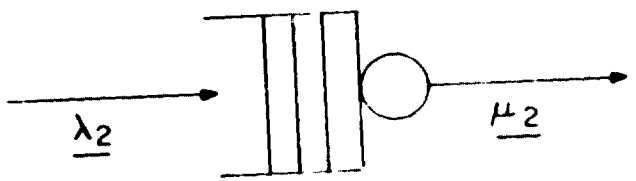


FIG. 3.14 PACKET PERFORMANCE VS. CIRCUIT-SWITCHED TRAFFIC.



$\underline{\lambda_2} = [\lambda_2(0), \lambda_2(1)]$

$\underline{\mu_2} = [\mu_2(0), \mu_2(1)]$

WHERE

$\lambda_2(i)$: PACKET ARRIVAL RATE WHEN THE SERVER OPERATES WITH CAPACITY $\mu_2(i)$, $i = 0,1$

FIG. 3.15 TWQ=DIMENSIONAL M/M/1 SYSTEM

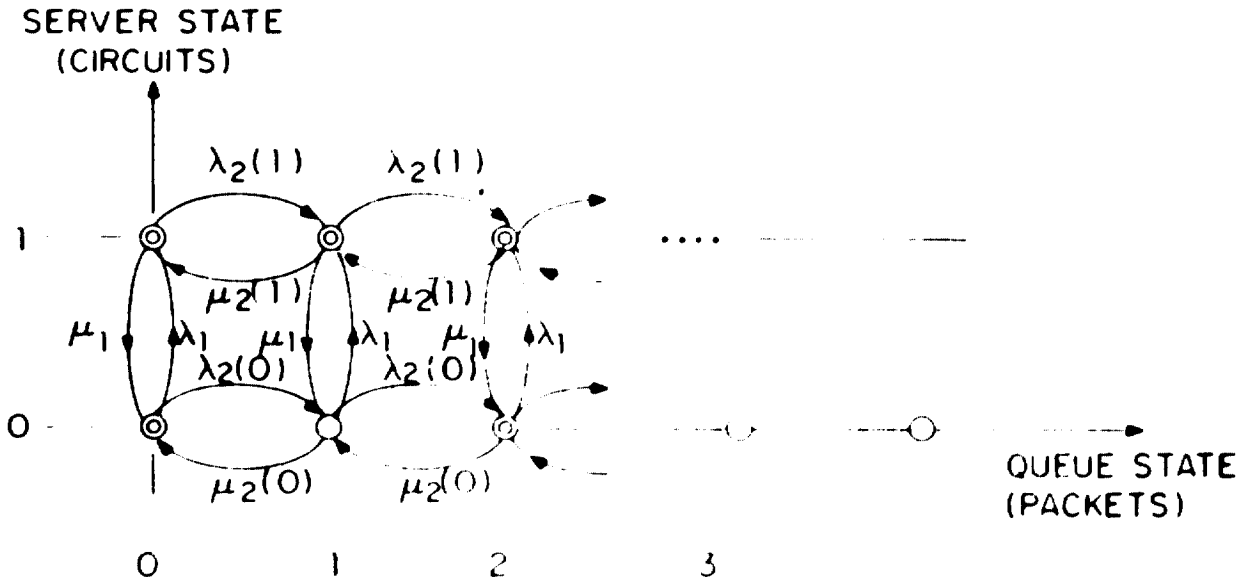


FIG. 3.16 TRANSITION STATE DIAGRAM OF THE TWO-DIMENSIONAL M/M/I SYSTEM

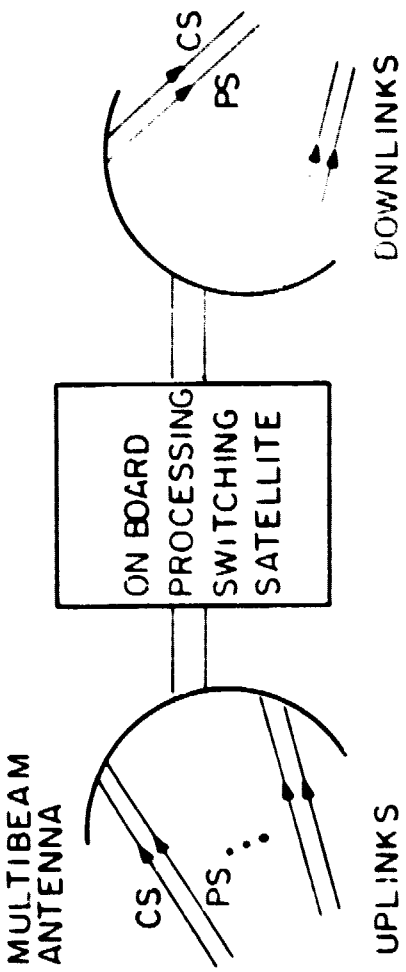


FIG. 3.17 INTEGRATED STORE-AND-FORWARD SATELLITE

A-EE-2358 JD

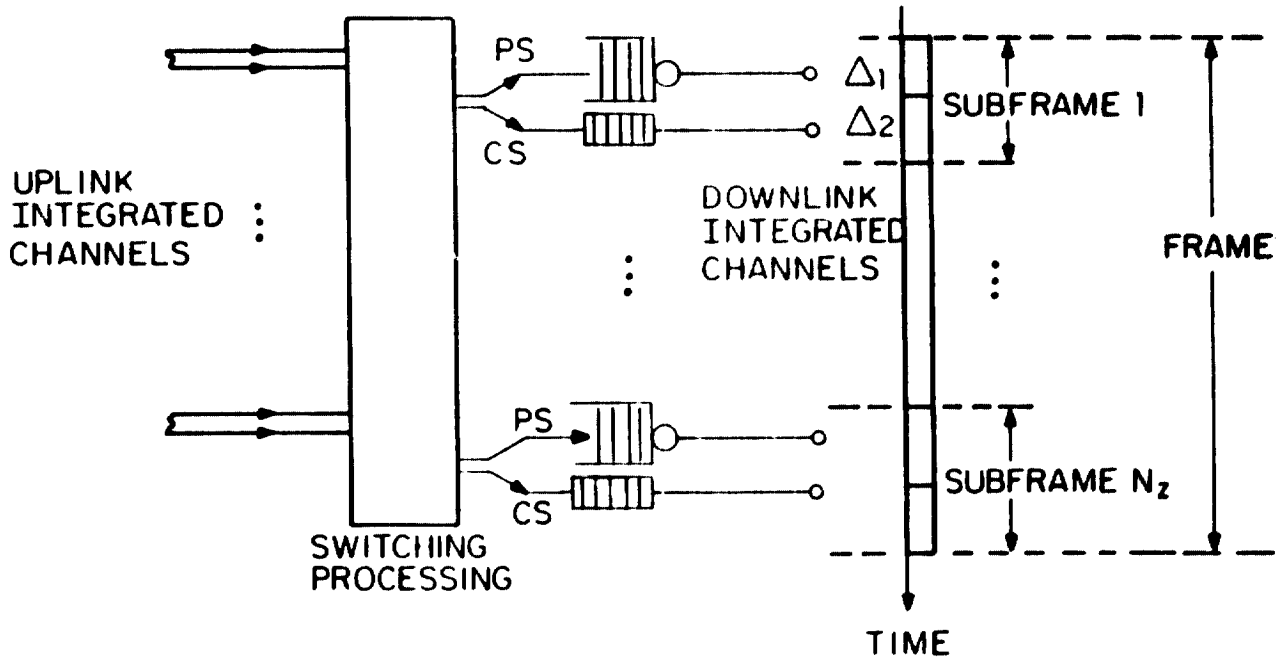
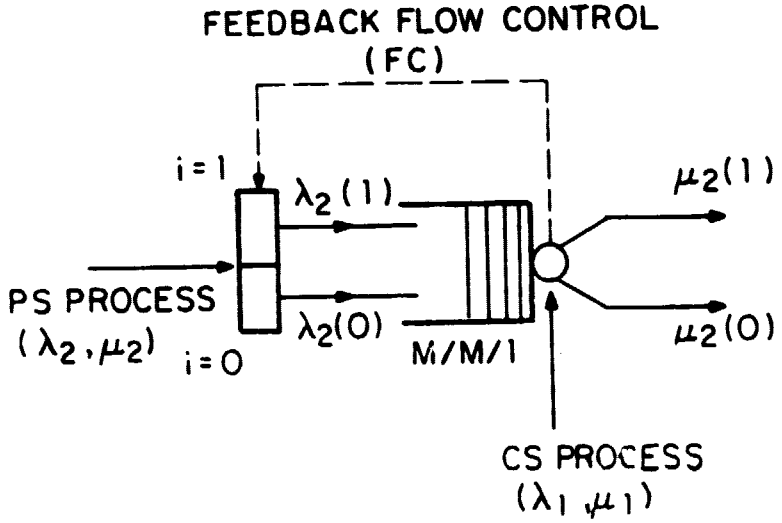


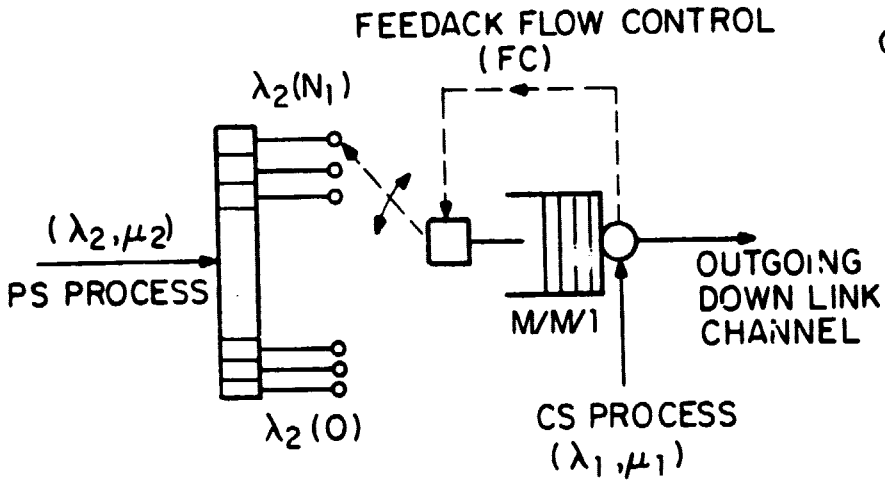
FIG. 3.18 INTEGRATED ACCESS OF THE DOWNLINK CHANNELS



CONDITION

$$\frac{\lambda_2(1)}{\mu_2(1)} = \frac{\lambda_2(0)}{\mu_2(0)}$$

(a) CIRCUIT SUBCHANNEL HAS ONE SLOT



CONDITION

$$\frac{\lambda_2(i)}{\mu_2(i)} = \frac{\lambda_2(i+1)}{\mu_2(i+1)}$$

FOR $i = 0, N_1 - 1$

(b) CIRCUIT SUBCHANNEL HAS N_1 SLOTS

FIG. 3.19 FEEDBACK FLOW CONTROL MECHANISM

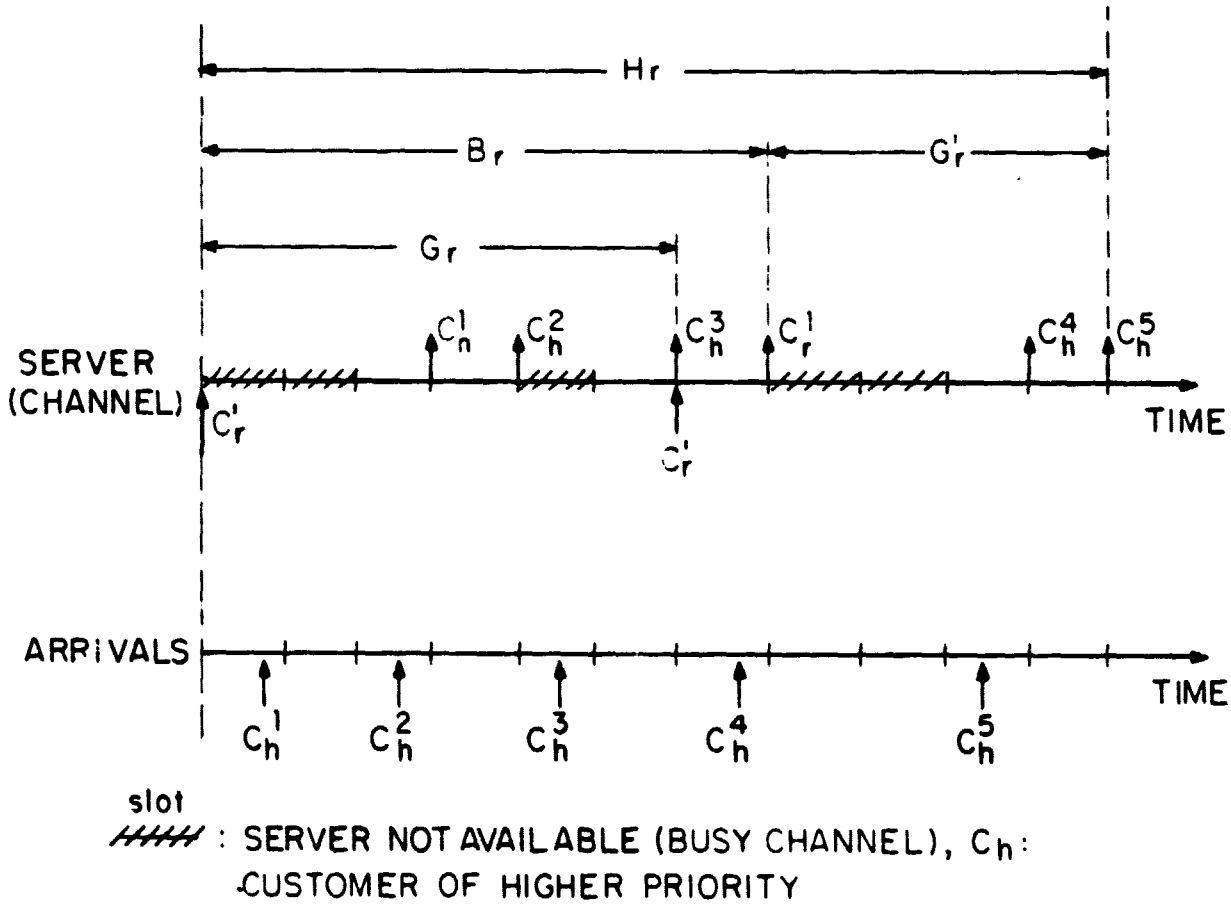
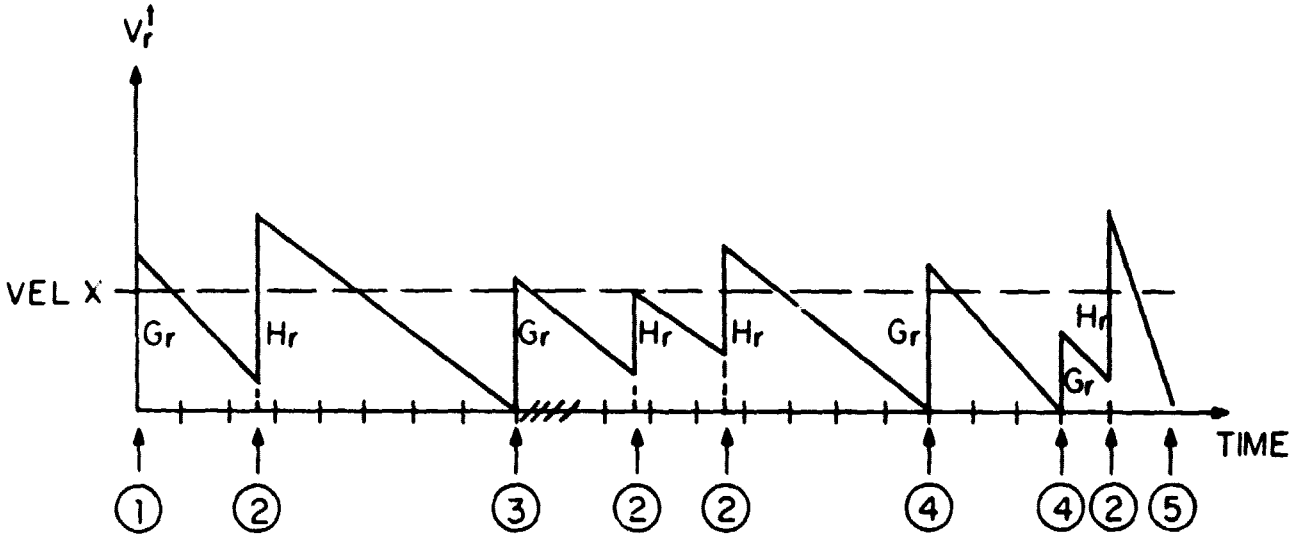


FIG. 3.20 SAMPLE SERVICE AND ARRIVAL REALIZATION



(1) CYCLE BEGINS AT $t = 0$ (EMPTY BUFFER)
(2) JUMPS OF V_r^t BY H_r CAUSED BY r-CUSTOMER ARRIVALS
(3) LOWER PRIORITY CUSTOMER's ATTEMPT FOR SERVICE (FAILURE)
(4) LOWER PRIOPITY CUSTOMER's ATTEMPT FOR SERVICE(SUCCESS)
(5) BUFFER EMPTY : END OF CYCLE

FIG. 3.21 SAMPLE WAITING TIME PROCESS REALIZATION

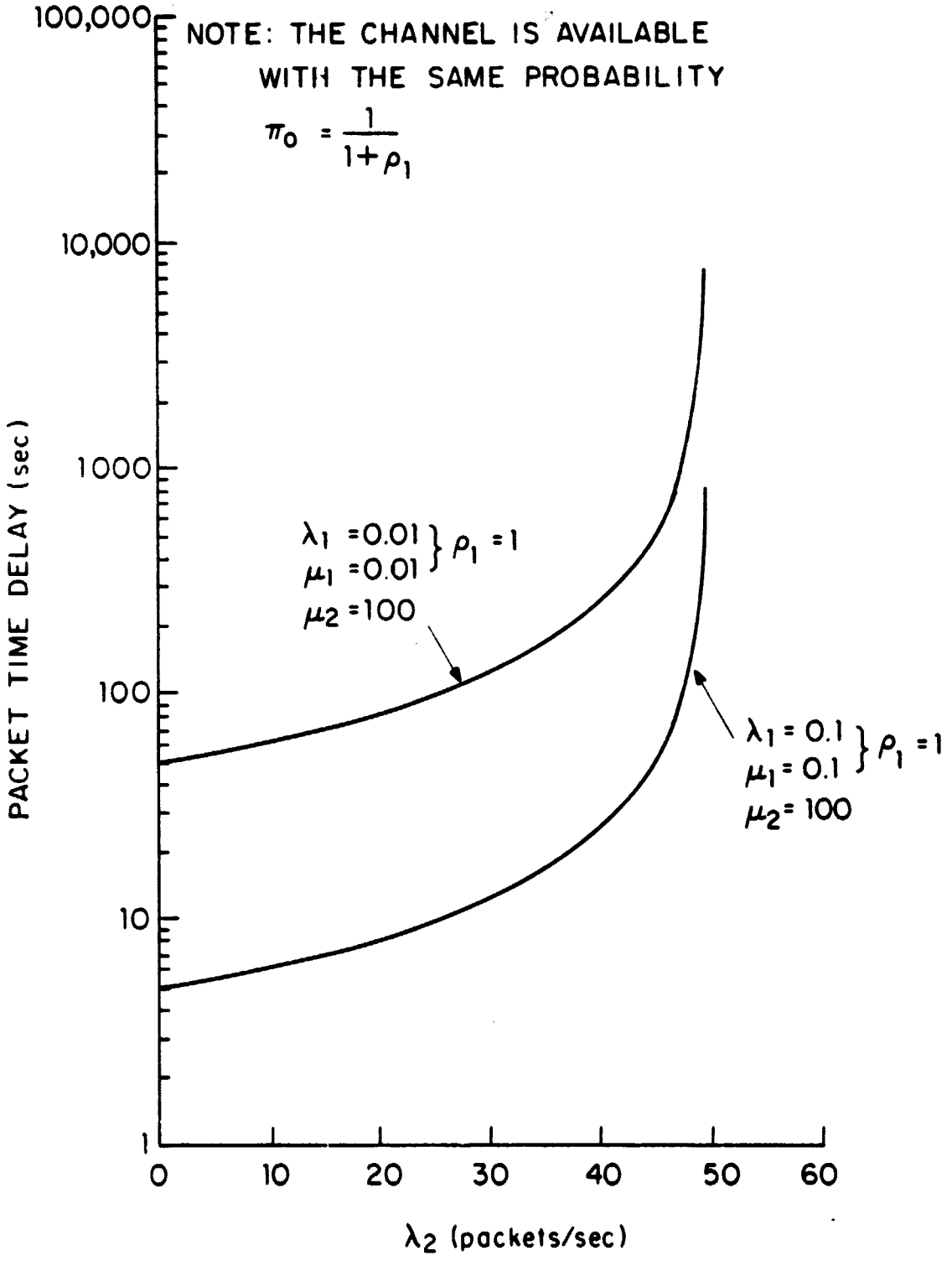


FIG. 3.22 EFFECT OF CS PARAMETERS (λ, μ) ON PS PERFORMANCE

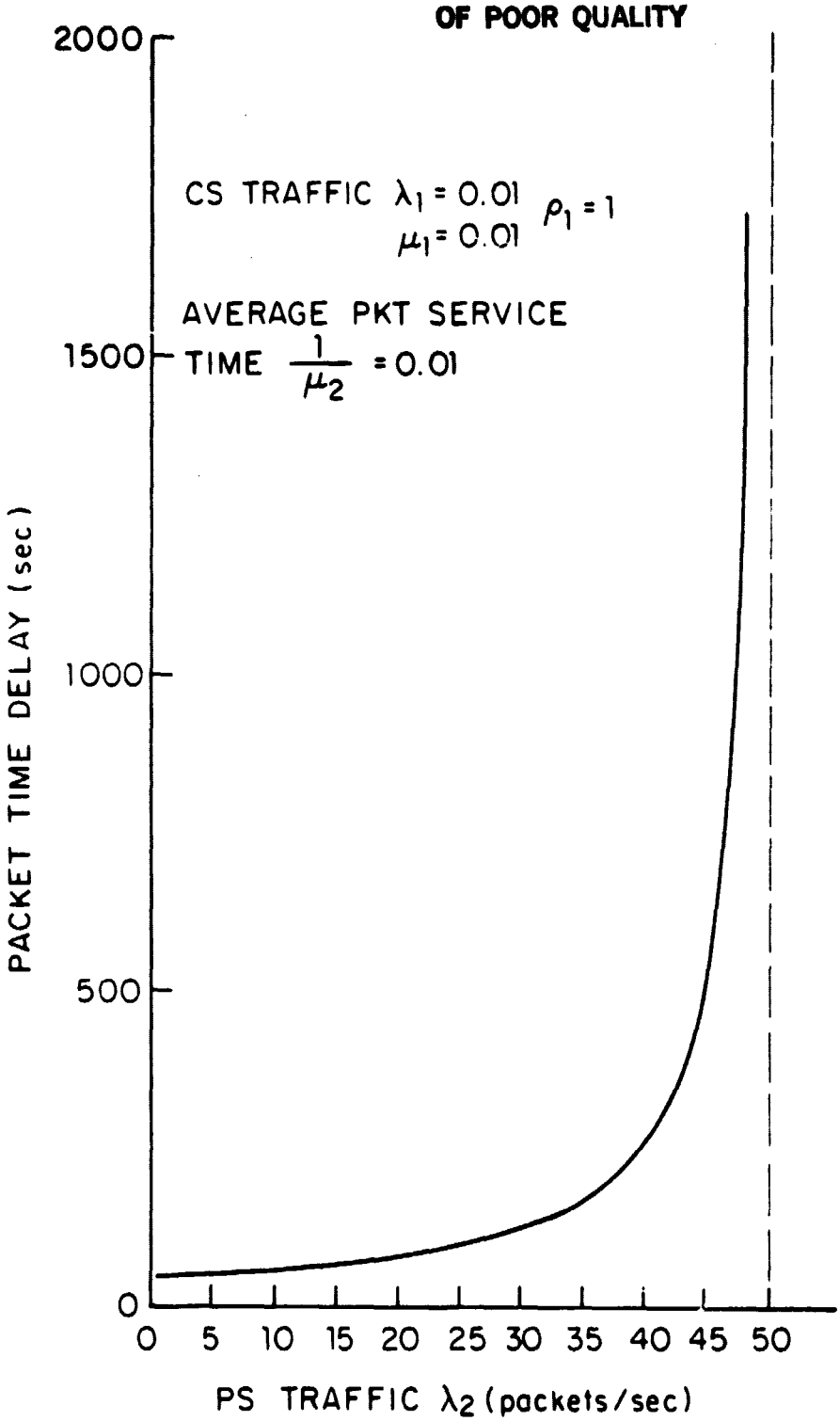


FIG. 3.23 PACKET PERFORMANCE VERSUS PS TRAFFIC

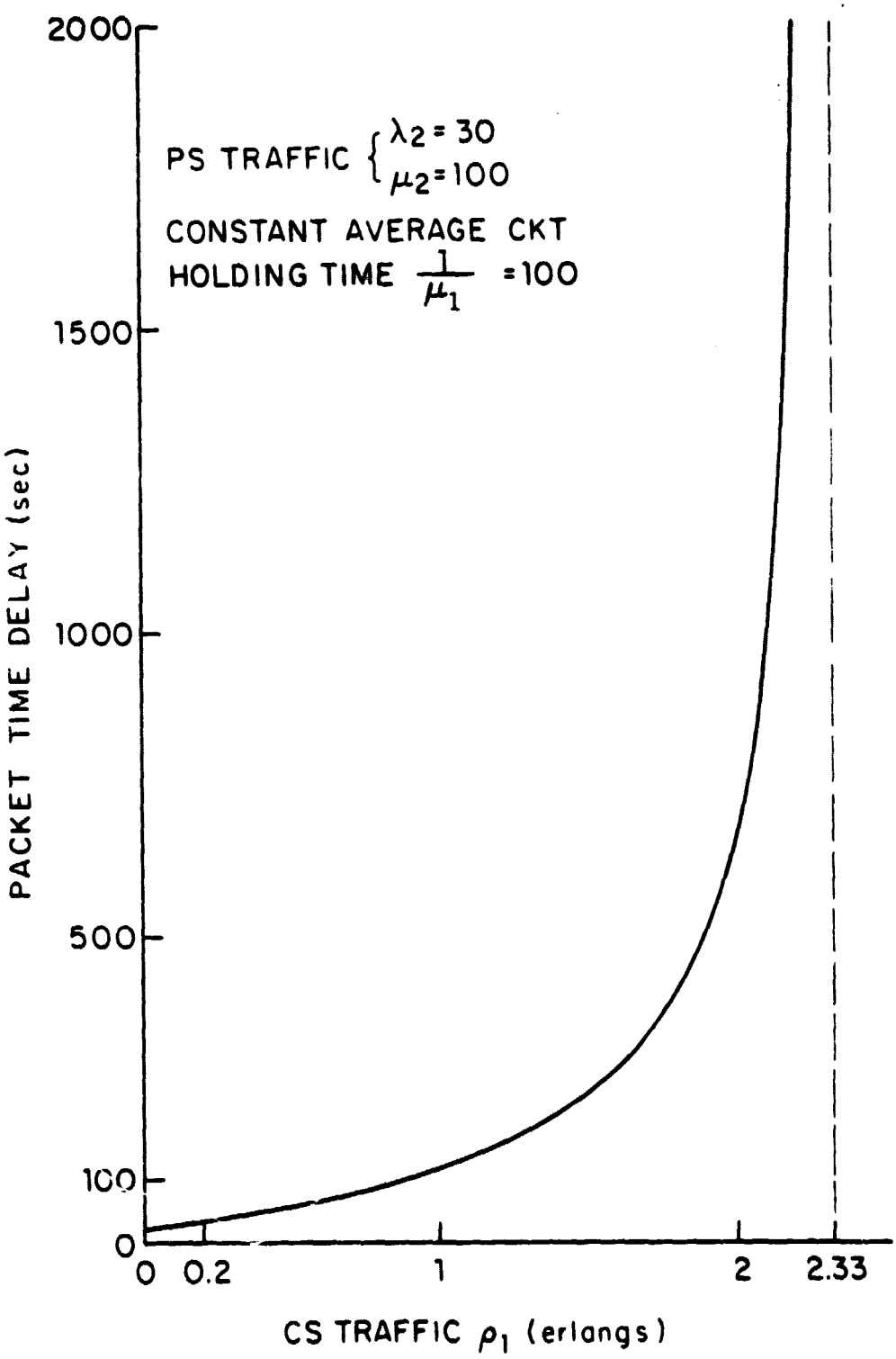


FIG. 3.24 PACKET PERFORMANCE VERSUS CS TRAFFIC

APPENDIX A

TWO-DIMENSIONAL MARKOV CHAIN MODEL OF THE INTEGRATED TDMA FRAME

The state transiton diagram for the joint CS/PS process arising in an integrated TDMA frame Fig. A.1 is given in Fig. A.2. Taking into account boundary conditions, balance equations can be written by equating the rate of entering a state and the rate of leaving that state. This leads to the following set of steady state difference equations [KWO 80].

- For $i < N_S$, we have

$$\left[(\lambda_1+\lambda_2+i\mu_1+j\mu_2)\right]P_{ij} = \lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j}+\lambda_2 P_{i+1,j}+\lambda_2 P_{i,j-1}+(j+1)\mu_2 P_{i,j+1} \quad (A.1)$$

$$\text{for } j = 0,1,\ldots\ F-i-1$$

$$\left[\lambda_1+\lambda_2+i\mu_1+(F-i)\mu_2\right]P_{ij} = \lambda_1 P_{i-1,j}+(i+1)\mu_1 P_{i+1,j}+\lambda_2 P_{i,j}+(F-i)\mu_2 P_{i,j+1}$$

$$\text{for } F-i \le j \le B+F-i-1 \quad (A.2)$$

$$\left[\lambda_1+i\mu_1+(F-i)\mu_2\right]P_{ij} = \lambda_1 P_{i-1,j}+\lambda_2 P_{i,j-1}+\lambda_1 P_{i-1,j+1} \quad \text{for } j = B+F-i \quad (A.3)$$

- For $i = N_S$, we have

$$(\lambda_2+N_S\lambda_1+j\mu_2)P_{N_S,j} = \lambda_1 P_{N_S-1,j} + \lambda_2 P_{N_S,j-1}+(j+1)\mu_2 P_{N_S,j+1} \quad (A.4)$$

$$\text{for } j = 0,1,\ldots\ N_B-1$$

$$(\lambda_2+N_S\mu_1+N_B\mu_2)P_{N_S,j} = \lambda_1 P_{N_S-1,j}+\lambda_2 P_{N_S,j-1} + N_B\,\mu_2 P_{N_S,j+1} \quad (A.5)$$

$$\text{for } j = 0,1,\ldots,\ N_B-1$$

$$(N_S\mu_1+N_B\mu_2)P_{N_S,j} = \lambda_1 P_{N_S-1,j}+\lambda_2 P_{N_S,j-1}+\lambda_1 P_{N_S-1,j+1} \qquad (A.6)$$

$$\text{for } j = B+N_B$$

Together with the normalization condition $\sum_{i,j} P_{ij} = 1$, Eqs.(A.1) through (A.6) can be solved recursively to determine the steady-state joint probabilities P_{ij}. However, for a realistic case, the number of equations involved in the computations becomes quite large. This number is computed in the following way:

For $Q_s = i$, we have $j = 0,1,\ldots, B+F-i$. This yields $B+F-i+1$ equations. Therefore the total number of equations is

$$\sum_{i=0}^{N_s} (B+F+1-i) = \left(\frac{N_s+1}{2}\right)(2B + 2N_B + N_S + 2)$$

For illustration, an example with $N_S = 10$, $N_B = 5$, $B = 10$ requires 231 equations to compute the P_{ij}, from which the quantities of interest in the performance criteria are determined.

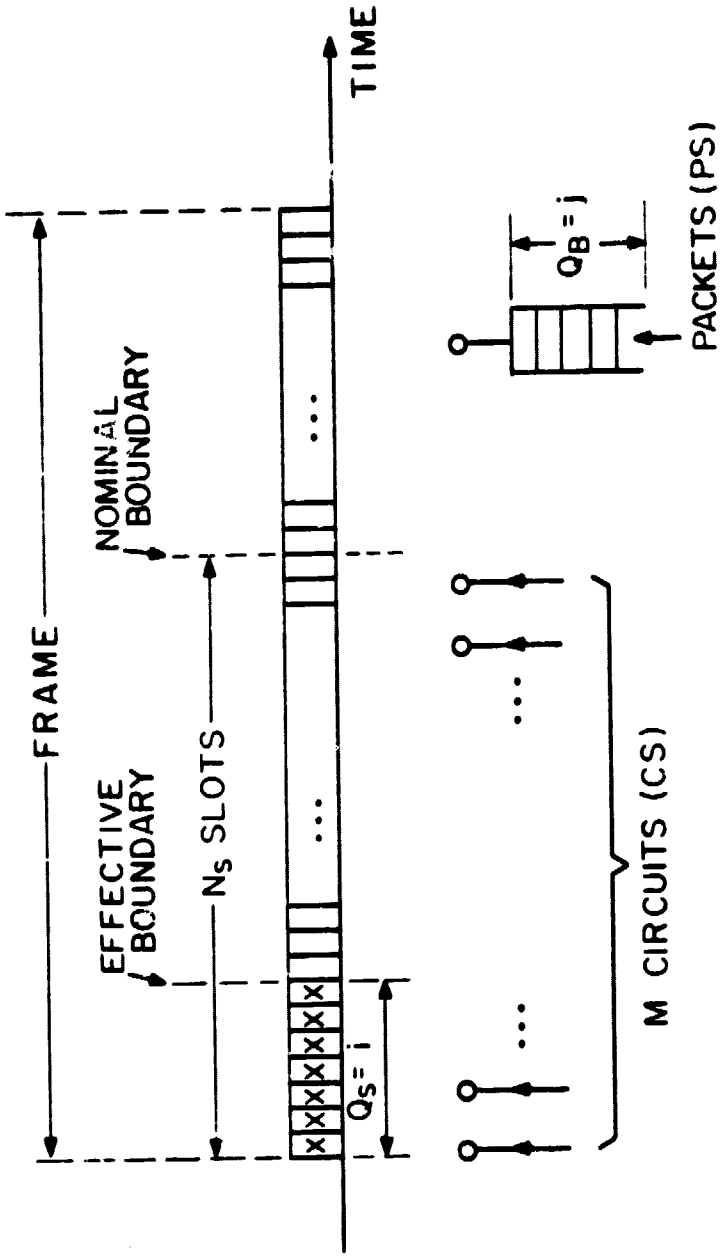


FIG. A.1 JOINT CS/PS INTEGRATED PROCESS

A-GC-2383 JD

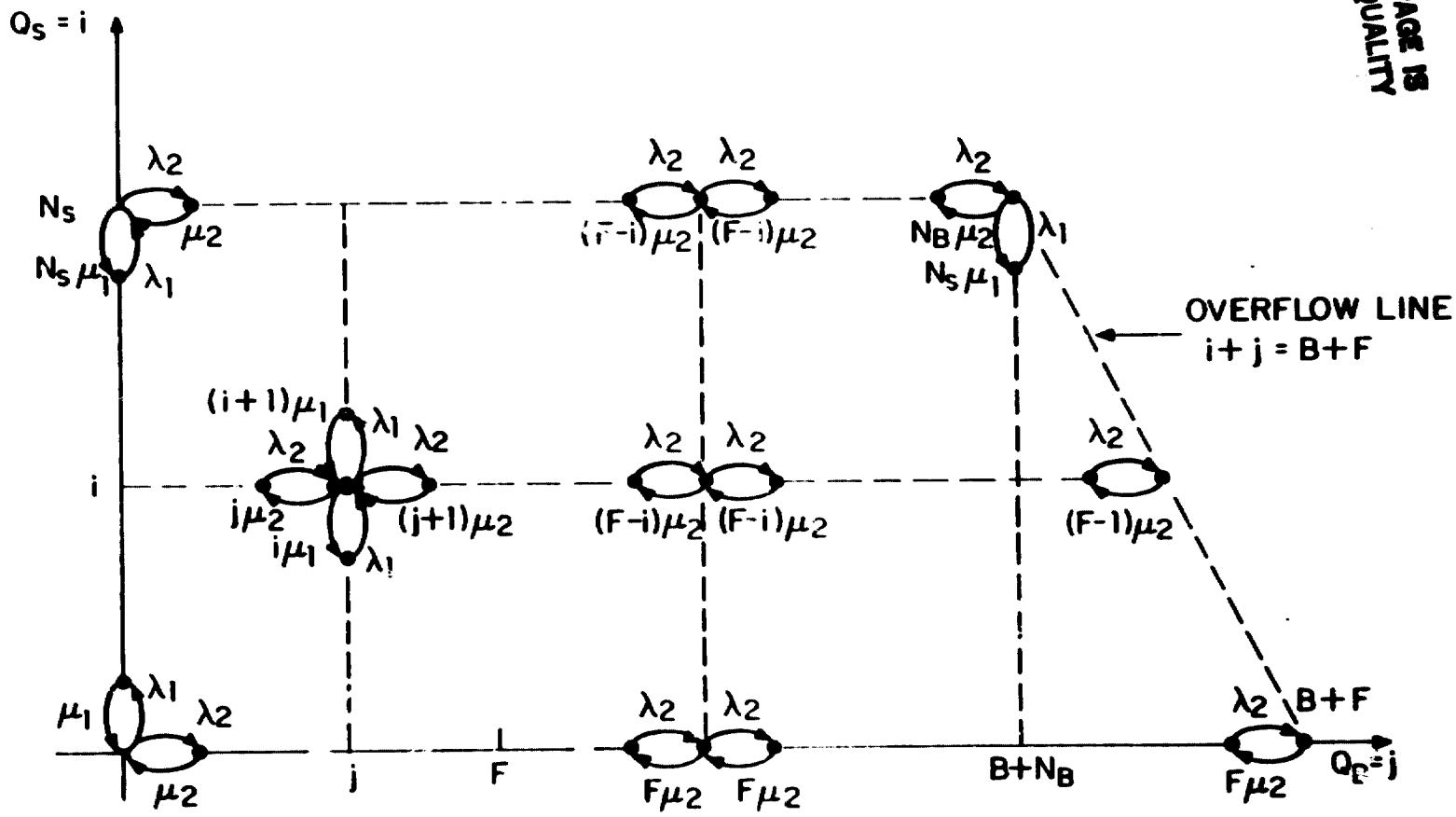


FIG. A.2 STATE TRANSITION DIAGRAM OF THE JOINT CS/PS PROCESS

APPENDIX B

ANALYSIS OF THE SHARED CHANNELS MODEL FOR THE CS SLOT ALLOCATION PROBLEM

The state of the queuing system presented in Section 3.1.1-d and depicted in Fig. 3.2(d) is described by the following closed form expression

$$P(n_1,n_2,\ldots,n_M)=P(\underline{n}) = P(\underline{o}) \prod_{i=1}^{M} \frac{\rho_i^{n_i}}{n_i!} \quad , \tag{B.1}$$

where $\rho_1 = \frac{\lambda_1}{\mu_1}$ and n_i is the # of circuits from staiton i in the system.

The existence of $P(\underline{n})$ can be proved by writing the following balance equation:

$$(M_u\lambda_1 + M_d\mu_1)P(n_1,\ldots,n_M) = \lambda_1 \underbrace{\left[P(n_1-1,n_2,\ldots,n_M) + P(n_1,n_2-1,\ldots,n_M)+\ldots\right]}_{M_d \text{ nonzero terms}}$$

$$+ \mu_1 \underbrace{\left[P(n_1+1,n_2,\ldots,n_M)+P(n_1,n_2+1,\ldots,n_M)+\ldots\right]}_{M_u \text{ nonzero terms.}} \tag{B.2}$$

It can readily be shown by substitution that Eq.(B.1) satisfies Eq.(B.2). Now, it remains to determine the normalization constant $P(\underline{o})$. This is done by summing $P(\underline{n})$ over all possible states, and setting the sum to 1, that is:

$$\sum_{S_0} P(\underline{n}) + \sum_{S_1} P(\underline{n}) + \ldots + \sum_{S_R} P(\underline{n}) = 1 \tag{B.3}$$

where state S_K means that K shared channels (out of R) are occupied.

The probability of the system being in state S_k, using symmetry, is

$$\sum_{S_k} P(\underline{n}) = \binom{M}{k} \sum_{n_1=N+1}^{N+N_A} \cdots \sum_{n_k=N+1}^{N+N_A} \sum_{n_{k+1}=0}^{N} \cdots \sum_{n_m=0}^{N} \left\{ \prod_{i=1}^{M} P(\underline{o}) \frac{\rho_1^{n_i}}{n_i!} \right\}$$

$$= P(\underline{o}) \binom{M}{k} \left\{ \prod_{i=1}^{k} \left(\sum_{n_i=N+1}^{N+N_A} \frac{\rho_1^{n_i}}{n_i!} \right) \prod_{i=K+1}^{M} \left(\sum_{n_i=0}^{N} \frac{\rho_1^{n_i}}{n_i!} \right) \right\}$$

$$= P(\underline{o}) \binom{M}{k} P^k q^{M-k} \tag{B.4}$$

where

$$P = \sum_{n_i=N+1}^{N+N_A} \frac{\rho^{n_i}}{n_i!} \tag{B.5}$$

and

$$q = \sum_{n_i=0}^{N} \frac{\rho_1^{n_i}}{n_i!} , \tag{B.6}$$

Therefore, using Eqs.(B.3) and (B.4), we get the following expression for $p(\underline{o})$:

$$p(\underline{o}) = \frac{1}{\sum_{k=0}^{R} \binom{M}{k} P^k q^{M-k}} \tag{B.7}$$

Now we proceed to calculate the probability that an arriving CS customer from one of the M stations (users) is blocked. This occurs in one of the two situations:

- The station has all its N dedicated slots busy and all the R channels are occupied by the (M-1) other stations.
- The station is using one shared channel (N_A slots) and all ($N+N_A$) slots are busy.

Without loss of generality, a customer from station 1 is blocked with probability P_B, obtained by summing over <u>all</u> blocking states.

$$P_B = \binom{M-1}{0} \sum_{n_2=0}^{N} \cdots \sum_{n_M=0}^{N} P(N+N_A, n_2, \ldots, n_M)$$

$$+ \binom{M-1}{1} \sum_{n_2=N+1}^{N+N_A} \sum_{n_3=0}^{N} \cdots \sum_{n_M=0}^{N} P(N+N_A, .)$$

$$+ \quad . \quad . \quad .$$

$$+ \binom{M-1}{R-1} \sum_{n_2=N+1}^{N+N_A} \cdots \sum_{n_R=N+1}^{N+N_A} \sum_{n_{R+1}=0}^{N} \cdots \sum_{n_M=0}^{N} P(N+N_A, .)$$

$$+ \binom{M-1}{R} \sum_{n_2=N+1}^{N+N_A} \cdots \sum_{n_{R+2}=N+1}^{N+N_A} \sum_{n_{R+3}=0}^{N} \cdots \sum_{n_M=0}^{N} P(N, n_2, \ldots, n_M) \qquad (B.8)$$

This can be written in a condensed form:

$$P_B = \sum_{k=0}^{R} \binom{M-1}{k} \left\{ \sum_{n_2=N+1}^{N+N_A} \cdots \sum_{n_{k+1}=N+1}^{N+N_A} \sum_{n_{k+2}=0}^{N} \cdots \sum_{n_M=0}^{N} P(N+k^+N_A, n_2, \ldots, n_M) \right\} \qquad (B.9)$$

where

$$k^+ = \begin{cases} 1 & \text{if } k \neq R \\ 0 & \text{if } k = R \end{cases} \qquad (B.10)$$

and

$$P(N+k^+N_A, n_2, \ldots, n_M) = P(\underline{o}) \frac{\rho_1^{N+k^+N_A}}{(N+k^+N_A)!} \cdot \frac{\rho_1^{n_2}}{n_2!} \cdots \frac{\rho_1^{n_M}}{n_M!} \qquad (B.11)$$

Using Eq.(B.7) for $P(\underline{o})$, we get the final expression for P_B:

$$P_B = \frac{\sum_{k=0}^{R} C(k)\binom{M-1}{k} p^k q^{M-1-k}}{\sum_{m=0}^{R} \binom{M}{m} p^m q^{M-m}} \tag{B.12}$$

where

$$C(k) = \begin{cases} \rho_1^{N+N_A}/(N+N_A)! & \text{if } k \neq R \\ \rho_1^{N}/N! & \text{if } k = R \end{cases} \tag{B.13}$$

and

$$P = \sum_{m=N+1}^{N+N_A} \frac{\rho_1^m}{m!} \tag{B.14}$$

$$q = \sum_{m=0}^{N} \frac{\rho_1^m}{m!} \tag{B.15}$$

The average number of slots occupied by customers from station 1 is calculated as follows:

$$E(n_1) = \sum_{n_1=0}^{N} n_1 \left\{ \sum_{k=0}^{R} \binom{M-1}{k} p^k q^{M-1-k} P(\underline{o}) \frac{\rho_1^{n_1}}{n_1!} \right\}$$

$$+ \sum_{n_1=N+1}^{N+N_A} n_1 \left\{ \sum_{k=0}^{R-1} \binom{M-1}{k} p^k q^{M-1-k} P(\underline{o}) \frac{\rho_1^{n_1}}{n_1!} \right\} \tag{B.16}$$

$$\frac{E(n_1)}{P(\underline{o})} = \sum_{k=0}^{R} \binom{M-1}{k} p^k q^{M-1-k} \left(\sum_{n_1=0}^{N} n_1 \frac{\rho_1^{n_1}}{n_1!} \right)$$

$$+ \sum_{k=0}^{R-1} \binom{M-1}{k} p^k q^{M-1-k} \left(\sum_{n_1=0}^{N+N_A} \rho_1^{n_1}/n_1! - \sum_{n_1=0}^{N} \rho_1^{n_1}/n_1! \right) \tag{B.17}$$

$$\frac{E(n_1)}{P(\underline{o})} = \sum_{k=0}^{R-1} \binom{M-1}{k} p^k q^{M-1-k} \left(\rho_1 \sum_{m=0}^{N+N_A} \frac{\rho_1^m}{m!} \right) + \binom{M-1}{R} p^R q^{M-1-R} \left(\rho_1 \sum_{m=0}^{N-1} \rho^m/m! \right) \quad \text{(B.18)}$$

Therefore, the effective boundary is given, on average, by:

$$E(\gamma) = \frac{M\, E(n_1)}{F} \quad \text{(B.19)}$$

Similarily, the second moment of the effective boundary (or the number of busy CS slots) is, after some calculation given by

$$E(\gamma^2) = \frac{P(\underline{o})}{F^2} \sum_{k=0}^{R-1} \binom{M-1}{k} p^k q^{M-1-k} \left[\rho_1^2 A(N+N_A-2) + \rho_1 A(N+N_A-1) \right]$$

$$+ \binom{M-1}{R} p^R q^{M-1-R} \left[\rho_1^2 A(N-2) + \rho_1 A(N-1) \right] \quad \text{(B.20)}$$

where $$A(m) = \sum_{i=0}^{m} \rho_1^i/i! \quad \text{(B.21)}$$

The parameter $E(\gamma^2)$ is needed to evaluate the packet time delay using the model of a single server with "breakdowns" (see Eq.(2.14)).

4. OPTIMAL CIRCUIT SCHEDULING IN SSMB SYSTEMS: PERFORMANCE BOUNDS AND COMPUTATIONAL COMPLEXITY

Arbitrary traffic matrices

Interference (overlapping beam) constraints

Multi-channel beams

Different uplink/down configurations

As indicated in earlier chapters, when system structure and constraints become more complex, the problems of design and performance analysis increase commensurately. Thus, we have generally limited our previous discussion of system designs to suboptimal ones when dealing with packet switched systems having such complexities as skewed traffic matrices and overlapping beams. Furthermore, our performance evaluation in these cases has been essentially by simulation rather than analysis, since in a packet-switching or integrated networks environment, any analytical models would be completely intractable. However, if we limit our attention to systems in which capacity is assigned on a slowly varying basis, as would be the case for dedicated or switched circuit assignments, it is possible to expand considerably the range of systems which are amenable to optimal design and analysis. In this chapter we shall examine the SS/TDMA scheduling problem in this context. The thrust of our work on this problem has been:

1) To determine the computational complexity of the scheduling problems with a given set of constraints.

2) To develop optional scheduling algorithms in cases where that is feasible, or suboptimal heuristics in cases where optimization is impractical.

We consider systems with arbitrary (i.e. skewed) traffic matrices, with possible overlap (i.e., interference) between beams and with one or more of the following additional types of complexity:

- different uplink/downlink beam configurations
- different power allocations in different beams
- frequency reuse within beams by polarization
- channelization within beams by FDMA, CDMA or other orthogonal multiplexing schemes
- onboard demod/remod
- fewer transponders than beams

In the following sections, we give a summary of our results and those of others on these problems. The discussion emphasizes general computational features of the scheduling problems. The specifies of the algorithms can be found in the cited references.

4.1. OPTIMAL SWITCHING FOR SYSTEMS WITH SINGLE CHANNEL BEAMS

We consider a system operatng with a fixed TDMA frame structure in which slots are to be assigned on a permanent (or at least relatively long term) basis to traffic defined by a given traffic demand matrix D. The assignment of a single slot is assumed to represent the allocation of "one unit" of capacity. The "units" might be in the form of digitized voice channels (i.e. 64 kbps channels) or in any other convenient size. The entry d_{ij} (an integer) in the D matrix indicates the number of units of capacity required for transmissions from zone i to zone j. This capacity might be allocated on a permanent basis (if we are dealing with leased channels) or it might represent a pool of resources set aside for demand assignment in a circuit-switched manner. In either case the slot allocation problem is the same. We consider first, a system which has N uplinks, switched (through an r-f switch in the satellite) to N downlinks, so that the demand matrix is Nx

N. (We treat below the case where there may be a different number of uplinks and downlinks.) The optimal switching problem consists of accommodating the specified traffic demand within a TDMA frame of minimum duration subject to the system constraints. These constraints include the possible conflicts within the switch, limitations on the number of active transponders, interference due to overlapping beams, and possibly others. A minimum duration frame corresponds to maximum transponder utilization. Thus, an optimized assignment corresponds to maximizing the throughput per frame. (In practice, the time duration of the frame might be fixed, in which case solving the optimization problem as stated here corresponds to packing a given demand as "tightly" as possible into the frame, thereby allowing as much as possible for additional demand, or more generally, freeing as much transponder capacity as possible for other use.)

Assuming that there are N transponders available and that there are no other constraints except those of the switch, it can be shown that a frame of length

$$L \geq \text{Max} \left[\max_i R_i, \max_j C_j \right] \qquad (4.1)$$

where

$$R_i = \sum_j d_{ij}$$

$$C_j = \sum_i d_{ij}$$

is necessary to accommodate all traffic. For example, for the 4 x 4 matrix of Fig. 4.1 the minimum possible frame size is L = 4. It can also be shown that this minimum length is achievable, and an algorithm has been developed for determining an optimal schedule. As an illustration, the

matrix of Fig. 4.1 can be scheduled as shown in Fig. 4.2. In the figure, each row might be considered to represent the transmissions of a given transponder associated with that uplink. Note that certain transponders are inactive during part of the frame (shown shaded in the figure), indicating that 100% transponder utilization is not (and cannot be) attained in this example. (Of course, the assignment is nevertheless optimal.) For the details of the optimal scheduling algorithm the reader is referred to [INU 79].

It is interesting to note that the assignment problem just described is related to a well-known problem in graph theory, known as the edge coloring problem in bipartite graphs. As will be seen below, many of the other problems we consider, involving more realistic constraints also have their counterparts in graph theory, suggesting that there is a large body of literature to draw on in developing algorithms for scheduling and capacity allocation.

The above problem can be generalized to the case where there are less active transponders than zones and the demand matrix is not square. A non-square matrix corresponds to a system with a different number of uplink and downlink zones. Such a configuration might result from a system design adapted to highly non-symmetric traffic distributions, in which antenna gains and power allocations are adjusted to define geographic coverages differently for uplink and downlink transmissions. Or it might be a design especially adapted to certain special broadcast (multi-destination) requirements. A typical example is shown in Fig. 4.3. In this context, the optimal scheduling problem is stated as follows. Given a (generally rectangular) demand matrix D, find a schedule which assigns all traffic in a frame of minimum duration using K transponders. (The limitation to K transponders corresponds to a constraint that no more than K simultaneous transmission can occur during any slot.)

It can be shown that the minimum frame length necessary to schedule all traffic in this case is given by

$$L = \text{Max}[\max_i R_i, \max_j C_j, \lceil \frac{T}{K} \rceil] \qquad (4.2)$$

where $T = \sum_{i,j} d_{ij}$

and $\lceil x \rceil$ = least integer $\geq x$

An algorithm for performing the scheduling with this value of L has been developed. As an example, with K = 2 transponders, the demand matrix of Fig. 4.4 can be scheduled within a frame of length L = 12, as shown in the figure. (This is the minimum possible frame length.) The details of the optimal scheduling algorithm in this case may be found in [BON80].

Sometimes it is advantageous to attempt to reduce to a minimum the number of times the switch is reconfigured within a frame. This might be the case for example, when switching requires a significant amount of time or power. Previous work [ITO77] has treated the scheduling problem with a more or less "informal" attempt to reduce the number of switchings Ns. However, we have found that a true optimal scheduling approach in which a constraint on Ns is included in the formulation, generally turns out to be an exceedingly difficult combinatorial problem. We comment briefly here on the computational complexity of this problem as determined from our recent

studies, restricting ourselves to the case of an N x N matrix D with K transponders. It is easily shown that the minimum number of switchings N_{min} required for any D is $\leq N$. Thus one approach would be to determine N_{min} and then attempt to determine an optimal schedule subject to the constraint, $N_s = N_{min}$. Some partial results have appeared in the literature on this problem. [CAL80] We have determined that the problem is NP-complete, i.e., its computational complexity in all likelihood grows exponentially with N. On the other hand, the techniques mentioned above, for optimal scheduling without constraints on N_s, are guaranteed to produce $N_s \leq N^2$. These techniques have a complexity of the order of N^5. For constraints on Ns between these two extremes, (N and N^2) it is uncertain at this point what the complexity of the problem is. This problem is still under study, and various heuristics have been developed for obtaining "good" schedules subject to switching constraints.

4.2 MULTI-CHANNEL AND DUAL POLARIZED BEAMS

So far it has been assumed that all uplink and downlink beams are of equal capacity, each permitting one unit of traffic to be transmitted in each time slot. In reality, each beam is likely to be broken down into subchannels in various ways. For example, independent transmissions could be superimposed in the same frequency band using two orthogonal polarizations. Or any one of a number of standard orthogonal multiplexing techniques such as FDMA, CDMA or other analog and digital methods could be used to to create several independent channels within the same beam, the number of channels being a function of bandwidth, antenna gain and power allocated to the beam. Such channelization might be useful for cases in which heterogeneous

populations of small and large stations exist within a single beam, and/or the traffic matrix is highly skewed and asymmetric. In the latter case, beams with a large number of channels could be configured to cover zones of high traffic demand, while zones of lower demand would be served by beams with a lower number of channels. A general model applicable to a wide range of such systems is shown in Fig. 4.5. As shown in the figure, the i-th uplink is assumed to be channelized into β_i subchannels each of one unit capacity, and the j-th downlink channelized into α_j subchannels of one unit each. The data streams arriving at the satellite are demultiplexed, so that a total of $\Sigma\beta_i$ subchannels are available at the input of the first switch, to be connected to the set of $\Sigma\alpha_j$ subchannels on the downlinks. The configuration shown in the figure (just one of many possible ones) uses a single transponder for each subchannel. Thus the maximum number of subchannels that can be active in any one time slot will be K, the number of available transponders. More generally, K represents the maximum number of simultaneous transmissions of which the satellite is capable. This might be a function of switch, power or other limitations, in addition to limitations on the number of transponders. We consider the general case of rectangular systems (M uplinks and N downlinks). Note that the demultiplexing and multiplexing might consist only of filtering and frequency translation (as in an FDMA system), or it might involve onboard demodulation and remodulation in more complex systems.

The problem of optimal scheduling in this system is similar to that treated above involving systems with a single channel per beam. (Illustrated

in Fig. 4.4). This time, however, the switch constraints are more complex. Now in a given time slot, a maximum of β_i units of traffic can be transmitted on uplink i, a maximum of α_j units on downlink j and the total transmissions in any slot cannot exceed K. As before, we start with a (rectangular) demand matrix D with non-negative integer entries, and wish to find a schedule for this traffic in a minimum frame duration L, respecting the switch constraints mentioned above. It has been shown that a lower bound on frame duration for this case is given by

$$L = \text{Max}\left[\left\lceil\frac{T}{K}\right\rceil\ ,\quad \max_i\left\lceil\frac{R_i}{\beta_i}\right\rceil\ ,\quad \max_j\left\lceil\frac{C_j}{\alpha_j}\right\rceil\right] \qquad (4.3)$$

Furthermore an algorithm has been developed for constructing a schedule which achieves this bound. The reader is referred to [GOP 81] for details of the algorithm. We will confine ourselves here to a simple illustrative example. Fig. 4.6 shows a demand matrix for a three beam system having various channelizations of its up and down links as indicated by the β's and α's in the figure, and having $K = 4$ transponders. For this example, the lower bound of (4.3) yields $L \geq 5$. By applying the algorithm a schedule can be constructed which achieves this bound, as shown in the figure. The traffic matrix D is broken down into a "feasible decomposition", which is a sum of matrices D_i, each of which represents the traffic transmitted in one time slot. The matrix D_1 in figure shows that in slot 1, a total of four units of traffic are transmitted (consistent with the constraint $K = 4$). It is distributed as follows: One subchannel in beam 1 is transmitting back down to beam 1, and the other is connected to a subchannel in downlink 2. Simultaneously, the two subchannels in uplink 2

are connected back down to two subchannels in beam 2. This completely exhausts all subchannels in the up and downlinks on beams 1 and 2, and all transponders, so that uplink 3 does not transmit any traffic in this slot. The remainder of the schedule is defined in a similar fashion by matrices D_2, D_3, D_4 and D_5.

4.3 ADJACENT BEAM INTERFERENCE

In this section we confine attention to N x N systems. In general, if a large contiguous area is to be covered by a number of beams in a multibeam system, there will always be some interference in the regions of zones affected by adjacent beam overlap, as illustrated by the shaded areas in Fig. 4.7. More generally, depending upon how well the antenna side lobes are suppressed, this interference may even spill over into non-adjacent zones. It is convenient to represent these interference effects by an "interference graph", the edges of which signify mutual interference between the vertices (zones) at their extremities. In a graph-theoretic sense such vertices are "adjacent", even though the corresponding zones may not be geographically adjacent. In the following, we will define "adjacent zone" interference in this more general sense. Fig. 4.8 shows an interference graph superimposed on the geographical interference picture shown in Fig. 4.7. Note that in the case where interference only occurs between geographically adjacent zones, the interference graph is planar. As will be pointed out below, planar interference patterns are more easily handled than non-planar ones.

It is clearly not feasible to simultaneously transmit to, or receive from two zones which mutually interfere. There are a number of ways of "resolving" such interference. If transmissions to/from all zones are

assumed to be reusing a common uplink and common downlink channel (as would be the case, for example, in an N x N system reusing the same frequency band N times, without exploiting orthogonal polarizations or other methods of channelization within a beam) then interference can only be resolved by scheduling the TDMA frame so that at most one of a mutually interfering set of beams is transmitting (receiving) at a time. In the example of Fig. 4.8, zones 2 and 3 can be scheduled simultaneously, but if 1 or 4 is active, all others must be inactive. Clearly, a schedule for a given demand matrix in a system which requires resolution of adjacent zone interference will yield a longer frame length than a schedule for a non-interfering system.

A second way of resolving interference is to make use of several orthogonal channels of the type proposed in Sec. 4.2. For example, if two orthogonal polarizations are available, pairs of interfering zones could be simultaneously active provided that they were using different polarizations. Similarly, if two or more separate frequency bands (or other forms of orthogonal channels) were available, they could be used for simultaneous transmissions to or from a mutually interfering set of beams. In general, then, we resolve interference by requiring that transmissions to or from a set of interfering beams be orthogonal. If orthogonality in time is used, as described above, then the time frame generally must be expanded to resolve interference. If orthogonality in frequency is used we are thereby "cutting up" the available frequency spectrum and limiting its reuse. In either case, resolution of interference will necessarily reduce the efficiency of utilization of the available communication resources. Since efficient resource allocation is a prime consideration in SS/TDMA systems, we have studied possible algorithms for optimal scheduling in the face of adjacent zone interference.

Consider, first, the problem of resolving interference by the use of two polarizations. The overall problem is: given a traffic matrix D, schedule this traffic in minimum time subject to both the switch/transponder constraints and the interference constraints.

One way to approach this problem is via a two-step approach:

1) Taking into account the traffic matrix D, assign polarizations to the various transmissions which reduce the interference conflicts to a minimum.

2) If all interference has not be resolved in step (1), expand the resulting schedule to a larger time frame by adding time slots in order to separate in time, the remaining interfering traffic.

In the case where the interference graph is planar we have formulated an algorithm for step (1) which requires polynomial time; i.e., it requires a computational effort which grows as a power of N. The algorithm is based on a procedure of Hadlock for solving a problem in graph theory known as the Max-cut problem.

In the case of non-planar interference graphs, and/or in cases where more than two orthogonal channels are available for resolving interference, we have shown that the scheduling problem is NP-complete. Since it would be impractical to implement an optimal scheduling algorithm in this more general case, we have developed several heuristics for constructing suboptimal schedules. By establishing a lower bound on the frame length for the optimal schedule with interference constraints, we have been able to estimate the efficiencies of these heuristics. Our simulation results averaged over a fairly large number of tests show that the frame lengths based on heuristic approaches average only about 15% more than the (not necessarily achievable) lower bound. This work is still in progress.

REFERENCES

[CAL80] Calo, S.B., and K.S. Natarajan, "Reference Time Slot Assignment in an SS/TDMA System with Minimum Switching," IBM Res. Report #RC 8342, 1980.

[GOP 81] Gopal, I.S., Bongiovanni, G., Bonuuccelli, M.A., Tang, D.T. and Wong, C.K., "An Optimal Switching Algorithm for Multibeam Satellite Systems with Variable Bandwidth Beams," IBM Res. Report RC 9006, 1981.

[INU 79] Inukai, T., "An Efficient SS/TDMA Time Slot Assignment Algorithm," IEEE Trans. Comm. vol. COM-27, October 1979.

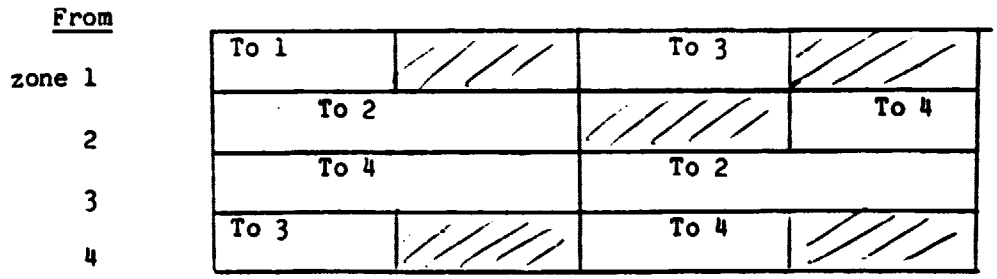
[BON 80] Bongiovanni, G., Coppersmith, D., and Wong, C.K., "An Optimal Time Slot Assignment Algorithm for An SS/TDMA System with Variable Number of Transponders," IBM Research Report RC8301, April 1980.

[ITO 77] Ito, Y., Urano, Y., and Yamaguchi, M., "Analysis of a Switch Matrix for an SS/TDMA System," Proc. IEEE, vol. 65, no.3, March 1977.

From uplink zone \ To downlink zone	1	2	3	4
1	1	0	1	0
2	0	2	0	1
3	0	2	0	2
4	0	0	1	1

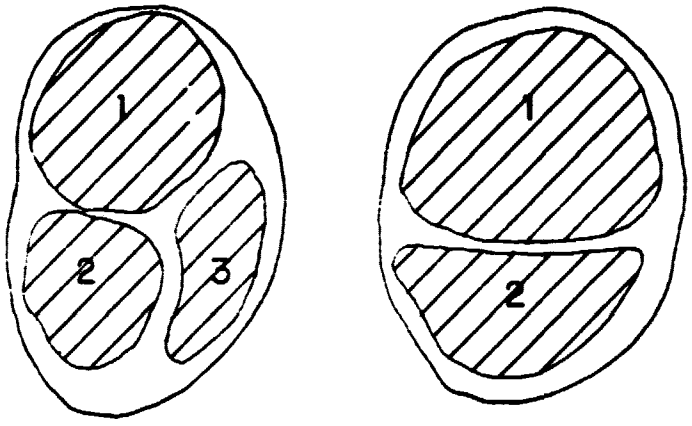
Demand Matrix

Fig. 4.1



Optimal schedule of duration 4

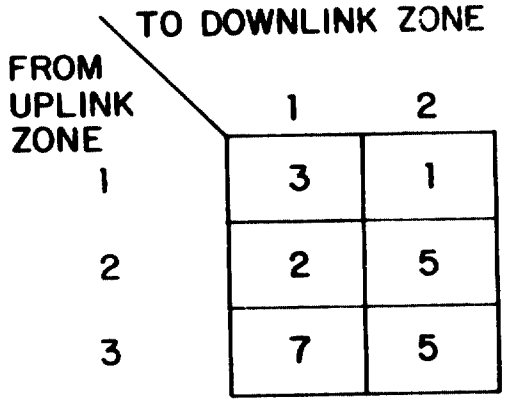
Fig. 4.2



UPLINK BEAM FOOTPRINTS (3 ZONES)

DOWNLINK BEAM FOOTPRINTS (2 ZONES)

Fig.4.3 Different uplink/downlink configurations



TRAFFIC MATRIX D

LOWER BOUND = 12

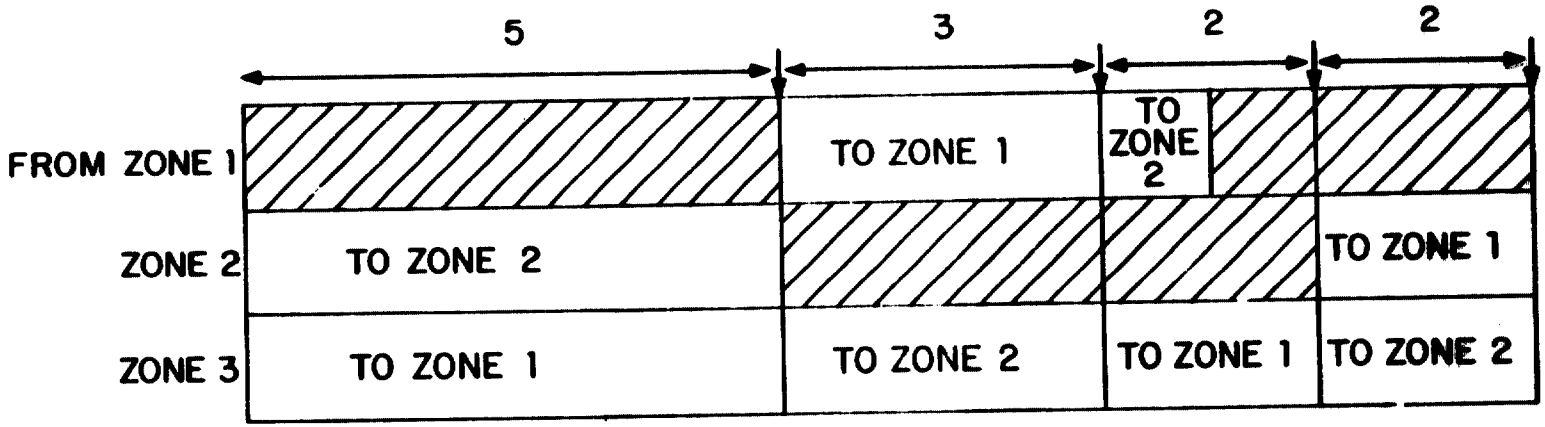


Fig. 4.4 Rectangular Traffic Matrix Scheduling

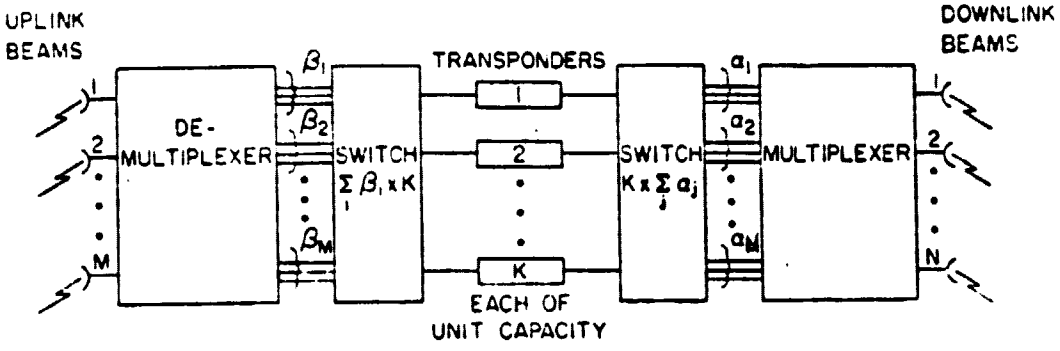


Fig. 4.5 System configuration

D

	α_1	α_2	α_3
β_1	2	6	0
β_2	1	6	1
β_3	1	0	3

$\beta_1 = 2$ $\beta_2 = 2$ $\beta_3 = 1$

$\alpha_1 = 1$ $\alpha_2 = 3$ $\alpha_3 = 1$

$M = 3$ $N = 3$ $K = 4$

LOWER BOUND, $L = 5$

FEASIBLE DECOMPOSITION:

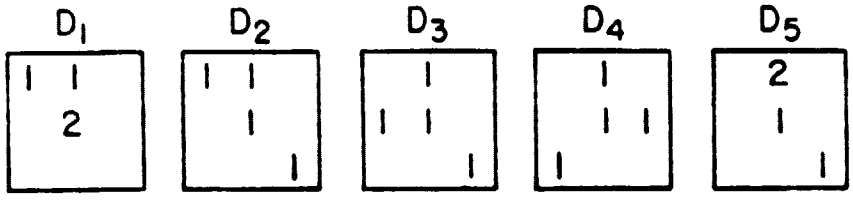


Fig. 4.6 Example of feasible decomposition

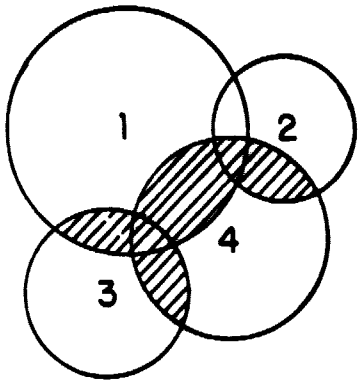


Fig. 4.7 Geographical interference pattern.

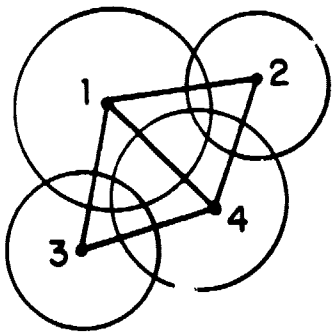


Fig. 4.8 Interference graph.

Research Participants

Faculty

T.E. Stern

M. Schwartz

H.E. Meadows

Graduate Research Assistants

S. Ganguly

B. Kraimeche

K. Matsuo

I. Gopal

Supporting Staff

S. Abdoo

B. Lim

J. McMillan