

Simulation of Hybrid Terrestrial-Satellite Networks for Service Restoral and Performance Efficiency

Martin Nesenbergs



U.S. DEPARTMENT OF COMMERCE
Robert A. Mosbacher, Secretary

Janice Obuchowski, Assistant Secretary
for Communications and Information

November 1991

PREFACE

The Institute for Telecommunication Sciences is conducting a series of projects concerned with the use of advanced satellite system technology to enhance rapid restoration of services provided by the Public Switched Networks (PSN) following a natural or manmade disaster, as well as to assess the potential beneficial roles of advanced communication satellites in the growing Integrated Services Digital Networks (ISDN). It is tacitly assumed that Public Switched Telephone Networks (PSTN) are dominant and therefore merit most attention among the public networks. Overall goals of this work are (1) to promote an effective integration of advanced satellite systems with future broadband terrestrial networks, (2) to perform studies that examine uses of advanced communication satellite systems to reduce national vulnerability to telecommunication outages, (3) to identify and recommend interface and functional standards required for integrated services, such as ISDN, in a terrestrial-satellite broadband transmission and switching environment, and (4) to use quantitative end-user service quality as the basic measure of network performance in all phases of the study. The Institute is working with the National Aeronautics and Space Administration, other Government organizations, universities and industry to perform the necessary research and development.

The purpose of the task addressed by this report has been to outline, define, characterize, and plan for a simulation program that will analyze the performance of terrestrial and satellite networks. The report addresses the modeling and simulation of circuit-switched networks. The actual network simulation tasks will constitute the next phases of the program. Because of the differences in simulation tools, the simulation is planned to proceed in steps from circuit-switched networks, to packet switched networks, and finally to ISDN.

Certain commercial systems, equipment, and software products are identified in this report to describe adequately the designs and conduct of research or experiments. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.

The views, opinions, and/or findings contained in this report are those of the author. They do not represent an official position of the National Telecommunications and Information Administration, the U.S. Department of Commerce, or of the National Aeronautics and Space Administration.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	vi
LIST OF TABLES	viii
LIST OF ACRONYMS	ix
ABSTRACT	1
1. INTRODUCTION	1
1.1 Vulnerabilities and Threats.....	2
1.2 Potential Solution: Hybrid Satellite-Terrestrial Network.....	4
1.3 Need for Modeling and Simulation.....	6
1.4 Outline of this Report.....	7
2. OVERALL PLAN AND OBJECTIVES	8
2.1 Terrestrial and Satellite Hybrid Networks.....	8
2.2 Survivability and Restoral Objectives.....	21
2.3 Traffic Classification.....	23
3. MODELING AND SIMULATION METHODOLOGIES	25
3.1 Discrete Event Methods.....	27
3.2 Aggregated Methods.....	31
3.3 Statistical Design and Processing.....	34
4. CIRCUIT SWITCHED NETWORKS	39
4.1 Planning Factors for Modeling and Simulation.....	39
4.2 Available Simulation Tools.....	41
4.3 Selection of Preferred Method.....	44
4.4 Temporal Aggregation: An Example.....	45
4.5 Simulation Scope and Performance Targets.....	75
5. CONCLUSIONS	98
6. REFERENCES	99

LIST OF FIGURES

	Page
Figure 1.	Illustration of a small, twelve-node, terrestrial network 9
Figure 2.	Proposed AT&T network architecture of the future 11
Figure 3.	The classical seven-layer model for Open Systems Interconnection (OSI) 12
Figure 4.	Physical view of the satellite support facilities 14
Figure 5.	Functional architecture of the satellite-terrestrial hybrid 17
Figure 6.	Distinction between physical links (a) and their logical connectivity (b) 20
Figure 7.	The roles of modeling and simulation in network performance evaluation 26
Figure 8.	Function $(1/x) \exp \Psi(x)$ 38
Figure 9.	Maximum likelihood gamma density fit to a given histogram 40
Figure 10.	Functional outline for time-aggregated, blocking GOS simulation of a circuit switched network 48
Figure 11.	Input: network specification 50
Figure 12.	Input: traffic specification 52
Figure 13.	Input: stress specification 54
Figure 14.	Initialization of simulation runs 55
Figure 15.	Add-calls routine (ACR) for a connected network 60
Figure 16.	ACR modification for a separated network 65
Figure 17.	Delete-calls routine (DCR) 68
Figure 18.	Output routines 71
Figure 19.	The parameter cube for network simulation 77

LIST OF FIGURES (continued)

	Page
Figure 20. Illustrative network model of existing facilities	78
Figure 21. A symmetric test network with $N=12$, $L=12$, and $X=4$	79
Figure 22. A symmetric test network with $N=12$, $L=18$, and $X=3$	81
Figure 23. A symmetric test network with $N=12$, $L=30$, and $X=2$	82
Figure 24. Preferred alternative routing scheme of today	84
Figure 25. The principle of DACS capacity switching	85
Figure 26. Different deployment of cross-connect devices can yield the same logical connectivity	87
Figure 27. Specification of the simulation model: Part 1.1, the terrestrial network	88
Figure 28. Specification of the simulation model: Part 1.2, the satellite network	90
Figure 29. Specification of the simulation model: Part 1.3, the network control system	91
Figure 30. Specification of the simulation model: Part 2, traffic	93
Figure 31. Specification of the simulation model: Part 3, stress	95
Figure 32. Network dynamics: (a) stress, (b) GOS, and (c) facility utilization	97

LIST OF TABLES

	Page
Table 1.	Description of OSI Layer Functions 13
Table 2.	Selection of Services and Switching Categories 19
Table 3.	Features on Which to Evaluate a Simulation Language 28
Table 4.	Examples of Standards SDL Symbols and Their Definitions 29
Table 5.	Specification Parameters for a Circuit-Switched Network Model 42
Table 6.	Example of Traffic Estimation on a Small, Five- Node Network 57
Table 7.	Example of Node-to-Node Calls on the Five-Node Network 59
Table 8.	A Simple Routine for Partitioning the Set of $x(n)$'s into $x(j,k)$'s 62
Table 9.	The Wichmann-Hill Algorithm 67
Table 10.	List Arrays for Time-Sampled Simulation of Circuit-Switched Networks 72
Table 11.	Matrix Arrays for Time-Sampled Simulation of Circuit-Switched Networks 74

LIST OF ACRONYMS

ACH	Automated Clearing House
ACR	Add-Calls Routine
ACTS	Advanced Communications Technology Satellite
ASDS	Accunet Spectrum of Digital Services
ATM	Automated Teller Machine, or Asynchronous Transfer Mode
AT&T	American Telephone and Telegraph Company
BER	Binary (or Bit) Error Rate (or Ratio)
BH	Busy Hour
BISDN	Broadband ISDN
CAAD	Computer-Aided Analysis and Design
CCC	Clear Channel Capability
CCITT	International Telegraph and Telephone Consultative Committee
CCS	Common Channel Signaling
CNS	Commercial Network Survivability
CPE	Customer Premises Equipment
CSI	Commercial SATCOM Interconnectivity
CSS	Commercial Satellite Survivability
DA	Demand Assignment
DACS	Digital Automatic Cross-Connect System
DCR	Delete-Calls Routine
DoD	Department of Defense
ESS	Electronic Switching System (such as ESS4)
EVT	Extreme Value Theory
FCC	Federal Communications Commission
F-T1	Fractional T1
GOS	Grade of Service
GTE	General Telephone and Electronics Company
INTELSAT	International Telecommunications Satellite Organization
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISO	International Organization for Standardization
ITS	Institute for Telecommunication Sciences
IVSN	Initial Voice-Switched Network
LAN	Local Area Network
LATA	Local Access Transport Area
MAN	Metropolitan Area Network
MCI	Microwave Communications, Incorporated
NCS	National Communications System
NETS	National Emergency Telecommunications Service
NLP	National Level NS/EP Telecommunications Program
NNI	Network Node Interface
NRC	National Research Council
NS/EP	National Security Emergency Preparedness
NSTAC	National Security Telecommunications Advisory Committee
NTIA	National Telecommunications and Information Administration
ORIN	Originating Route Identification Number
OSI	Open Systems Interconnection

LIST OF ACRONYMS (continued)

PSN	Public Switched Network
PSTN	Public Switched Telephone Network
QOS	Quality of Service
QTCM	Queuing Traffic Congestion Model
RBOC	Regional Bell Operating Company
SATCOM	Satellite Communications
SDH	Synchronous Digital Hierarchy
SDL	System Description Language
SDN	Software Defined Network
SHAPE	Supreme Headquarters Allied Powers in Europe
SS	Signaling System (such as SS6 or SS7)
STP	Signal Transfer Point
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
VLAN	Very Large Area Network
VSAT	Very Small Aperture Terminal
WAN	Wide Area Network

SIMULATION OF HYBRID TERRESTRIAL-SATELLITE NETWORKS FOR SERVICE RESTORAL AND PERFORMANCE EFFICIENCY

Martin Nesenbergs*

Motivated by recognized vulnerabilities of the terrestrial public networks, this report addresses the question whether an appropriate introduction of advanced satellite systems would or would not benefit the telecommunication services for the currently existing terrestrial infrastructure. What the satellite subnetwork should be, and what performance gains are to be realized, are two key issues. Given voice, data or integrated services traffic, the survivability and restoral effectiveness of different network configurations is likely to vary considerably for different crisis scenarios. It is concluded that answers to these and other complex, performance-related, questions can only be gotten by means of computer modeling and simulation. Today there seem to be sufficient simulation tools available for the task. The report reviews the overall plan and simulation objectives for circuit-switched networks. From the many proposed methodologies, the discrete event and temporal aggregation methods are emphasized. The importance of simulator inputs is demonstrated through needs for relatively detailed specifications of the terrestrial and satellite networks, their interfaces, the offered traffic, stress (i.e., network damage and traffic overload), and service performance measures required in the simulation output.

Key words: advanced satellites; blocking grade-of-service; circuit-switched networks; modeling; network damage; performance; satellite-terrestrial hybrid; service restoral; simulation; stress; temporal aggregation; traffic

1. INTRODUCTION

The recent growth of information and telecommunication industries has been beneficial to U.S. economies and the society at large. At the same time, however, one must admit that many public and private enterprises are becoming increasingly, if not excessively, dependent on the services of various communications networks. Sudden unavailability of the accustomed instant connectivity or capacity to distant sites can be detrimental to

* The author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80303-3328.

both government and business. As time goes by and new service demands and facilities emerge, the future risks can be expected to increase at a rate faster than the anticipated growth rate of service benefits - unless, of course, something is done to remedy the threats.

1.1 Vulnerabilities and Threats

Consider the banking industry. Telecommunications networks serve over a billion Automated Clearing House (ACH) and approaching 5 billion Automated Teller Machine (ATM) transactions per year. Resultant bank office savings in staff, paperwork, and customer time may well be equivalent to billions of dollars per year. Associated industrial productivity and business efficiency gains are harder to estimate, but may very well be one order of magnitude larger. The securities industry is estimated to spend around \$2 billion (or 5% of gross revenue) each year on telecommunications, as over 90% of all security transactions involve one or more networks.

Other industries, such as manufacturing, trade, merchandising, insurance, publishing, transportation, travel, recreation, and so on, also use communications networks to a considerable, albeit varied, degree. The Federal Government, through its Departments of Defense, Treasury, Transportation, Health and Human Services, etc., constitutes perhaps the world's largest user of owned, leased, and measured-use communications networks. The Federal, state, and other governments' expenditures and cost savings are not readily available, but are known to be huge.

In view of this unprecedented dependence on telecommunications, it is only reasonable and prudent to ask about the reliability, survivability, and restorability of various network services. The question has been raised in several studies. The growing vulnerabilities of the Public Switched Telephone Networks (PSTN) and the more general Public Switched Networks (PSN) have been assessed by the National Research Council (NRC, 1986 and 1989). In summary, the NRC studies identify the following nine areas of concern:

1. Proliferation of private networks would be of more benefit if the assorted networks could interoperate with each other.
2. New and improved inter-network gateway architectures are needed for network interoperability.
3. Critical network facilities (i.e., nodes and trunks) should be more secure against a variety of threats.
4. As a general policy, concentration of network traffic should be limited to avoid widespread consequences of a single facility outage.

5. Despite inherent costs and inefficiencies, the number and diversity of alternate routes should be increased.
6. Critical users require priority service.
7. There is a need for advance rehearsal or simulation of disaster and recovery scenarios to have reliable strategies ready for all possible stress events.
8. Networks must better protect their operational and management software from hostile penetration and manipulation.
9. Signaling systems, including all their hardware and software, should be more robust.

These problem areas are expected to expand in the future. An example could be excessive concentration of network databases in single central locations. In addition to specific facility failures or outages, service can be degraded by intolerable surges or mixtures of traffic, as well as inappropriate resource sharing. The resultant Grade or Quality of Service (GOS or QOS) can be reduced by congestion in the existing networks, and perhaps even more so in future integrated services networks. Combinations of degradations may reach a level that is equivalent to a service outage for end-users (Yokoi and Kodaira, 1989).

Threats to service continuity come in various forms. The most familiar ones are physical damage events to network facilities. Recent publicized accounts list such man-made or natural events (Zorpette, 1989; Braun, 1989; Linfield, 1990; Bell and Zorpette, 1991) as:

- * Fire in downtown business district of Minneapolis, MN (November, 1982).
- * ESS4 outage in Dallas, TX, accompanied by a failure of its hot standby (February, 1987).
- * Central office fire in Hinsdale, IL (May, 1988).
- * New Jersey cut of optical-fiber cables that carried 200,000 calls per hour (November, 1988).
- * Hurricane Hugo in North and South Carolinas (August, 1989).
- * Earthquake in San Francisco, CA, area (October, 1989).
- * One of the 10 busiest optical-fiber links of AT&T's long-haul network, between New York City and Newark, NJ, disabled by AT&T's own workmen (January, 1991).

Other incidents, caused by construction crews (e.g., backhoe operators), flooding, tornados, vandalism, etc., may be individually less serious, but they tend to occur more often as part of normal peacetime affairs. In times of war many threats of different destruction levels have been postulated (NCS, 1988a and 1988b). Caused by military activities and/or terrorism, these threats can disable switching nodes and trunk lines.

A network's ability to provide adequate service can be impaired by flaws or intrusions related to control software. There are two vivid and recent examples:

- * AT&T toll network slowdown (January, 1990), allegedly caused by incompatibilities between Signaling Systems 6 and 7 (i.e., SS6 and SS7).
- * Another software bug in Signaling System 7 was blamed for telephone outages that plagued 10 million customers in Los Angeles, Washington, DC, Pittsburgh, and nearby areas (July, 1991).

Both flaws were corrected in less than 24 hours, but one can easily conjecture other potential mishaps, faults, or less than benign penetrations by software "virus, worms, or bombs." Authorities, public officials, and significant users are being assured by the carriers that appropriate security measures are being taken to protect the network controls and operations.

Finally, service can also be degraded by either widespread or locally focused network congestion. Such congestion is caused by extreme traffic surges that overwhelm the available facilities and bar new incoming calls from receiving service. Typical examples with almost predictable stress levels occur every year on Mother's Day and Christmas. Other traffic surges associated with some sensational, regional, or national events are less predictable in intensity and duration.

1.2 Potential Solution: Hybrid Satellite-Terrestrial Network

Faced with all the above threats to the survivability of the largely terrestrial public networks and their services, one should be obliged to seek practical alternatives to reduce the risks for end-user services. Many measures can be taken. One can start with extensive physical hardening and 24-hour protection of existing facilities, or addition of redundant multihoming terrestrial lines and trunks (perhaps in the form of more dispersed optical fiber plant), or more control centers, or distributed data bases, or enhanced software security, and so forth.

The potential solution considered here is the introduction of satellite network connectivity. This additional new network is assumed to have a to-be-specified, sufficiently fast switching capability in orbit for individual point-to-point circuits and high bandwidth messages. Potential candidate networks could be designed using both geostationary or low-orbit satellites. A promising low-orbit, 77 satellite, global digital network called the Iridium

(Williams, 1991) has been proposed largely for personal services. Motivation based on Intelsat nonswitching satellite experiences is given by Kinzie (1989), whereas recent crisis restoration demonstrations have been reported by Boensch and Sogegian (1989) and SunGard (1989). Furthermore, for several years AT&T has been offering interactive, tariffed satellite data-network services under the title of SKYNET (Sanchez, 1988).

However, a more likely candidate may very well turn out to be some adaptation of the stationary, switch in the sky concept of the Advanced Communications Technology Satellite (ACTS), see Naderi and Campanella (1988), Palmer and White (1990), or Wright et al., (1990).

Because of the satellite switching capability, satellites like ACTS would be more advanced than the Japanese DYANET system (Morihiro et al., 1990; Ohnuki et al., 1990; Nakashima et al., 1990; Kato et al., 1990). The DYANET system, also called the "common alternative routing system," has been in orbit and operational since 1988. It combines the traffic-handling capabilities of both terrestrial and satellite Time Division Multiple Access (TDMA) circuits. The former carries stable traffic, while the latter carries the far more variable overflow and emergency traffic from throughout Japan.

Similar work on Time Division Multiple Access satellite systems with Demand Assignment (TDMA/DA) has been reported from Italy (Butto et al., 1989). Satellites also appear to be "fully prepared" for operation in the ISDN environment of the future (Patacchini and Galante, 1987; CCITT, 1989a). The switching satellite is envisioned to provide backup connectivity when and where such backup happens to be needed. Under normal unstressed conditions, the satellite circuits can be part of the regular service network. In this report, the resultant terrestrial plus satellite configuration is called the satellite-terrestrial hybrid network.

The use of satellite support networks has been considered by many in the United States. In 1983, the National Security Telecommunications Advisory Committee (NSTAC) issued recommendations pertinent to Commercial Satellite Survivability (CSS). Among other matters, NSTAC recommended that the following survivability issues be addressed:

- * Interoperability among different existing commercial satellite networks.
- * Emergency restoration (e.g., plans, procedures, and general coordination) of commercial satellite communication services in a crisis.

Specific, recent and comprehensive, Federal Government initiatives have been supported and coordinated by the National Communications System (NCS) under the heading of the National Level NS/EP Telecommunications Program (NLP). The NLP program of NCS is divided into three parts: the Nationwide Emergency Telecommunications Service (NETS), the Commercial Network

Survivability (CNS), and the Commercial SATCOM Interconnectivity (CSI).

While the NETS program has been largely focussed on switching and control nodes of the PSN, and CNS on the terrestrial transmission routes, CSI pertains to satellite networks (FCC, 1988).

CSI is mandated to seek commercial satellite networking support when connectivity crises occur on the terrestrial AT&T, MCI, or GTE-Sprint interexchange networks. The implementation of CSI is achieved through compatible commercial Earth stations placed sufficiently close to important switching centers. Since the CSI Earth stations may be owned and operated by different carriers, including some owned or operated by the Government, one of the main requirements for CSI is interoperability. The Earth stations in question must either cooperate with or operate as compatible gateways between the commercial satellite networks and the terrestrial networks. The architecture, size, growth plan, and other technical questions for CSI development have not been resolved at this time. A reconstituted task force of CSS is addressing these issues.

The standards community, such as the T1Q1, has also initiated a discussion of network and service survivability in the United States (Kaudel, 1989).

1.3 Need for Modeling and Simulation

The broad objective here is to establish how effective or ineffective a switched satellite network would be for the support role of a large terrestrial network under stress. Offhand this may sound simple enough; however, the really germane questions still remain to be asked. For example: What terrestrial networks are to be supported? With what satellite facilities? What traffic, network damage, and other scenarios must be included in the analysis? A closer look at the problem reveals this performance-assessment task to be very complicated. The following factors may have to be included as the very minimum in the problem statement:

- (1) Specification of the unstressed (unperturbed) terrestrial network. Included are lists of nodes, connectivity matrix (or its equivalent), link capacities, delays, and relevant network control (e.g., routing) algorithms.
- (2) Specification of satellite subnetwork or network alternatives. Included are number of satellites, their data rate (bandwidth) capacities, signal switching rates, number of fixed, steerable, and switched beams, beam switching rates, number of Earth stations, gateways, control systems, and their capabilities.

- (3) Specification of the normal (unstressed) offered telecommunications traffic load by amount, type, and service priority.
- (4) Specification of different stress scenarios. Included are damages to nodes or links, traffic overloads (geographically or temporarily).
- (5) Specification of performance parameters to be used in network evaluation. Included should be blocking probability (or blocking grade of service (GOS)), delay, carried traffic, queue statistics, facility (e.g., trunk group) utilization, and so forth.

Individually, these five factors are already quite complex. For example, the terrestrial PSTN consists of more than 20,000 switching offices and on the order of a million trunks (Nesenbergs, 1989). The main interexchange backbone has over 100 ESS4's and a correspondingly smaller number of trunk miles. Even without the satellite part, this topology is much too large for paper-and-pen analysis. As another example, consider network performance. The to-be-ascertained network performance improvement must translate into some quantifiable service advantage for the end-user community, perhaps uniformly over the entire network area or to a different degree in different regions. There may also be different dynamic (time variable) effects on different user classes, such as those assigned to different priority levels.

Mutual interactions between the above five factors make the problem even more difficult to analyze. Appropriate computer representations, modeling, and simulation seem to be the only available tools to handle network problems of such complexity. Fortunately, computer modeling and simulation tools have become more powerful and more prevalent in recent years.

1.4 Outline of this Report

The remainder of this report is structured as follows.

Section 2 outlines the overall modeling and simulation scope, which includes: the types of hybrid satellite-terrestrial networks to be considered, objectives for service survivability under different stress conditions, service restoration after facility damage as well as recovery from severe congestion, and classification of traffic types by either end-user service (e.g., voice, data, video, etc.) or by switching category (e.g., circuit switching, packet switching, etc.).

Section 3 presents an overview of different modeling and simulation methods. It starts with a detailed discrete event-by-event (call-by-call or message-by-message) depiction of the network processes, which can be the most truthful representation of the real world, but also results in the longest simulator runs. Attempts to simplify by some means of aggregation (grouping or clumping) of events, by analytical tools derived from traffic theory, and by such statistical processing of simulator output data

as suggested by Extreme Value Theory (EVT), may yield considerably shorter simulation runs. It is important, however, that adequate statistical levels of confidence be maintained in implementing these shortcuts.

Section 4 addresses the simulation of circuit-switched networks. The service provided by these networks can be viewed as largely voice telephony. However, circuit-switched data traffic and other services should not be excluded from consideration here. Because of the size of the national circuit-switched networks (e.g., the PSN or the interexchange network) and the volumes of traffic carried, special and very efficient simulation methods appear necessary. The most commonly used performance parameter for circuit-switched networks is the blocking GOS. Not to be overlooked, however, are such other parameters as probability of lost connectivity (either to the entire network or to given parts of the network), duration of lost connectivity (i.e., outage time), excessive delays, and a few other secondary measures.

Finally, Section 5 is a summary conclusion of work performed and work remaining, while Section 6 is a bibliography of references cited in this effort.

2. OVERALL PLAN AND OBJECTIVES

2.1 Terrestrial and Satellite Hybrid Networks

The term "network" is used here in the conventional long-haul or interexchange network sense. Consider the small terrestrial network of Figure 1. It is shown to have 12 nodes and 18 links. The node functions, such as switching, take place at the locations indicated by circles. However, when it comes to definition of links, i.e., the connecting lines between the circles, more individual care may be required. After all, depending on the network architecture, one may have to distinguish between physical links (e.g., direct switch-to-switch trunks or spans), logical links (e.g., not necessarily direct trunks, or ORIN's in circuit switching terminology), or virtual links (e.g., in computer communications).

For practical purposes, the small illustrative network of Figure 1 has limited value. It may always be used to test different modeling or simulation tools. However, in the real world of nationwide telecommunications much larger network models are needed. AT&T's backbone network interconnects 114 ESS4's with terrestrial spans that number in the thousands. The regional service areas of many ESS4's may contain around a hundred local central offices and local tandem switches. The now outdated, but highly survivable, AUTOVON polygrid network had plans for 67 switches and over 500 direct links (Joel, 1982). According to recent information, AT&T's Common Channel signaling (CCS) network--a large packet-switched network--has 26 nodes called Signal Transfer Points (STP).

Whereas, it would be of real operational interest to simulate the behavior of 1,000 or even 10,000 node networks in every detail, experience of others tells us that so large a task may be

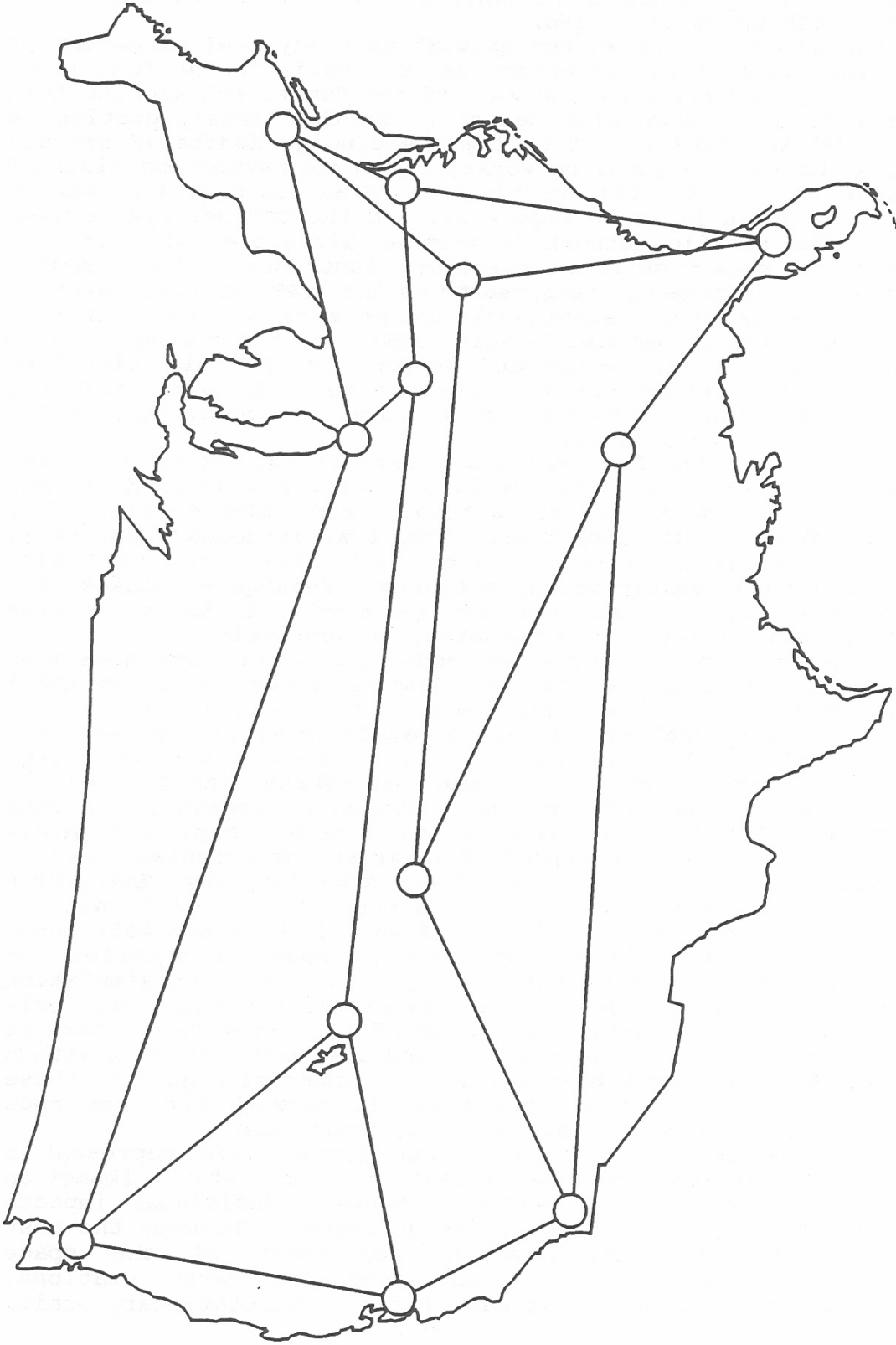


Figure 1. Illustration of a small, twelve-node, terrestrial network.

impossible today. A more realizable goal may be the simulation of networks with up to 100 nodes.

Instead of looking at the network as a physical placement of facilities, it is often advantageous to view it in the functional domain. Figure 2 presents one view of the functional architecture of the proposed future AT&T network. Another interpretation is given by Timko (1987). Notice that, excluding dedicated private networks and international networks, the basic network consists of two parts. They are: the traffic or information carrying network (denoted by solid lines in Figure 2), and the CCS network (broken lines). The traffic network is further divisible into circuit-switched and packet-switched transport functions. The circuit-switched part is commonly referred to as the AT&T Switched Network. It includes as secondary subnetworks the private switched networks, such as the Software Defined Network (SDN), and the coming ISDNs or BISDNs (CCITT, 1989a; Kearns and Mellon, 1990). The signaling networks of the future are projected to be fast packet-switched networks with Signaling System No. 7, commonly abbreviated as SS7 (Modarressi and Skoog, 1990).

Since the traffic volumes, statistics, distributions, processing methods, and other characteristics are quite different for circuit-switched, packet-switched, and ISDN traffic, our overall plan is to consider these three traffic modes separately. The first simulation phase to be discussed here will deal with circuit-switched, mostly voice, networks. Subsequent phases will consider the very different modeling and simulation tools required for packet-switched and ISDN networks, respectively.

An internationally accepted model for data communications network architectures is the Open Systems Interconnection (OSI) network reference model (CCITT, 1989b). It is still too early to tell how the OSI model may or may not apply to voice traffic. For nondata traffic the progress has been slow, much standards development work remains to be done, and substantial further OSI model changes may be expected for telephony. However, for data networks the OSI model is already quite mature and, with minor changes, rather broadly accepted in a variety of networks, such as Local Area Networks (LAN), Metropolitan Area Networks (MAN), Wide Area Networks (WAN), and other almost "universal" network applications (Austin et al., 1989). Figure 3 shows the well-known seven-layer structure of the OSI model. Specific functions of individual layers are summarized in Table 1. For the simulation program proposed here, layers 4 to 7 are of little interest. Only layers 1 to 3, namely the Physical, Data Link, and Network, need to be considered. As seen from the intermediate node processes within the communication subnet boundary in the center of Figure 3, these lower three layers suffice to define all network link and node functions, their outages, and necessary protocols.

The addition of a satellite subnetwork would represent a change for the terrestrial network that may have distinct impact on each of the three traffic switching modes. Individual impacts depend on the resolution of many design issues. Perhaps the most crucial issues are the physical deployment of the space platform(s), their orbit(s), and associated Earth stations. Figure 4 depicts a case of two satellites in geostationary orbit.

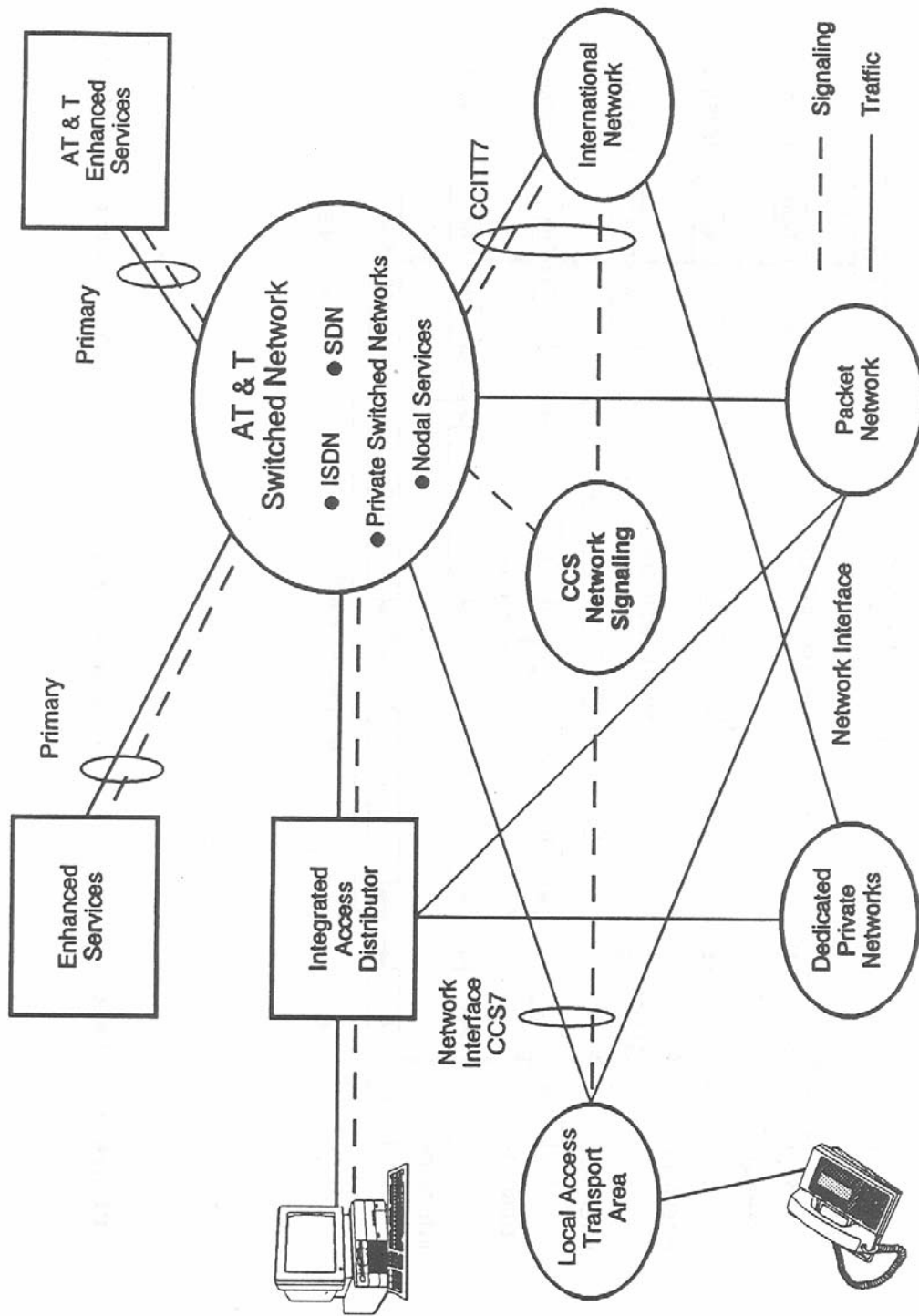


Figure 2. Proposed AT&T network architecture of the future.

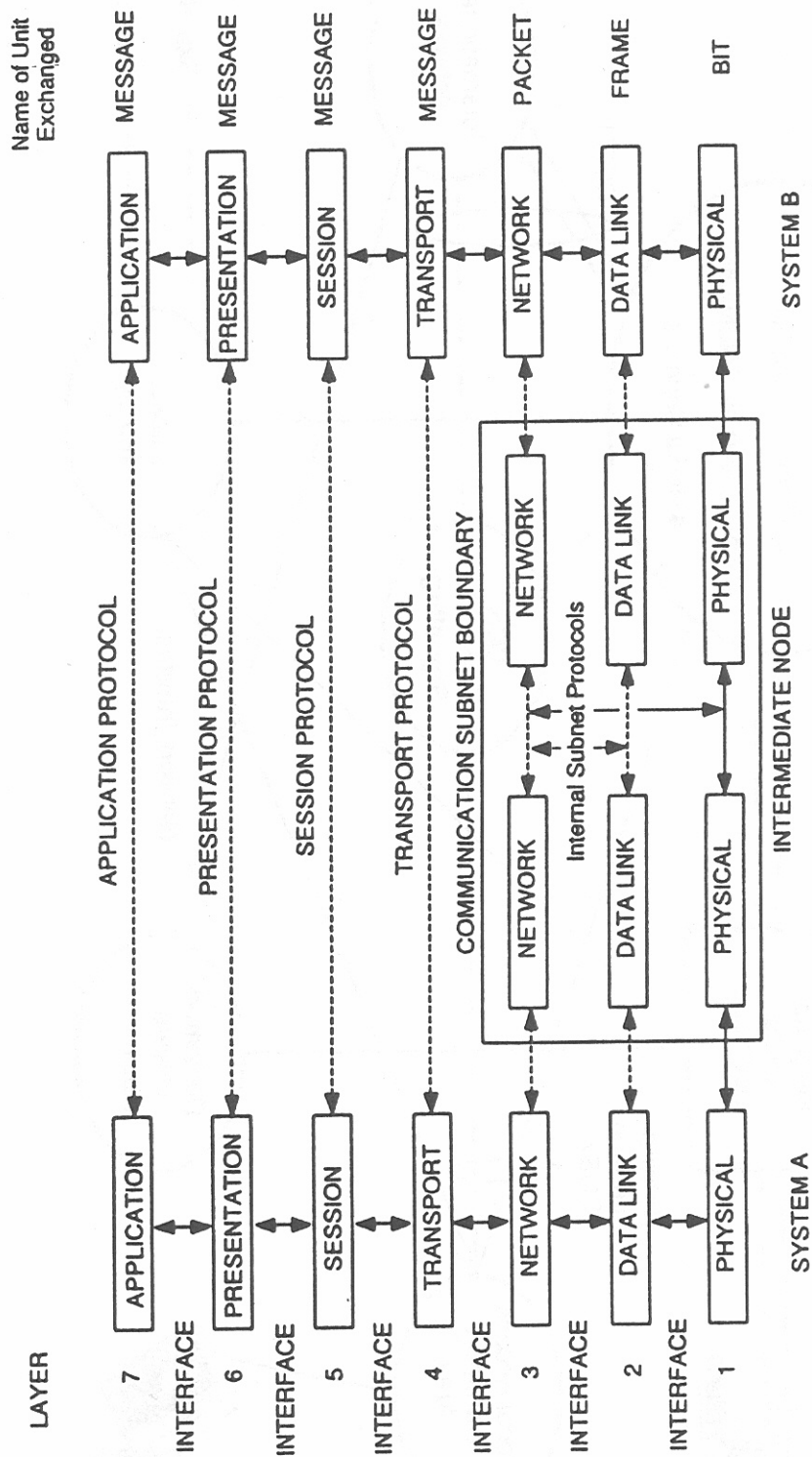


Figure 3. The classical seven-layer model for Open Systems Interconnection (OSI).

Table 1. Description of OSI Layer Functions

Layer Number	Layer Name	Layer Function
7.	Application	Selection of appropriate network service for user's application
6.	Presentation	Code conversion and data reformatting of the user's application
5.	Session	Coordination of interaction between user application processes on different hosts
4.	Transport	Control of network end-to-end processes, such as ensuring data integrity through the network
3.	Network	Switching and routing of information, establishment of logical association for remote hosts, indication of remote connection
2.	Data Link	Ensuring error-free physical links and information transfer over said links
1.	Physical	Provision of physical medium for information flow, devices, and electrical interfaces for bit level data flow

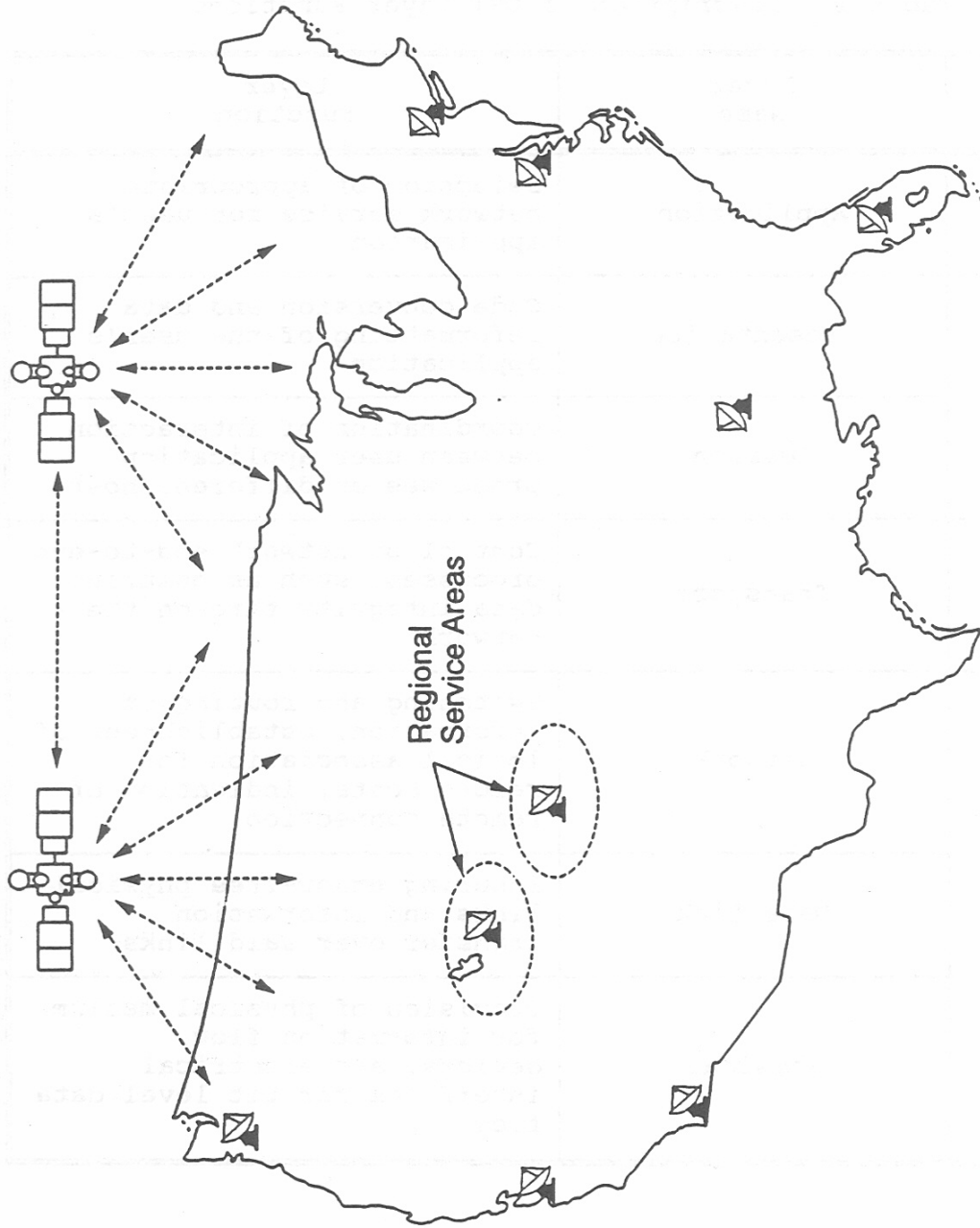


Figure 4. Physical view of the satellite support facilities.

This, of course, need not be so in general. It is an unresolved analysis or simulation issue to determine: a reasonable number of satellites in orbit, the satellite capacities (in terms of bandwidth, switching speed, beam configurations, etc.), their orbit specification (geostationary or otherwise), number and location of Earth stations, gateways, network control stations, and so forth.

The number of Earth stations seems to represent two mutually contradictory requirements. On one hand, since no one knows where the coming crisis events will occur and what their nature will be, a safe strategy may be to place an Earth station next to every major switch in the country, or one per each regional service area. This argument favors a large number of Earth stations. On the other hand, the satellite capacity must inescapably be relatively small compared to the huge capacity of the terrestrial (largely fiber optics) network. In 1989 (Bell, 1990) the AT&T long-distance telephone network is alleged to have carried nearly 225 billion call-minutes. If one assumes the average call to be approximately 3 minutes long, then there must have been around 200 million such calls during the typical 24-hour day. Under normal circumstances the carried busy-hour (BH) load can therefore be estimated to be around 40 million calls or 2 million Erlangs. Yet, in crises or disasters both the offered and carried loads are known to be even larger. This brings the fixed transponder capacity into play. Thus, given any crisis or no crisis, only a small number of stations in a few regions can receive significant satellite traffic at the same time. If the number of Earth stations is large, then most of them must nearly always be idle. The resolution of this efficiency question is one of several simulation objectives.

Computer networks that contain terrestrial links plus one or more satellites have been classified as Very Large Area Networks or VLAN's. Efforts to simulate and analyze VLAN performance have been reported by Wolf and Ghosh (1988).

Another question pertains to the placement of Earth stations. The site selection, a complex task for any reasonably large number of Earth stations (say 10 or more), can be based on several principles. Possible selection principles may involve the following considerations:

- * Traffic statistics (e.g., volumes, national criticality, suitability for existing carriers, etc.).
- * Survival of facilities and overall network connectivity.
- * User services: their availability and quality (e.g., outage probability, outage duration, blocking GOS, delays, QOS).
- * Economics (e.g., costs, market projections).

Economics will not be considered in this report. For network planning based on costs and market projections, see Cruz et al., (1989). Primary emphasis here will be on the characterization of

user services. But, since traffic volumes and facility survivability also affect service quality, both traffic and survivability factors must be part of the overall Earth station siting problem.

Sophisticated algorithms are known for designing survivable networks. As an example, heuristic Lagrangean-relaxation and partial branch-and-bound techniques have been implemented in AT&T's Enhanced Interactive Network Optimization System (E-INOS). AT&T uses E-INOS to design voice, data, and integrated networks for their customers (Agarwal, 1989). On a smaller geographic scale and with less analysis, survivable fiber optic network architectures for intra-LATA use have been studied by Wu et al., (1988). Their general conclusions about the potential protection afforded by facility diversity (as in self-healing network topologies) appear quite valid also for nonfiber networks and for inter-LATA applications.

Whether the network is terrestrial or a satellite-terrestrial hybrid, an essential instrument for efficient message transport is a correctly chosen routing algorithm. Many different routing schemes are used now and many more may be used in the future (Chang, 1989; Cain et al., 1989).

Overall, the satellite-terrestrial hybrid network is expected to have a functional architecture similar to the terrestrial network. Figure 5 shows a possible and likely future architecture. Note that in this particular model the satellite subnetwork avoids the Local Access Transport Areas (LATA) by interfacing on the toll side of the interexchange switches. (Other architectures can be defined that place satellite interfaces within LATA's.) The satellite subnetwork carries both switched network traffic (including packet messages) and CCS signaling. The latter is needed for satellite on-board switch management, the satellite net control center, and all internetwork gateways and interfaces.

Gateways between satellite and terrestrial subnetworks often pose design challenges in the real world of internetworking. An Earth station that serves one or more terrestrial switches may encounter signals from different generation systems, by different manufacturers, and in agreement with one of many protocols and existing standard signaling systems (Joel, 1982). Agreement on a common CCS with SS7 would appear to be very beneficial here.

For data and computer network interconnection, standardization status is more advanced. In the United States, the Department of Defense (DoD) and the ensuing INTERNET community have advanced the standard Transmission Control Protocol (TCP) and the Internet Protocol (IP) (Groenbaek, 1986). Internationally, the CCITT has developed its own concatenation scheme for virtual circuits. Moreover, according to the latest ISO announcements, the two IP's will soon be capable of interoperation (Sunshine, 1990). A recent and comprehensive bibliography of network interconnection has been prepared by Biersack (1990).

Simulation program is planned for voice and data services. Voice transmissions can be both analog and digital, but only the digital version will be considered here. Three basic types of switching can serve the voice service: dedicated (or nonswitched), circuit switched, and packet switched. A breakdown of services and

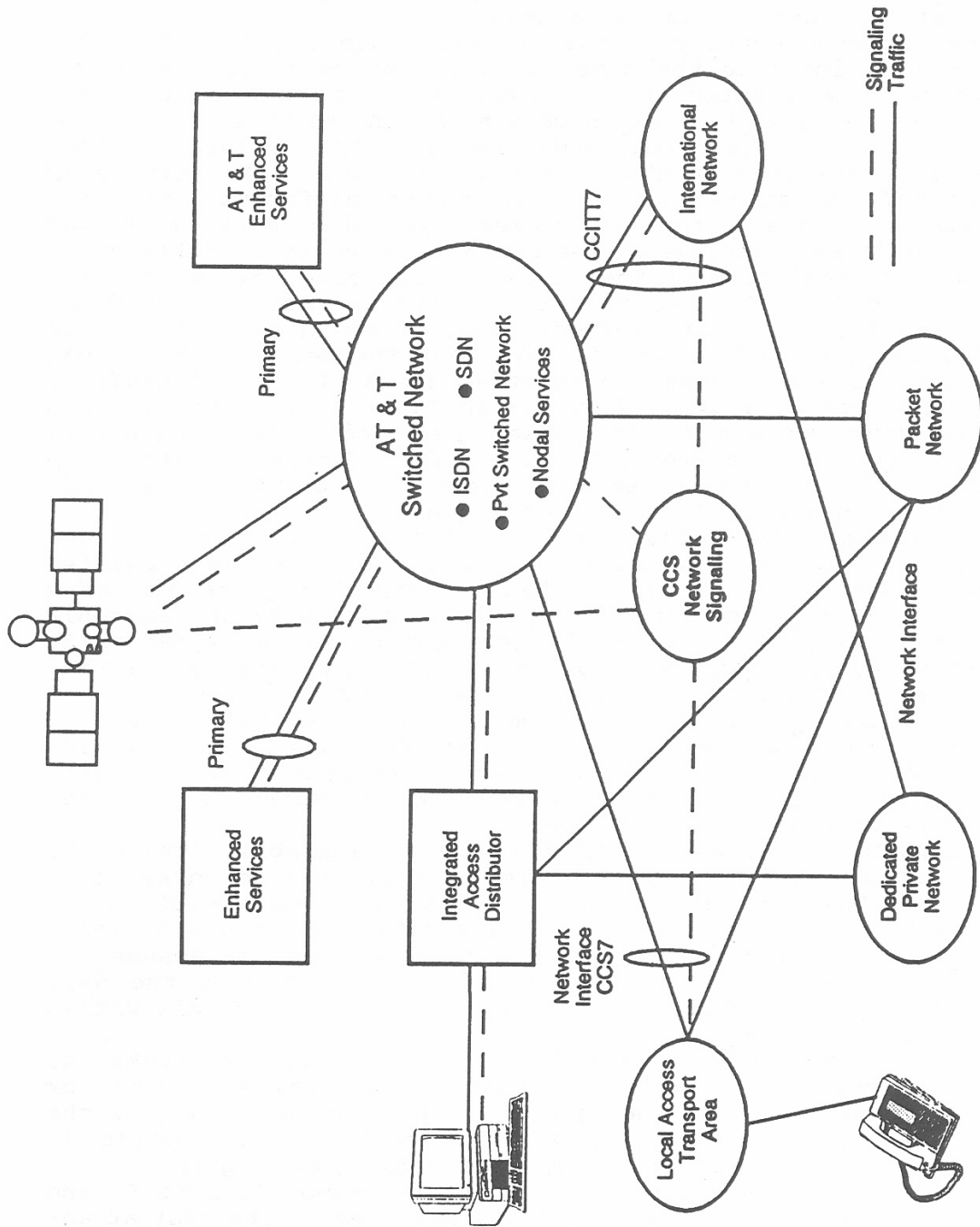


Figure 5. Functional architecture of the satellite-terrestrial hybrid.

switching classes is presented in Table 2. The left-hand column shows the three basic services. The other columns refer to network switching. Checkmarks indicate service/switching combinations that exist, but will not be simulated here.

The numbered entries in parentheses, such as [1], [2], [3], [4], and [5], indicate the order in which these particular cases are likely to be treated in the proposed ITS simulation program. Thus, in the initial phase of voice simulation, 64 kb/s (or perhaps 32 kb/s) calls will be circuit-switched over the network. Later, this will be followed by the second phase, where data service is accommodated by either or all of three different switching arrangements. The first is packet switching over dedicated channels of fixed capacities. The second is circuit-switched data, which appears analogous to the digital voice case, except for the necessarily different interarrival and service duration (holding) times. The last pure data service case, if one goes that far, is one of sending packets over channels that themselves are circuit switched in or out to meet the relative needs of offered traffic. In the final phase, indicated by [5] in Table 2, integrated voice and data services are to be packet switched. More extensive service integration, as for ISDN with video, facsimile, etc., can be either inferred from these five cases or developed as another step in the modeling and simulation sequence.

The network topologies to be simulated will be expanded versions of the oversimplified U.S. domestic backbone suggested earlier in Figure 1. A representative network may consist of N nodes and L links, with topology specified either by the commonly used connectivity matrix C or the less common distance matrix D . Other matrices or lists are used, as needed, to define bandwidths, capacities, relative facility use "costs", and other characteristics of network nodes and links. Note that, if one is interested in networks with N as large as 100, then C and D are 100×100 dimensional. To avoid long and time consuming routines, which may arise in routing processes and elsewhere, such large matrices require special algorithms.

The number of links, L , can also be considerable. While the general bounds $N-1 < L < N(N-1)/2$ are valid for all networks, they are too wide to be of direct use. For instance, when $N=100$ then L must be somewhere between 99 (for a tree network) and 4,950 (for a fully connected network). A specific topology must be assumed to pin down a value for L . This issue is illustrated by the $N=12$ network of Figure 1, where the actual value $L=18$ falls well within the range $11 < L < 66$.

When talking about network links (or channels, or trunks, or lines), one must also clarify the distinction between actual, or physical, connectivity on one side and logical connectivity on the other. Perhaps the easiest way to demonstrate this point is through a graphical plot of a simple network. See Figure 6.

Part (a) shows six circles or switches, numbered 1 to 6, and links or circuits that interconnect the switches. Note that at any given switch, the circuits may be of two types: those that terminate on the switch and those that merely pass through it. Take for instance switch #6. There are three trunks that terminate on #6 and one that does not. The three terminating trunks go to

Table 2. Selection of Services and Switching Categories

Service	Dedicated Circuits	Circuit Switched	Packet Switched		
			On Dedicated Channels	On Circuit Switched Channels	On Integrated Channels
Analog Voice	✓	✓			
Digital Voice	✓	[1]	✓	✓	[5] (Joint)
Data	✓	[3]	[2]	[4]	

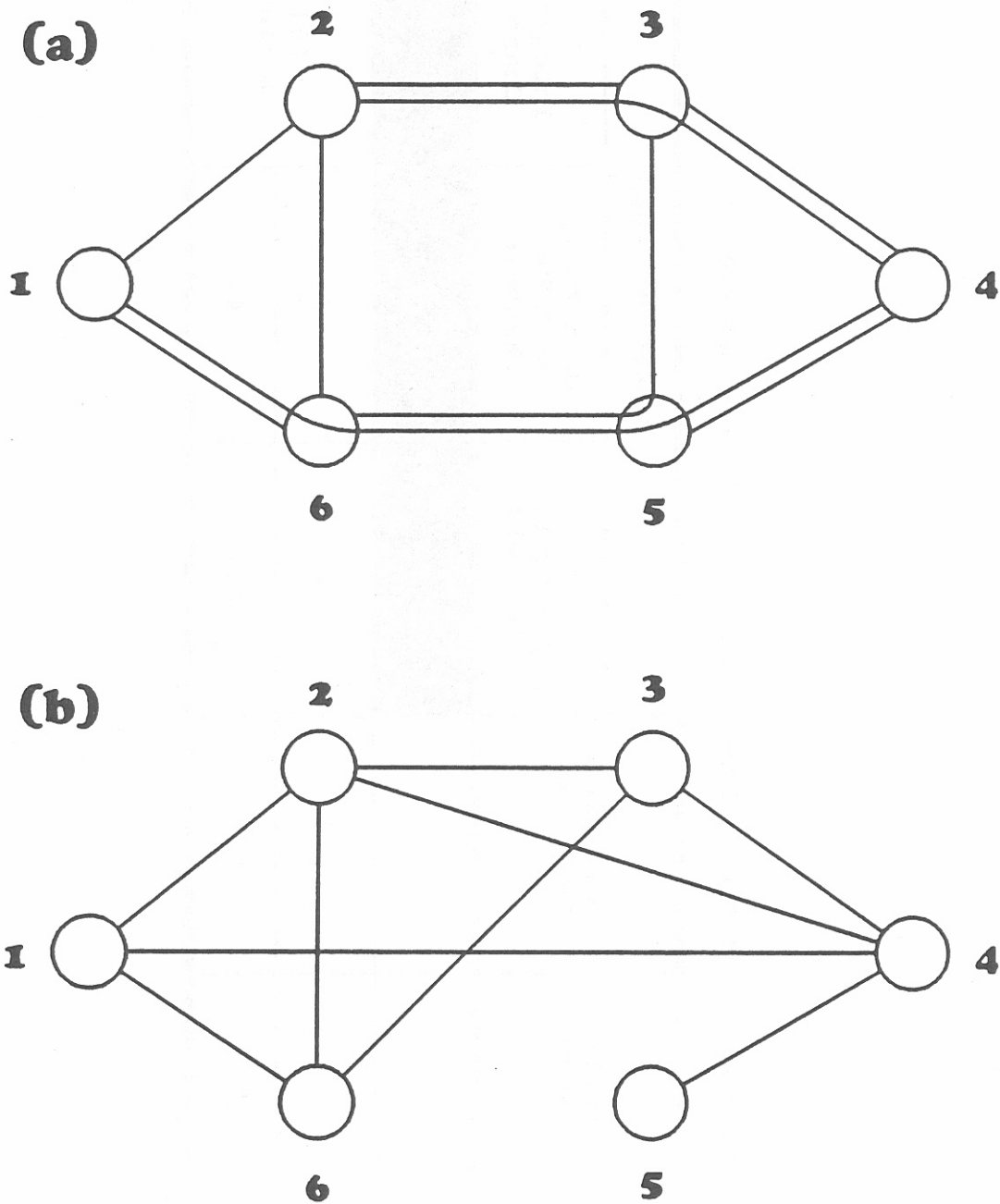


Figure 6. Distinction between physical links (a) and their logical connectivity (b).

nodes #1, #2, and #3, while the fourth trunk is dedicated to through traffic between #1 with #4. Traffic that arrives on the three terminating trunks is either switched to another terminating trunk or is serviced by a local user terminal. The other traffic, i.e., the one on the dedicated trunk, is not switched at node #6. Likewise, it is also not switched at other nodes where the line is shown to pass right through the circle.

The aggregation of all direct links between two switches is called a span. Specification of all spans and their terminations uniquely determines the physical connectivity of a network. Thus, part (a) of Figure 6 represents the physical connectivity of the network.

Part (b) is the logical equivalent of the same physical network. This logical topology is constructed from the physical topology in part (a) by deleting trunking through all the nonswitched tandem nodes. The result is a direct connectivity map, as shown in (b). Observe that, whereas the #5-#6 span consists of two physical links, there is no direct logical connection between #5 and #6. All the traffic that arrives at a switch in the logical topology (b) must be either switched or it must terminate at that switch.

Trunks suffer physical damages and operational outages usually as a consequence of cut or otherwise destroyed spans. In the logical domain this may translate to a deletion of several logical links. For example, a cut of the #5-#6 span in (a) is equivalent to cuts of #1-#4 and #3-#6 links in (b). In survivability and restoration modeling of networks one finds both physical and logical representations useful for different purposes. Physical stress or damage starts in the physical model. Simulation of traffic and network processes is, however, more likely to take place almost entirely in the logical domain.

To simplify matters, let us make the following assumptions. Unless otherwise stated, all the switches considered from now on will be ideally nonblocking and all the links will be full duplex channels. However, their capacities (data rates, bandwidths, number of equivalent voice channels) will be allowed to vary, so as to represent different network node facilities and transmission media.

2.2 Survivability and Restoral Objectives

Questions about survivability usually fall into one of two categories. First, in most user's mind, is the survival of end-user service. The concern here is with service availability, as well as the quality, blocking, delays, and other degradation characteristics of the surviving service (Richters and Dvorak, 1988; Brush and Marlow, 1990). The second category of survivability pertains to survivability of network facilities. That includes links, network connectivity, nodes (switches, data bases, and others), their capacities and speeds.

The objective here is to assume physical survival, or its converse--outages--, as a given initial condition for the simulation process. In a sense, we start by assuming a specific

damaged or undamaged network status. This status could be the result of various crisis scenarios, such as those postulated in the NCS (1989c) network connectivity analysis model (NCAM) or elsewhere (Wu et al., 1988; Newport and Varshney, 1989; Schroeder and Newport, 1989). Given the assumed network status, one next seeks by simulation the service characteristics for different end-user categories: between specific nodes, geographic regions, over the entire network, for certain terminal types, different priority classes, and so forth.

Restoral also falls into the same two categories. There is the restoral of end-user services and the restoral of network components and functions. Each category has a number of descriptive parameters that broadly divide into terms that pertain to restoration speed (time for reestablishment of certain features or functions) and restoration extent (provision of capacities, coverage regions, etc.).

The new element in restoral simulation is time. The objective here is to study the dynamic or time-varying statistics during the restoration process. To do so, the stress (damage, traffic overload, etc.) must also be a variable in time. The scope of damage seems to be adequately represented by one or more step functions in time. Traffic overload can also be a step function, with a duration that is of the order of one or more hours. The output of simulation runs is intended to show how a performance parameter (blocking GOS, delay, throughput, etc.) behaves before, during, and after crisis events. In other words, how effective are existing terrestrial and proposed satellite-terrestrial hybrid networks in restoring services? And at what specific capacity levels?

Service maintenance and service restoral in damaged networks depend on several other factors that may or may not be significantly enhanced by redundant terrestrial topologies. One of the most helpful factors may turn out to be the proposed additional satellite network. Other factors belong to the algorithmic world of network and traffic controls. Routing, whether it is local or global, fixed, alternate or adaptive, affects the network performance (Gifford, 1987; Mouldin et al., 1989). A recent example is the emulation of flooding (or send everywhere) techniques on the Telecom Canada long haul network. Using the so-called self-healing features of the topology, it has been shown that optimally fast restoration is possible without any need for centralized control, disaster pre-plans, or data bases of any type (Grover et al., 1990).

Different control and routing algorithms will be defined and used as part of the simulation packages. Optimal restoration by rerouting to satellite backup facilities should be automated and fast, say in a matter of seconds (not in minutes, hours, or days, as required for manual switchovers). Because the nature and scope of any network damage can hardly be predicted in advance, an adaptive algorithm, or perhaps a family of different adaptive algorithms, should be kept on hand to suit varying circumstances. Local recovery controllers would be tasked to select automatically the most appropriate of the routing algorithms.

The normal terrestrial controls, as exemplified by the CCS subnetwork, may quite likely encounter interface difficulties when interacting with satellite links. One important issue has to do with protocol conversion and availability of standards and products. This study plans to ignore the many difficult problems of protocol conversion, at least in the initial phases of the simulation program. Subsequent simulation studies may have to address more detailed performance questions of standards and protocols.

Another set of difficulties is apt to be caused by the 0.25 second up-and-down propagation delay through a geostationary satellite. Under heavy traffic, the signal transfer points (STP) could be severely slowed, if and when they were required to wait that long for inter-STP signal exchanges. That slowdown will introduce call (message) set-up delays, overload the STP registers, and very likely cause software-related timeout problems. Perhaps further simulation will reveal under what conditions CCS packets should be allowed to use satellite paths, and when not. Of course, a low orbit satellite network may alleviate this delay problem.

2.3 Traffic Classification

From the end-user applications point of view, the offered traffic consists of voice, data, video, facsimile, and other less prominent services. In the network, which can be considered to be the PSN (and especially so in its interexchange segment), traffic is seen in a different light. An information transaction, be it a telephone call or a data message, occupies a specified bandwidth channel for a certain period of time, and requires handling by the network control resources.

The actual time assignments are determined by the switching regime of the network, combined with the statistics of the service requests (i.e., interarrival and call or message duration distributions). These statistics are sufficiently different for both voice and data, and circuit and packet switching, so that entirely different models and analytical methods are necessary for the representation of the different classes.

Historically, all data communications volumes, costs, and revenues have been growing (Dunn and Johnson, 1989). Data traffic constitutes the most significant traffic growth component today. Public data services constituted about 1% of total common carrier revenues in the early 1960s. In 1987 that percentage had grown to 10% and, if the present trend of more than 20% annual growth rate continues, public data revenues should reach 50% of the total by 1997. Without more exact numbers from the industry, the 1990 percentage revenue for data can be estimated to be between 15% and 20% of the total.

The remaining 80% of today's revenues (and almost the same percentage of today's traffic) come from voice telephony. Almost all of this voice traffic is circuit switched.

For all categories of traffic loads, the simulation objective can be phrased as one or more of the following questions:

- * How effective (or ineffective) is the adaptive switched satellite network likely to be?
- * To meet specified effectiveness levels, what must be the capabilities of the satellite network?
- * What is the optimum or nearly optimum satellite network configuration?

To answer any of these questions, one must realize that the satellite support will have to deal largely with traffic that the terrestrial system fails to handle (e.g., overflow traffic). This overflow is apt to have at least one of four rather unique characteristics:

- (a) The average overflow traffic level is very low or actually zero for long periods of time.
- (b) The offered overflow traffic is characterized by infrequent but high peaks. Thus, it must be classified as irregular traffic with a variance that is much larger than the mean.
- (c) The overflow traffic can be geographically focused on a few dispersed, perhaps randomly changing, locations.
- (d) The overflow traffic characteristics may be unsuitable for normal handling by the PSTN. Examples could be requirements for ISDN, excessive bandwidth, very high data rate, special format, etc.

Various statistics, such as means, variances and peakedness factors, have been used by traffic theorists for the overflow traffic, as well as for cases where the offered load is changing (i.e., time-varying or time-dependent) to a different degree (Jagerman, 1975; Akimaru et al., 1986; St. Jaques and Stevens, 1989). Generally, the peakedness factor of variance/mean tends to be in the range between 1 and 3. However, exceptional cases are possible where the factor is less than unity or higher than 3. To avoid obtaining misleading results from sampled data, it seems important to sample traffic sufficiently often. Data derived from 1-hour sampling intervals can lead to severe underestimates of traffic variability. Half-hour data appears to yield a marked improvement (Bridges and Sen, 1990), while 15-minute data may be the most workable solution for normal voice traffic with an approximate mean holding time of 3 minutes (Yunus, 1987).

Finally, standardized and structured traffic representation has been recommended by the CCITT (1989c).

3. MODELING AND SIMULATION METHODOLOGIES

In the realm of network performance estimation, the terms of modeling and simulation are often used concurrently (Balaban et al., 1984; Koval et al., 1986). Yet the functions performed and products generated by the two activities are quite distinct. Figure 7 indicates this distinction for a general network whose performance is to be ascertained.

The network is described by its topology (see Section 2.1). That topology includes a possible detailed specification of all nodes and links, including their capabilities and restrictions. Traffic and stress are two other factors that strongly influence the operational status of a network. Here traffic can refer to the normal, engineered offered load on the network, while stress can identify physical facility damage status plus excessive, crisis generated, traffic surges.

Modeling is a process of network representation. Typically it consists of tables, parameter definitions, distributions and moments for random variables, statistics, equations, executable functions, logical rules, and assorted computer programs that, in one way or another together represent the "network" and its behavior. All these descriptors must be suited for the computer environment, facilities, data bases, and other processing tools used in the eventual simulation.

Many telecommunication networks are known to be large and complex. It is therefore important that the models be made as simple as possible, yet retain the key features essential to network evaluation. To reduce the computational complexity (computer time requirement), various approximations must often be considered. More about approximate models will be said later. Finally, analytical shortcuts (such as well established properties of certain traffic classes, random number generation, routing algorithms) should be exploited to shorten the simulation runs.

Simulation is the actual execution of software programs that represent basically two things: the network model and the application objectives. These objectives are almost always associated with performance. In the specification of the simulator it is important (for practical reasons) to find a fast, flexible, compatible, and available software. Similar features are essential for hardware, with processor speed playing perhaps the foremost role.

Performance numbers are the outputs of one or more simulator runs (Ilyas and Mouftah, 1985). These numbers and their accuracy are a direct result of the model used. Many performance-evaluation criteria are used. Quite broadly these criteria fall into at least three categories: end-user (applications) oriented, manager (or network operator) oriented, and network-designer (planner or implementor) oriented.

In all modeling and simulation exercises there seems to be room for questions and arguments about the validity of the performance numbers. How realistic are they? What are their levels of confidence? What interpretation and extrapolation is consistent with the assumed model? Perhaps another model should have been used? What is the best model under given network and

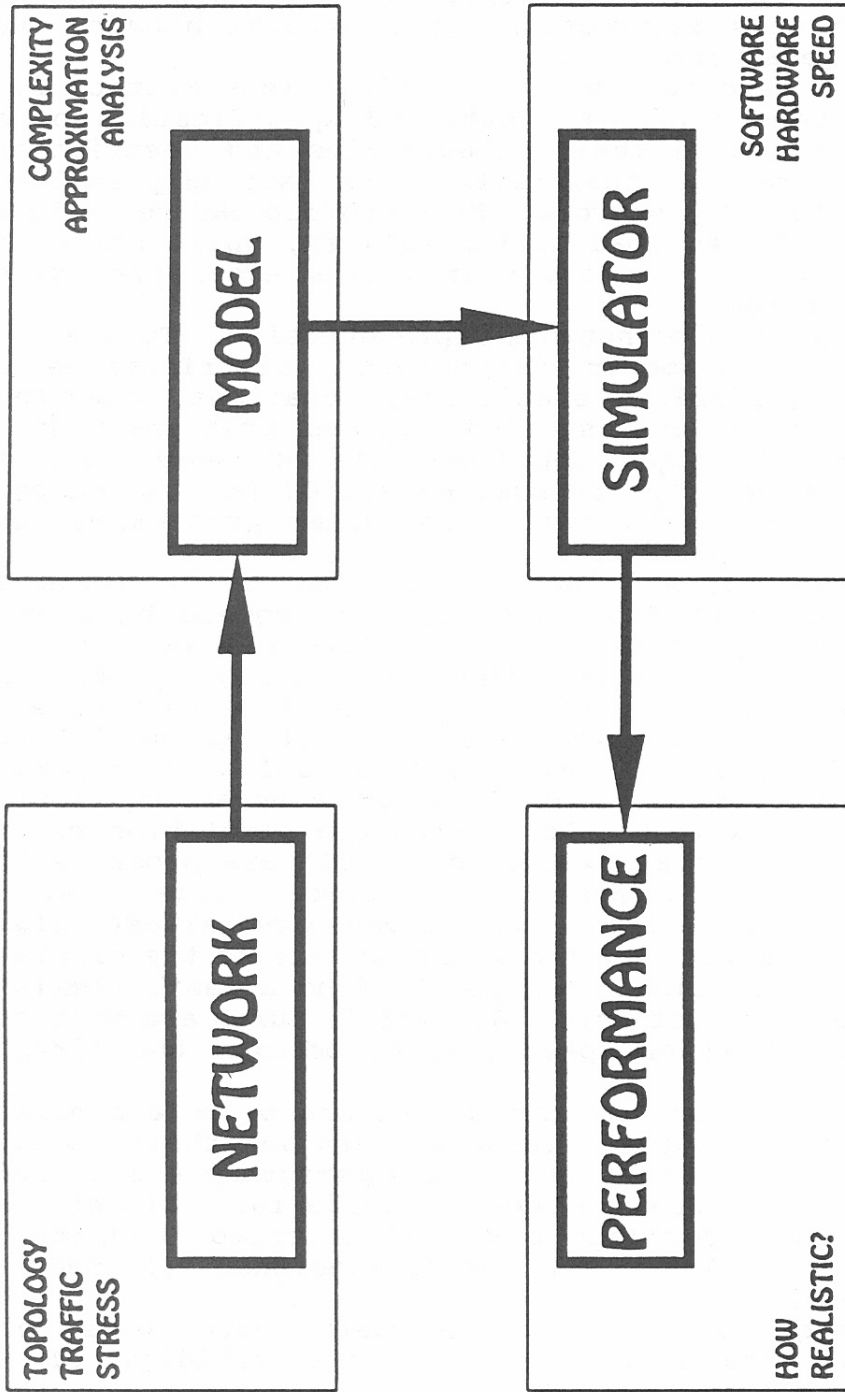


Figure 7. The roles of modeling and simulation in network performance estimation.

application constraints? It seems clear from this that performance assessment is closely tied to the original selection of the network model.

The importance of modeling extends naturally to the software used in both model representation and in the simulation process. The software, including the languages for simulation, can be selected on the basis of operator preference (familiarity), availability, cost, and several operational criteria (Mouftah and Shanmugan, 1987). Table 3 presents a brief outline of key evaluation factors for selection of simulation languages. Note that in this table the featured items in left and right-hand columns need not have a one-to-one correspondence. An item, such as CODING CONSIDERATIONS on the left side, may be reflected in Ease of Coding, Code Support, Self-Documentation, as well as Ease of Debugging on the right side, and so forth.

In years past, such higher level languages as FORTRAN and Pascal were frequently used in simulation software packages. Today, however, the language preferred by most may be the C or C++, combined with the Unix operating system (Kernighan and Ritchie, 1978; Kochan, 1983; Waite et al., 1983; Harbison and Steele, 1987). Unfortunately, by themselves these languages appear inadequate for simulation applications to telecommunication networks. Special pre-processor languages are used to free the operator from repetitive network representation tasks, for example, inputs of topologies and block diagrams. Since these simulation languages tend to be application specific, more about them will be said in Section 4.

As a final comment on general modeling methods, one should mention the relevant standards work by CCITT (Saracco and Tilanus, 1987). For more than 20 years, CCITT has been developing a general specification and description language that satisfies two objectives:

- * To specify precisely the functional features provided by existing systems.
- * To describe as precisely as possible the anticipated functional features of future systems.

The result is the standard CCITT systems description language or SDL (CCITT, 1989d; Moretti, 1990). Table 4 shows selected examples of network-oriented SDL symbols and their standard definitions.

3.1 Discrete Event Methods

Depending on the nature of the traffic, the term "event" is an abbreviation for one of many, clearly defined, network occurrences (moments or epochs) associated with a unique single information transaction. In the case of successful telephone calls, separate events can refer to start of dialing, start of conversation, end of conversation, and various events associated with access and disengagement delays. In the case of blocked, busy, or otherwise

Table 3. Features on Which to Evaluate a Simulation Language (from Mouftah and Shanmugan, 1987)

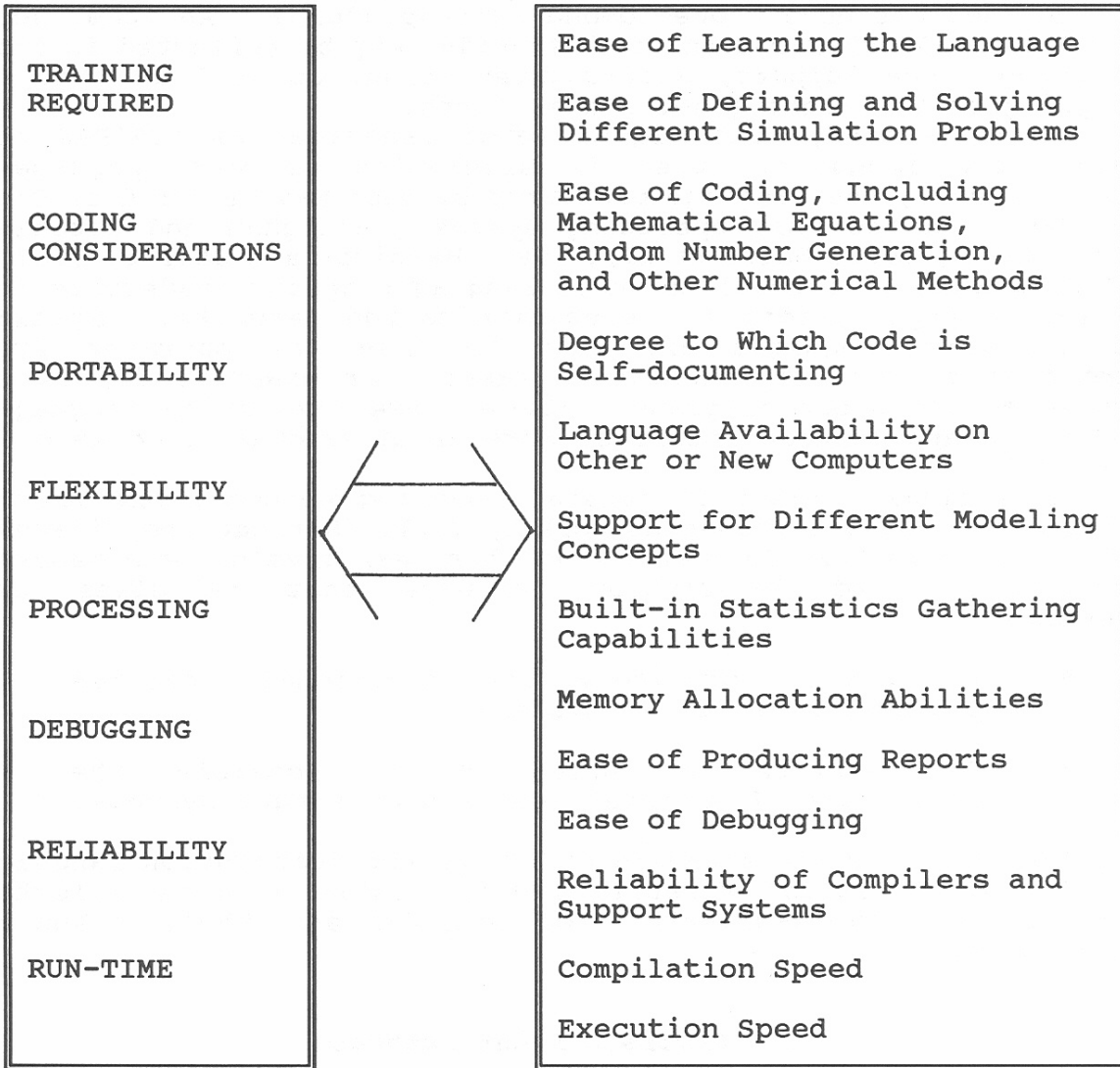

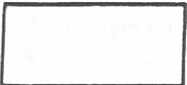










Table 4. Examples of Standard SDL Symbols and Their Definitions

	<p><u>Signal route symbol</u>-- A signal route symbol leads either from one process to another, or from a process to the origin end of a channel, or from the destination end of a channel to a process.</p>
	<p><u>Task symbol</u>-- A task is used to represent operations on data performed on a transition, with the exception of output signal generation and decisions.</p>
	<p><u>Input symbol</u>-- An input symbol attached to a state means that if the signal named within the input symbol arrives while the process is in this state, the transition that follows the input symbol should be interpreted.</p>
	<p><u>Output symbol</u>-- An output symbol represents the sending of a signal from one process to another.</p>
	<p><u>State symbol</u>-- A state is a point in the process where no actions are being performed but where the input queue is monitoring for the arrival of incoming signals.</p>
	<p><u>Decision symbol</u>-- A decision is an action within a transition that asks a question regarding the value of data items available to the process at the instant of executing the action.</p>
	<p><u>Option symbol</u>-- The option symbol is used to describe several alternative process behaviors with one diagram.</p>
	<p><u>Connector symbol</u>-- Any flow line may be broken by a pair of associated connectors, with the flow assumed to be from the out-connector to the in-connector.</p>
	<p><u>Return symbol</u>-- This symbol implies a return to the initial process state.</p>
	<p><u>Stop symbol</u>-- This symbol implies the end of the process.</p>

incompleted calls, one can likewise distinguish several detailed events. Or, if the network resource and time losses are negligible, one can refer to the whole incomplete attempt as one event. Similarly for data messages, different event categories can define start of transmission, end of transmission, start of reception, end of reception, as well as specific instances of entering and exiting network queues.

When model simplicity is of importance, the number of event categories is reduced. Thus, for a large network carrying mostly telephone traffic, the start and end times of a phone call (plus the statistics on blocked calls) may be the most significant events worth retaining.

Discrete event methods have been used most successfully for small networks or those with relatively low traffic. Thus, private data networks, LAN's, and the like, have seen the largest benefit from discrete event-by-event simulation. The large public networks with millions of subscribers may require excessive simulation run times. So far, event-by-event or even call-by-call models have not been useful for larger networks such as the nationwide PSN.

Consider the following simple example. Let:

N = Number of network nodes

T = Number of terminals (telephones) per node

C = Average number of calls originated by a single terminal per busy hour (BH)

B = Average number of basic machine instructions executed for a single call

S = Speed of the simulation machine, expressed in number of basic machine integer instructions per second (MIPS).

Then the average time needed to simulate one network BH can be inferred from the time expansion factor

$$\frac{\text{Simulator time}}{\text{Network time}} = \frac{NTCB}{3600 S} .$$

Due to random variations in offered traffic and perhaps in the ways the network handles traffic, the actual simulation times for repeat runs are likely to be different. However, for network times of one busy hour or longer, these random variations are not expected to be significant.

In a typical PSN scenario one could have

N = 1,000
T = 10,000
C = 5
B = 2,000
S = 1,000,000 .

Then the expansion factor turns out to be near 28. This means that a single simulator run lasting 28 hours would represent one hour of real time on the network. Often many runs are necessary to assess the effects of different network parameters, such as the logical connectivity, traffic, stress, network control algorithms, and so forth. Assume that one wants to evaluate 5 parameters at 4 values or levels each. Then $4^5 = 1024$ simulation runs must be made and the total uninterrupted simulator time would amount to some 28,000 hours or 3.2 years!

One of the earliest reported call-by-call simulators for circuit switched networks was UNISIM (Gimpelson and Weber, 1964). The UNISIM principle of treating each call as a separate entity applies both to event-by-event and call-by call simulators.

COMNET, GSS, OPNET, and BONEs are examples of four relatively recent telecommunication network simulation products. Section 4.2 will give more detailed descriptions of these simulation products.

There exist many other modeling and simulation software packages that have been specifically designed for computer networks. As such, they may only be of marginal interest for circuit switched networks. However, they may be extremely useful for applications to packet or queueing networks. Most of the existing tools, such as RESQ (Sauer et al., 1984), PET (Bharath-Kumar and Kermani, 1984), NASSIM (Mouftah and Bhatia, 1984), and GENESIM (Doner, 1988), appear to oriented towards theoretical research.

3.2 Aggregated Methods

Several techniques have been devised to shorten the execution times of simulator programs. The principal method is to reduce the number of separate events that need to be represented and processed in the sequence of repetitive simulation tasks. To that end one can aggregate (or lump together, or scale, or compress) either the traffic or network facilities, or one can substitute known system performance numbers (from tables or equations) for certain previously established conditions. Traffic can be aggregated:

- (a) in time (e.g., per second, minute, or hour);
- (b) topologically (e.g., over local regions);
- (c) by service category (e.g., priority or type, such as voice, data, video);
- (d) by functional processes exercised within the network (e.g., different switching, routing, processing regimes);
- (e) by scaling, with the proportionality constant to be determined, both traffic and capacity (or the number) of network facilities;
- (f) by any combination of the above.

Traffic aggregation methods (a) and (b) appear most promising for voice telephony, and hence for circuit-switched networks. Methods (c) and (d), on the other hand, may have more promise for large packet-switched and ISDN networks.

A useful variation of the aggregate methods for telephone networks is the Initial Voice-switched Network (IVSN) model and its enhanced version, the Queueing Traffic Congestion Model (QTCM). The original work on the model was done by S. S. Katz at Bell Telephone Laboratories (Katz, 1967). The subsequent improvements that led to IVSN and QTCM were done at Bell Telephone Laboratories, the Supreme Headquarters Allied Powers in Europe (SHAPE) Technical Centre in the Hague, Netherlands, and at Booz-Allen & Hamilton in Bethesda, MD (NCS, 1989b).

The Katz model (e.g., IVSN) takes a reverse approach to network performance assessment. The actual executed functions of the model are rather complicated due to the complex handling of routing algorithms and interactions between physical and logical circuits in the network (see Figure 6). What follows is a very simplified overview of the IVSN network-handling process.

At various service nodes of the network, target blocking grade of service (GOS) values are assumed. (To simplify, the same GOS can be assumed everywhere.) Then, using the offered load numbers and the inverses of blocking probability formulas known from traffic theory, the model estimates the minimum circuit requirements between various nodes in the given connectivity. If all the traffic substreams destined for different trunk groups were purely random, that is, with variance equal to mean, the old Erlang formulas would suffice (Kleinrock, 1975). Unfortunately, networks that permit alternate routing must occasionally pass overflow traffic from one trunk group to another. This overflow traffic is known to be nonrandom in the sense that its variance is often much larger than its mean. The equivalent random method of Wilkinson and related tools are then utilized (Wilkinson, 1971).

For a given network and GOS objectives, an IVSN run consists of a number of sequential steps. At each step, based on previous blocking and overflow numbers, a new network configuration is established. It generates a new set of blocking and overflow estimates. In principle this type of procedure could continue indefinitely. Convergence tests and truncation rules are used to terminate the process.

The following five assumptions appear most basic in the Katz (1967) model:

- (1) At all nodes the offered traffic obeys the Poisson call arrival and exponential service duration distributions.
- (2) The carried loads on different trunk groups are mutually independent and in statistical equilibrium.
- (3) The probability distributions of all simultaneous number of calls, on all trunk groups, are

adequately represented by their means and variances.

- (4) If all routing attempts fail for a call, that call is said to be blocked. Blocked calls are assumed lost without any aftereffect to the system.
- (5) Switches and other nodal facilities are nonblocking. This means that blocking can only be caused by a shortage of server trunks.

To use the IVSN or the QTCM simulator, the following four inputs are required:

- (1) Topology: locations of all nodes (switches) and their interconnecting--physical and logical--trunk groups.
- (2) Routing algorithms.
- (3) Estimates of average offered load for each node-to-node pair.
- (4) Minimum and maximum, or just maximum, blocking GOS bounds for end-users at each node-to-node pair. Note that this leaves the blocking numbers at tandem switches arbitrary, to be manipulated by the simulation program.

For present purposes, it must be mentioned that IVSN (NCS, 1989b) and presumably also QTCM software can be modified to analyze, that is to estimate, the end-user GOS for any fixed network, instead of merely seeking trunk group capacities to meet given GOS numbers. Such an application has been demonstrated at the MIT Lincoln Laboratories to analyze a circuit switched network (Lippmann, 1982).

A unique system modeling tool, borrowed from natural sciences, such as biology, medicine and ecology, is compartmental modeling (Garzia and Lockhart, 1989). This technique has been recently applied to the study of nonhierarchical circuit-switched networks. The approach is based on the concept of a "compartment" and flows (or exchanges) between compartments. For communication networks a compartment may represent one particular category of traffic. The totality of all compartments, that is, of all possible traffic types (by origination, by destination, and either offered, carried, lost, etc.), represents the network state vector. Assume that the state vector has k compartments. Then the k time derivatives of the state vector show all possible rates of transfer between different traffic categories. For circuit-switched networks this leads to k difference equations, solutions of which yields both the transient and the steady-state behavior of the network.

Various network topologies and node functions (e.g., routing algorithms, processing speeds, etc.) can be incorporated into the set of difference equations. However, what is typically lost in

compartmental modeling is spatial or topological resolution. Because, to achieve reasonable solution times for large networks with many compartments, the compartments are assumed to be homogeneous (i.e., with uniform distributions over the network). Whereas this premise of traffic homogeneity does complicate the interpretation of the simulation results, the compartmental model may nevertheless be quite useful. It is capable of demonstrating the time-dynamics of traffic that would occur in various sudden stress events, such as temporary traffic surges, probable node outages, different retrial models for blocked calls, activation of different routing and access controls, and so forth (Garzia and Garzia, 1990; Garzia, 1990; Garzia and Lockhart, 1990; Lockhart, 1990).

3.3 Statistical Design and Processing

It has been stated above that simulation of large and complex networks consumes large amounts of time and resources. In practice, of course, both time and budget are limited assets. Nevertheless, simulation runs have been known to generate large volumes of data with recognizable features that for a variety of reasons may be difficult to interpret statistically. The situation is forcefully summarized in Pawlikowski (1990), from which we quote:

Applying simulation to the modeling and performance analysis of complex systems can be compared to using the surgical scalpel (Shannon, 1981), whereby "in the right hand [it] can accomplish tremendous good, but it must be used with great care and by someone who knows what they are doing." One of the applications in which simulation has become increasingly popular is the class of dynamic systems with random input and output processes, represented for example by computer communication networks. In such cases, regardless of how advanced the programming methodology applied to simulation modeling is, since simulated events are controlled by random numbers, the results produced are nothing more than statistical samples. Therefore, various simulation studies, frequently reported in technical literature, can be regarded as programming exercises only. The authors of such studies, after putting much intellectual effort and time into building simulation models and then writing and running programs, have very little or no interest in a proper analysis of simulation results. It is true that "the purpose of modeling is insight, not numbers" (Hamming, 1962), but proper insight can only be obtained from correctly analyzed numbers. Other modes of presenting results, for example, animation, can be very attractive and useful when the model is validated, but nothing can substitute the need for statistical analysis of simulation output data in studies aimed at performance analysis; see also Schruben (1987).

In the stochastic simulation of, for example, queuing systems "computer runs yield a mass of data, but this mass may turn into a mess." If the random nature of the results is ignored, "instead of an expensive simulation model, a toss of the coin had better be used" (Kleijnen, 1979). Statistical inference is an absolute necessity in any situation when the same (correct) program produces different (but correct) output data from each run.

The bottom line is the following: simulation, like other experiments that deal with random entities, should be planned in advance. This planning can benefit from various approaches. One approach advocated by mathematical statisticians goes under the heading of statistical design of experiments (DOE), which, among other things, is concerned with identification of unbiased estimators, estimates or bounds on variance components, and realization of useful output confidence intervals (Kempthorne, 1952; Cochran and Cox, 1957; Wilks, 1963). Factor designs, either composite or orthogonal designs, Latin squares, and cluster analysis are some of the tools employed in DOE. When properly applied, these tools are claimed to shorten the test runs and/or make them more meaningful. The main application of DOE is to systems where multiple system parameters or statistical factors affect several output (performance) parameters and their interactions must be resolved.

In toll quality networks, blocking events and other malfunctions occur infrequently under normal circumstances. They can be classified as rare events that correspond to some random variables (e.g., number of calls to the same area) assuming rather extreme values. Extreme values, either in the sense of largest or smallest, sample values of typically independent random variables have been studied (Gumbel, 1958; Weinstein, 1973) and applied to communications problems by several investigators (Jeruchim, 1976; Berberana, 1990).

The most useful results appear to be the following. Based on the, so called, "stability postulate," the extreme value theory (EVT) permits only three possible asymptotes to which an extreme value distribution can converge as the number of samples n becomes increasingly large. (The postulate assumes that the appropriate convergence exists as n tends to infinity.) Each asymptote is characterized by very few simple parameters. They can be used in estimation problems, where otherwise either unknown or quite complicated distributions would have to be used. Let

$$F(x) = \text{Pr}[\text{random variable is less than or equal to } x].$$

Then for positive s the three stable distribution types are:

- (i) $F(x) = \exp [\exp (-x)]$ (all x),
- (ii) $F(x) = \exp [-x^{-s}]$ (positive x),
- (iii) $F(x) = \exp [-(-x)^{-s}]$ (negative x).

Corrected page

Literature refers to these as the (i) Gumbel, (ii) Frechet, and (iii) Weibul distributions, respectively. They are also called the classic EVT types, to distinguish from the more recent, and perhaps more applicable, generalized EVT's where x is replaced by some monotonic function of x .

Berberana (1990) reports on the application of the Gumbel form (i) to simulation of buffer overflow in experimental BISDN networks that employ the ATM technology. A total of M samples of the variable in question (i.e., queue size in this buffering application) are partitioned into N groups of n samples each. The maxima of these groups, denoted by X_i ($i=1, \dots, N$), are assumed to have a linear relationship to the random variable x that possesses one of the above asymptotic distributions. Thus, both n and $M=Nn$ must be large enough for the EVT model to hold. If the linearity is expressed as

$$x = a_n X_i + b_n,$$

then it remains to estimate the best values of parameters a_n and b_n . Berberana (1990) uses the maximum likelihood method and obtains the following set of equations for optimal a^* and b^* (the dependence on n being understood),

$$a^* = \mu - A(1)/A(0),$$

$$b^* = -a^* \ln A(0),$$

where μ is the sample mean of X_i and for $k=0,1$ the entities $A(k)$ are defined as

$$A(k) = (1/N) \sum_{i=1}^N X_i^k \exp(-X_i/a^*).$$

Iterative numerical methods have been used to solve these equations. Or, if one approximates,

$$\exp(-X/a^*) = \exp(-\mu/a^*) \cdot [1 - (X-\mu)/a^* + (X-\mu)^2/2a^{*2}],$$

then an explicit solution results:

$$a^* = (\mu_3/4)^{1/3} [(1 + \sqrt{1-2/27\gamma^3})^{1/3} + (1 - \sqrt{1-2/27\gamma^3})^{1/3}],$$
$$b^* = \mu - a^* \ln(1 + \sigma^2/a^{*2}),$$

where σ^2 is the sample variance, μ_3 stands for the third central moment of the sample,

$$\mu_3 = (1/N) \sum_{n=1}^N (X_i - \mu)^3,$$

and γ denotes the coefficient of skewness

$$\gamma = \mu_3/\sigma^3.$$

Both group size n and maxima sample size N appear to be critical factors in the quality of the estimators. In the case studied (Berberana, 1990), values around $n=10,000$ and $N=100$ seem to be adequate, if not optimal.

There is concern and even valid criticism due, when a two-sided (i.e., negative and positive) distribution is forced to fit data that are naturally one-sided, as is the case for nonnegative queue lengths in network buffers. Perhaps, the Frechet distribution (ii) should have been used in the above example. Unfortunately, the maximum likelihood and other parameter estimation methods appear to be far more difficult for both the Frechet and Weibull asymptotic distributions.

To avoid this difficulty, some EVT workers have attempted a direct fit of the gamma distribution (Ratz, 1988). The resultant fit to data points appears to be quite good, while the reduction of simulation runs is considerable.

It its simplest form, the gamma density function for nonnegative random variable x is defined as

$$g(x) = [m^a x^{a-1} / \Gamma(a)] \exp (-mx) ,$$

where both parameters a and m must be larger than zero. If one is given sample values X_i ($i=1, \dots, N$), then there is always a unique maximum likelihood solution a^* and m^* . This solution must satisfy

$$(1/a^*) \exp \text{Psi} (a^*) = \mu_g/\mu_a,$$

$$m^* = a^*/\mu_a,$$

where $\text{Psi} (x)$ is the psi- or digamma function (Abramowitz and Stegun, 1964), while μ_a and μ_g are the arithmetic and geometric means of the sample:

$$\mu_a = (1/N) \sum_{i=1}^N X_i ,$$

$$\mu_g = [\prod_{i=1}^N X_i]^{1/N},$$

For nonnegative sample values, the well-known "arithmetic-geometric" inequality (Hardy et al., 1964) states that μ_g/μ_a cannot be less than zero, nor can it be larger than unity. At the same time, the function $\exp \text{Psi}(x)/x$ is also monotonically increasing over the same range of values. See Figure 8. Thus, whatever the value of μ_g/μ_a , there exists a unique a^* . The optimal m^* follows immediately from a^* .

As an illustrative example, consider an experimentally obtained sample of 25 discrete data points:

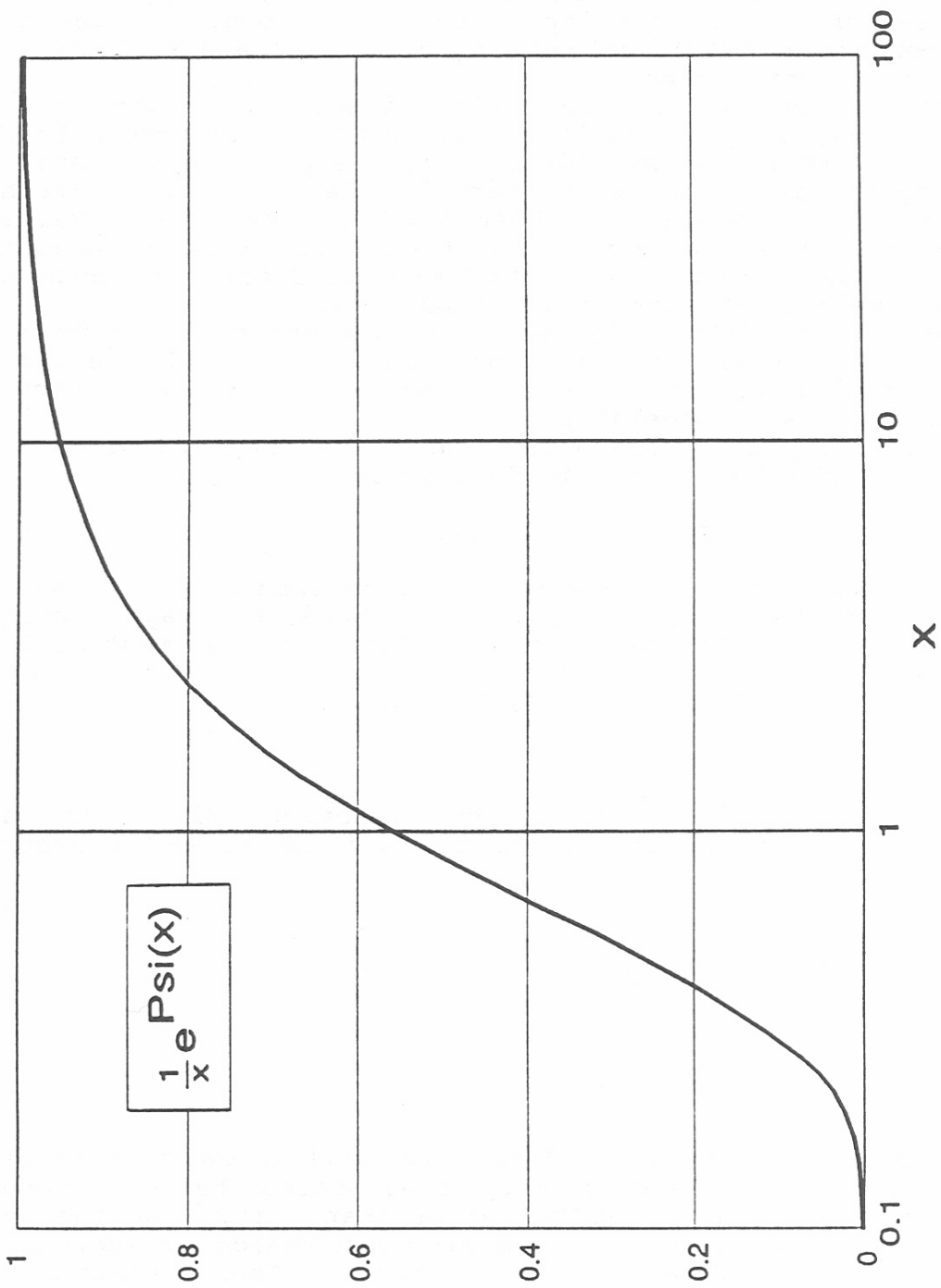


Figure 8. Function $(1/x) \exp \Psi(x)$.

7	13	19	27	40
10	13	21	28	40
12	14	24	32	42
12	15	24	33	47
13	17	25	36	51

The arithmetic mean of this sample is $\mu_a = 24.60$, while the geometric mean is $\mu_g = 21.52$. Their ratio $\mu_g/\mu_a = .875$. With the aid of Figure 8, one obtains the maximum likelihood parameter values

$$a^* = 3.90,$$

$$m^* = 0.158.$$

The corresponding gamma density function is

$$g(x) = 0.000141 X^{2.9} \exp (-0.158X),$$

for nonnegative x .

Figure 9 illustrates the given sample histogram and its gamma density approximation. The main advantage of this approximation is the ability to estimate the low probability, extreme value, tails of the distribution. The validity of these estimates depends on exponential character and stability conditions associated with EVT (Ratz, 1988).

4. CIRCUIT SWITCHED NETWORKS

4.1 Planning Factors for Modeling and Simulation

Planning for simulation starts with a parameter definition and numerical specification of the five categories of network factors identified earlier in section 1.3:

- (1) Terrestrial object network.
- (2) Satellite support network.
- (3) Normal offered traffic.
- (4) Stress scenario.
- (5) Performance parameters.

Note that this short list intentionally avoids a separate item called "network functions." The functions in question include a spectrum of operational, management, control, etc., procedures, rules, and algorithms. Generally they are associated with network software. Together with network hardware, functions and their implementations are assumed to belong under the above items (1) or (2), or both.

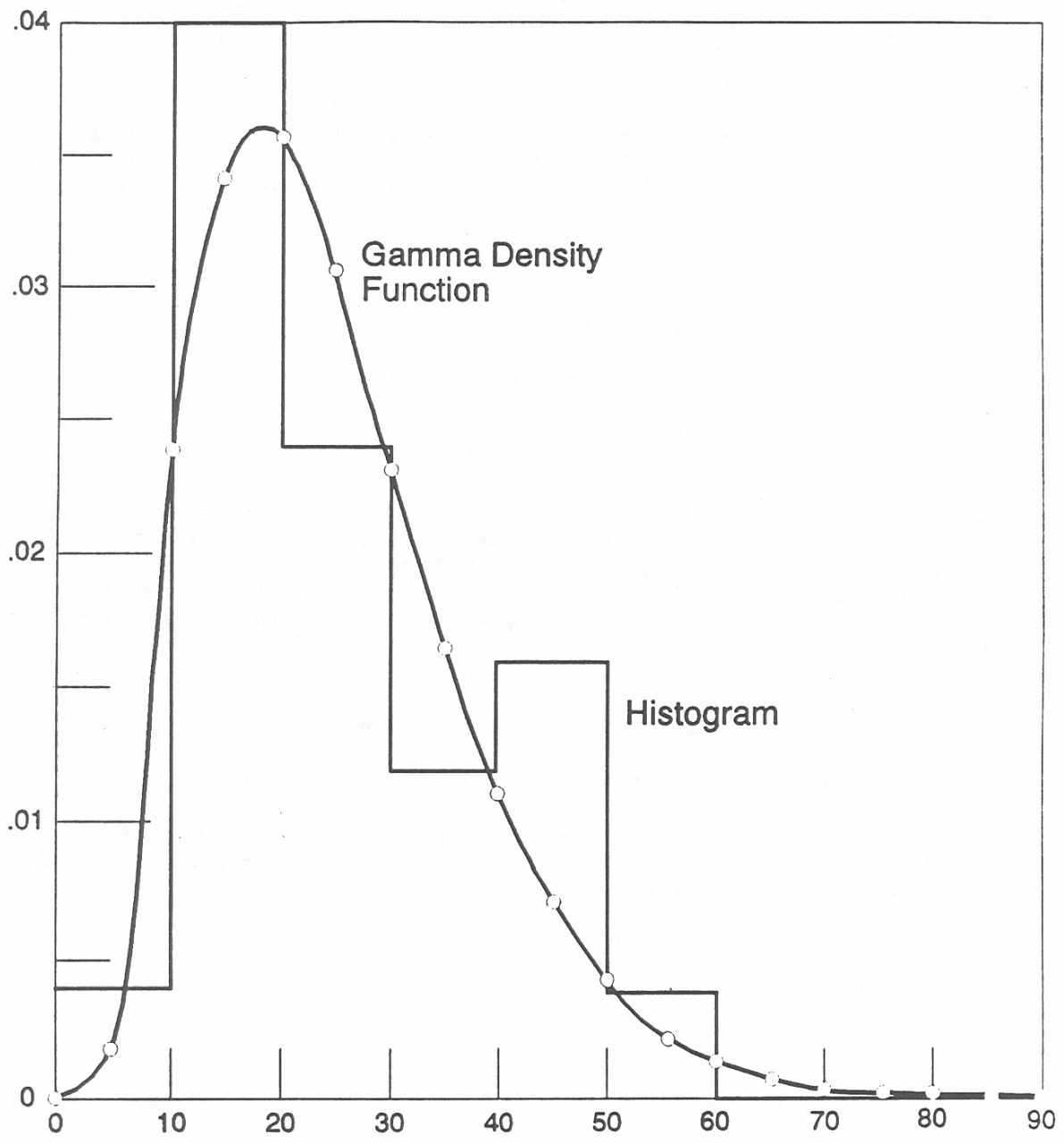


Figure 9. Maximum likelihood gamma density fit to a given histogram.

Each of these categories requires the specification of several dozen descriptors (parameters). Examine, for example, the circuit-switched telephone network. Table 5 shows a possible specification of the circuit-switched network model. Note that the length of such a specification table grows appreciably as more and more parametric details are included.

Assuming that all the necessary ingredient values are specified for the model, one big question remains. That question is the choice and implementation of the simulator and its support facility. Due to the present market availability of many fast computer units, software, and database systems that are claimed to be appropriate for simulation applications, the problem seems to be the selection of the most suitable hardware-software products. Or, if a suitable product cannot be found on the market, then certain custom designed tools would have to be generated from scratch.

A crucial criterion for the suitability of a simulation system is its speed. A system that is too slow will be useless for large networks, no matter how accurate its basic underlying mode or how inexpensive its procurement. To this end, shortcuts based on experience or theory, aggregation of events (either by region, by time, or by service category), and various expedient approximations could be useful tactics in simulator implementation.

4.2 Available Simulation Tools

The first of these, COMNET II.5, is based on SIMSCRIPT II.5 language and SIMGRAPHICS visuals. It offers a family of simulation tools. Besides COMNET II.5, there is a CACI system called NETWORK II.5. It appears that NETWORK II.5 is almost exclusively tailored to local area network (LAN) modeling needs and not those of nationwide PSN's. COMNET II.5, however, is a call-by-call simulator applicable to circuit-switched networks. The speed of COMNET II.5 remains a question. Even in terms of rough estimates, it is hard to tell how fast COMNET II.5 would simulate 100-to-1000 node circuit-switched networks in a typical work station or main-frame environment.

General Simulation System (GSS) is a communication modeling and simulation (analysis and design) tool developed by the Prediction Systems, Inc. (PSI) of Manasquan, NJ. So far GSS has found applications to mobile and tactical communications networks.

Simulation product called OPNET is applicable to parallel computer networks, mobile networks, LAN's, WAN's, and global networks. Most of the reported applications of OPNET deal with protocol evaluations for packet switched networks. It is not entirely clear how fast and effective OPNET would be for simulation of large circuit-switched networks.

Block Oriented Network Simulator (BONeS) was originally developed at the University of Kansas, Lawrence, KS (Shanmugan et al., 1989). Together with other simulation products, BONeS provides a graphical simulation environment through the use of model building blocks. These blocks specify the network topology, traffic, protocols, and data structures. They apply to a wide

Table 5. Specification Parameters for a Circuit Switched Network Model

No.	Category	Parameters
1	Terrestrial Object Network	Number of nodes Number of physical links Link capacities Physical/facility connectivity Logical (DACS/ORIN) connectivity Cross-connect capacities and speeds Switching capabilities Node and link delays Routing algorithms Relevant network controls
2	Satellite Support Network	Geo-stationary and/or low orbit deployment Number of satellites Orbit effect: delay and availability Number of transponders Capacity: data rate, bandwidth Signal switch rates Beam types: fixed, spot, scan Beam switching rates Earth stations: number and types Gateways: number and types Number of control centers Network reconfiguration controls Routing algorithms
3	Normal Offered Traffic	Definition of busy hour Engineering load Voice and data traffic mix Number of users: finite or infinite Priority classes Interarrival distributions Holding time distributions Erlang B, C, or other models for blocked service requests
4	Stress Scenario	Disabled or damaged nodes Disabled or damaged links Logical connectivity outages Control center outages Geographic extent of traffic overload Duration and time profile of traffic overload

Table 5. continued

No.	Category	Parameters
5	Performance Parameters	Blocking GOS: global, local, for specified user classes Service unavailability: global, local, for specified classes Percentage of carried traffic Over- and underutilization of switches, control centers, STP's, data bases, links, cross-connects, routes Service delays Probability of interruption Probability of message loss Propagation delays Echoes, distortions, SNR, BER

range of networks, including packet-switched networks, circuit-switched networks, LAN's, and WAN's.

Another interactive, discrete-event simulator for telephone networks, called GENSIM, (Mathis, 1989) has been used as a prototype simulation tool with limited objectives. GENSIM is intended to demonstrate the utility of real-time expert systems for network monitoring and control.

Sim++ is an object-oriented development toolkit based on C++. The U.S. Naval Research Laboratories are using Sim++ as part of their communications protocol and message routing evaluation program.

The performance of packet switched networks, which represent a significant part of all data and computer networks, is typically characterized by the dynamics of interacting queues (as in queueing networks). Many models and simulation tools for such queueing networks have been developed here and abroad (O'Reilly and Hammond, 1984; Spirn et al., 1984; Chiarawongse et al., 1988). Relevant methods and simulation tools will find application in future work (Kurose and Mouftah, 1988; Frost et al., 1988; Cassandras and Strickland, 1988).

4.3 Selection of Preferred Method

For circuit switched networks, as for other networks, one can proceed to implement a simulation capability in two quite distinct ways.

The first method turns to software vendors and acquires already existing simulation tools. The versatility of these system tools then dictates, not only the simulation process itself, but also the modeling options available to the system analyst. As an example, one may select the OPNET simulation tools from the candidates listed in the previous section. The menu of OPNET programs then will determine the menu of realistic models open to simulation.

The second basic approach is to define one's own network models, or a reasonable parametric range of the models, and to implement simulation tools custom tailored to the models proposed. This method has the potential of better satisfying individual simulation objectives and requirements. Unfortunately, it also entails considerable time, manpower, and costs in the design, development, and debugging of software.

Many planning factors are important. However, if a short list is desired, the following factors appear to be among the most important:

1. Network size: Number of nodes (switches), links, user terminals.
2. Network complexity: Characteristics of facilities, their functions, protocols, speeds, capacities (data rates or bandwidths).

3. Traffic: Volumes, percentage profiles (voice, data, etc.), message handling (circuit switched, packet switched).
4. Stress: Facility damage, traffic overloads, recovery.
5. Performance measures: Selection of parameters, confidence objectives.
6. Simulation system: Computer speed, memory (data bases), software (operating system, languages).
7. Simulation job: Division and sequencing of tasks.
8. Event aggregation: To shorten simulation runs, if needed, by clumping events either according to traffic type, geography, facilities, or in time.
9. Statistical analysis of output data.
10. Postmortem: Identification of biases (introduced by model used or by event aggregation) and their correction.

4.4 Temporal Aggregation: An Example

This section presents a simulation example that appears to be useful in selected applications. Assume that the network, like the PSN, carries mostly circuit-switched telephone traffic. For motivation of the model, assume that the network is so large and the traffic consists of so many service (call) events that a faithful event-by-event simulation is far too complex (i.e., it simply takes too long). Given that one prefers not to ignore any call events, let us consider a model that aggregates events or lumps them together "in time."

Perhaps the simplest lumping method samples at regularly spaced points in real (network) time. (The simulator that processes one sample point after another, however, need not run according to fixed simulator clock periods. It can proceed to the next sample point at any time after finishing with the previous sample.) Other "in time" lumping methods can vary the interval length between samples. One reason for this could be the objective for a certain number (minimum to maximum) of events to constitute a lump. The variable-interval methods seem to be more complex, as they require extensive time management and associated statistical controls. In what follows, the periodic sampling point scheme is assumed.

A heuristic justification for periodic or discrete representation of offered and carried calls can be found in the early work on telephone traffic. Periodic scanning and recording at switches has been less costly than continuous observation. As a result, telephone traffic measurements were usually carried out

by averaging sequences of instantaneous periodic observations of the switch status (e.g., number of calls present, etc.), rather than by continuous time averaging. It has been shown (Benes, 1957) that with such sampling, essentially no statistical information is lost for the purpose of estimating the following four telephone traffic statistics at a service node:

- * The number of calls that exist at the start of the n-th observation interval, $E(n)$.
- * The number of calls that arrive during the n-th observation interval, $A(n)$.
- * The number of calls that terminate (i.e., hang up) during the n-th interval, $T(n)$.
- * The average number of calls that last throughout the interval, $L(n)$.

These four basic statistics appear sufficient to model and to simulate circuit-switched network performance. For instance, the number of blocked calls, $B(n)$, during the n-th interval can be estimated from the conservation of "lost versus gained calls,"

$$E(n) = L(n) + T(n),$$

$$E(n+1) = L(n) + A(n) - B(n),$$

$$B(n) = [A(n) - T(n)] - [E(n+1) - E(n)],$$

which is clearly valid for sufficiently short observation intervals.

Previous experimental studies have established that under various conditions call durations are exponentially distributed. Also, as far as newly arriving (not overflow or network rerouted) traffic is concerned, the interarrival times are also exponential random variables. The mean of the first variable is known as the holding time, while the reciprocal of the second is called the calling rate.

As a consequence of the exponential property, the number of offered (or arriving) calls in a fixed interval must have a Poisson distribution. Likewise, another Poisson must be the distribution for the number of departing (terminating or hanging-up) calls.

A random variable X is said to have a Poisson distribution (Kleinrock, 1975), if

$$\Pr(X=k) = (N\lambda t)^k \exp[-N\lambda t]/k!,$$

where

- N = Number of idle users at a service facility,
- λ = Calling rate for an individual user,
- t = Duration of the sampling interval.

The mean and variance are the same for the Poisson distribution, that is,

$$E(X) = \text{var } X = N\lambda t.$$

Useful additive properties follow from this. For instance, if one service region with N users is joined to another region with M equivalent users, then the combined region also has Poisson traffic arrivals with mean and variance equal to $(N+M)\lambda t$.

Introductory Simulator Outline

Figure 10 shows a general outline for time-clumped simulation of a network. (A more detailed simulator representation will follow after this outline.) The network can be terrestrial, satellite, or a combination of the two. However, it is assumed to carry voice or similar circuit-switched traffic.

The human operator interfaces with this simulator at three functional levels that are called Inputs, Initialization, and Outputs in the figure.

At the Input level, the operator defines the target network, the offered traffic, and the stress scenario. The latter may include physical or logical damage to the network, as well as traffic surges and overloads.

At the Initialization level, the operator sets the time interval for traffic sampling, the initial traffic status, the desired output parameters, and other run controls (such as diagnostic snapshots and exits). The initial traffic setting could be all-zero (i.e., an empty network), but this would warrant a considerable run time to build up to a steady state. It is therefore advisable to start either with a previously generated and stored traffic state or with some estimate of what a reasonable steady-state loading may look like.

The output parameter to be studied in this example is the blocking or lost calls grade-of-service (GOS). Therefore, at the Output level, simulator data (such as counts of attempted, blocked, and completed calls) are collected, stored, and processed both for statistical significance and for various reports, graphs, etc.

The actual simulation process takes place in the, so called, simulation engine of Figure 10. Once entered, the engine periodically cycles through three routines:

- I. The statistical add-calls routine.
- II. The event counters for output GOS parameters.
- III. The statistical delete-calls routine.

The period of one cycle corresponds to the sampling interval for the network traffic. The cycles repeat as

I - II - III - I - II - III - I - II - III - . . .

until the simulation run is completed, or a pre-assigned program interrupt is reached.

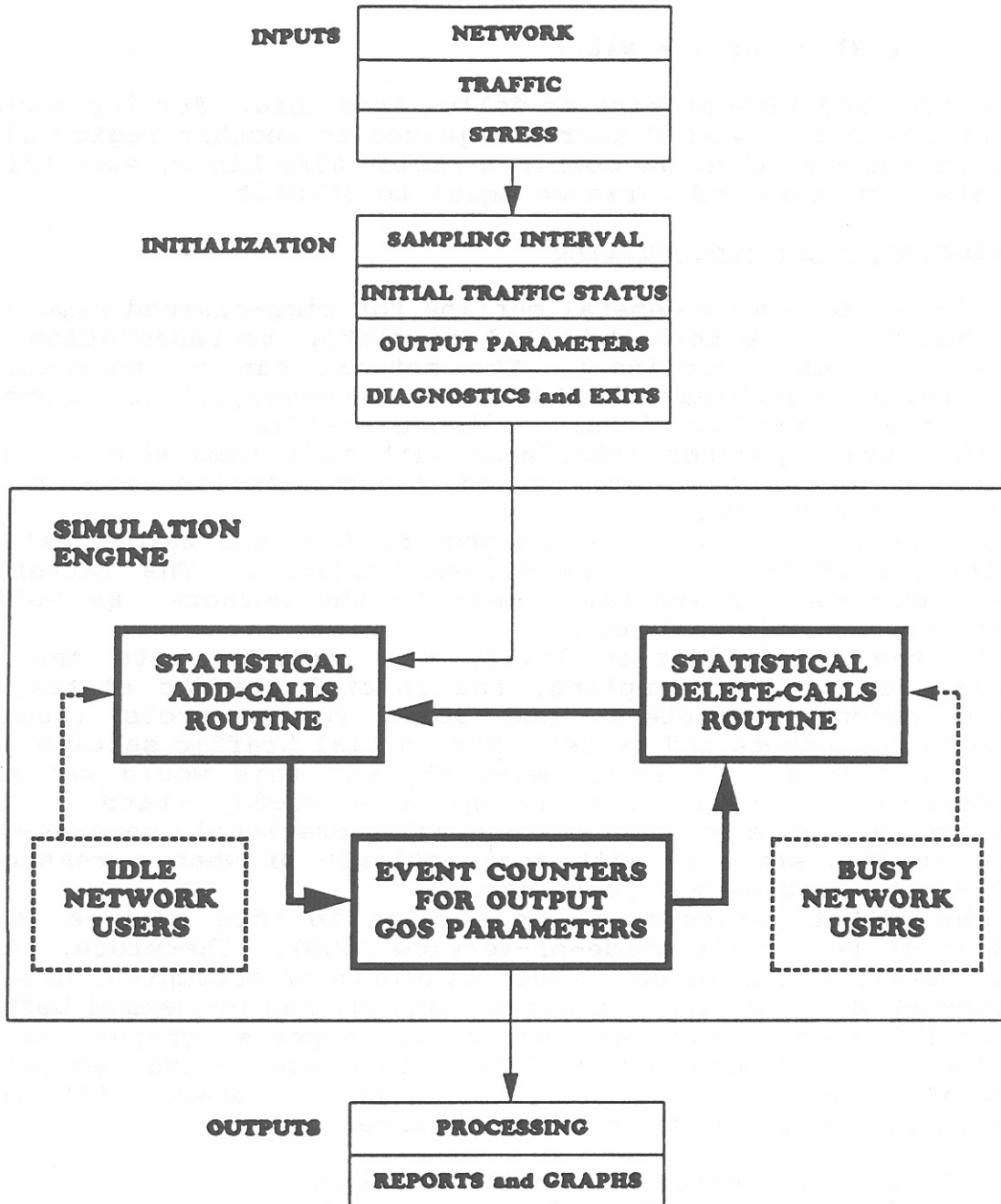


Figure 10. Functional outline for time-aggregated, blocking GOS simulation of a circuit switched network.

In the statistical add-calls routine, a number of new calls are originated for the network and for all network service nodes. These offered calls come exclusively from the set of idle users. Random number generators create the add-call numbers according to the Poisson distribution, where N represents the number of idle users in the appropriately defined service domain, such as any subnetwork or the entire network. More about distribution-dependent random number generation and source-to-destination call assignments for clumped traffic will be said later.

The statistical delete-calls routine performs the complementary or inverse function of the add-calls routine. It deletes the number of calls that terminate (i.e., hang-up) in the assumed sampling interval. The calls to be deleted come exclusively from the set of busy users. Whether at individual switches or in the entire network, the number of delete-calls are random variables with different Poisson distributions. More about delete-call details will follow.

The event counters identify specific performance events and maintain accumulated records of their numbers. In this GOS example, the counts must reflect the number of total call attempts and the number of calls blocked by the network (not by a busy called party). Depending on system applications, the counts may be further classified according to different criteria, such as node-to-node, region-to-region, over the entire network, by service or user class blocking, or for specified time intervals (e.g., for each 1-minute interval of the busy hour).

Function Diagram for the Simulator

To show the proposed simulator functions in more detail, this section presents a series of nine diagrams that, all together, represent the functional diagram of the simulator.

Input: Network

Network specification function is the starting point, as illustrated in Figure 11. The set of nodes, SN , is defined first. The number of nodes is denoted by N .

Next, one defines the set of links or SL . The number of links, L , deserves a special comment. Depending, whether one deals with physical spans or trunk groups between switches, or logical links created by cross-connect devices, L may have different interpretations and different values. For mathematical and topological aspects in simulation, the logical connectivity is more intrinsic. However, for facility damage or other stress events, the physical connectivity is the natural medium. To accommodate both points of view, Figure 11 shows parallel roles for the physical and logical connectivities. There is also a direct interaction (or translation) arrow between the two, as changes in one may warrant changes in the other.

Another parallelism binds together the two topological network representations: the connectivity matrix C and the distance matrix D . Matrix C is more familiar and more used than matrix D . However, D can have efficiency and speed advantages for certain

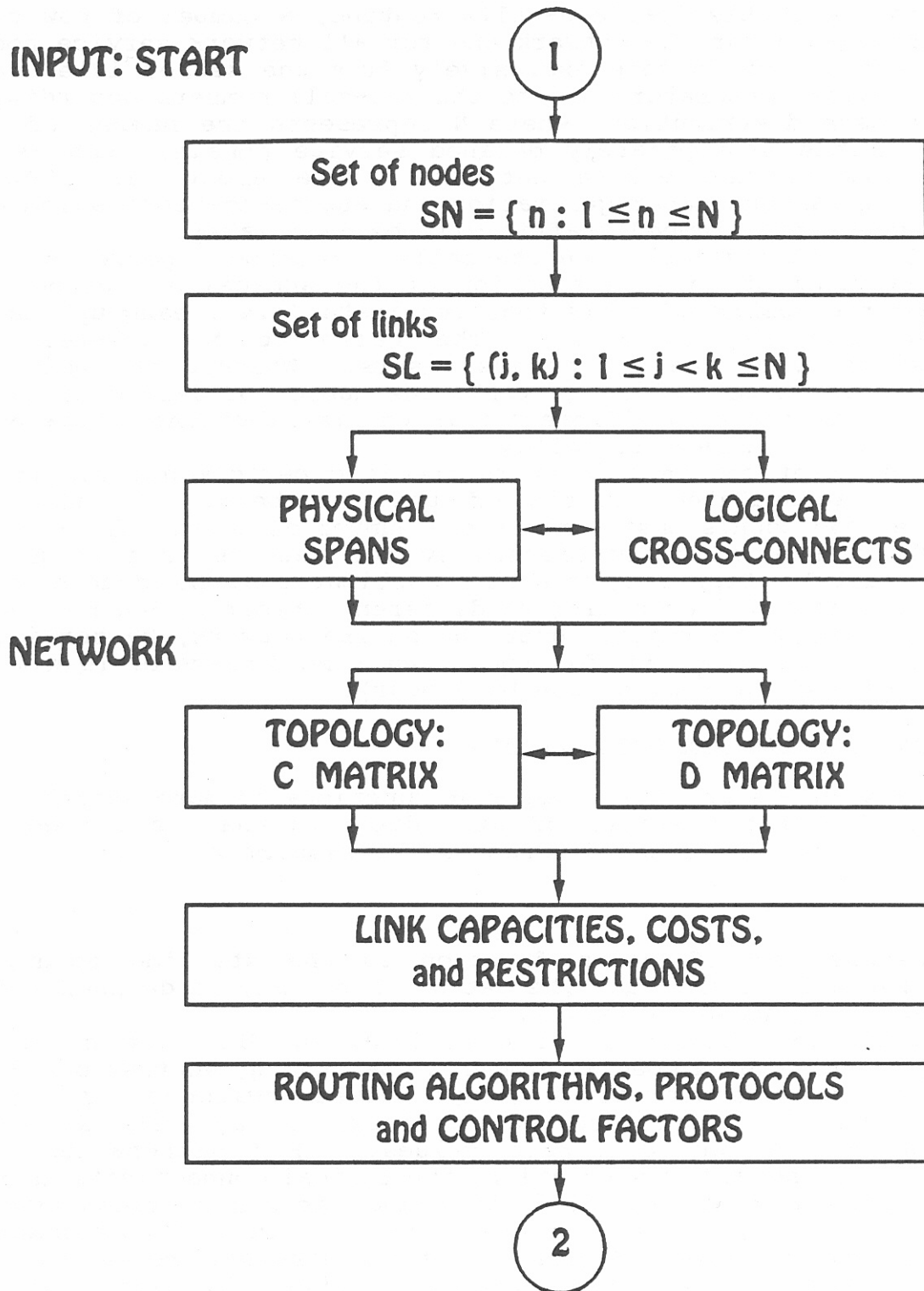


Figure 11. Input: network specification.

routing protocols. One can always translate C into D and vice versa. The step from D to C is trivially easy for all network sizes. The step from C to D becomes more difficult as the number of nodes, N , increases.

Link capacities, in terms of voice channels, data rates (for digital voice), or bandwidth, are also specified as part of the network input functions. This should be accompanied by link cost numbers, especially for applications where least-cost routing is considered. Restrictions on link use may be also be included here. They may list traffic categories that cannot tolerate echoes or path delays on satellite links, or that must avoid exposed circuits for security reasons. Or the restrictions may identify one-way facilities, such as incoming-only or outgoing-only trunks at designated nodes.

Finally, routing algorithms, relevant protocols, control functions, and perhaps even certain network-management rules must be included in the initial network definition. Emphasis here is on automated algorithms intended for operation under both normal and stress conditions.

Input: Traffic

Traffic specification is the second input task. The main voice or telephony traffic elements are outlined in Figure 12.

The list starts with the specification of the offered traffic. Since the traffic volume is proportional to the number of idle individual users, the proportionality factor (namely, the previously introduced calling or arrival rate, λ) must be numerically defined. This factor can depend significantly on local user communities and stress conditions. The rate, λ , may therefore be different at individual originating switches. On the other hand, homogeneous user communities can be represented by the same constant arrival rate over the entire network.

Note also that, if the total number of users is considerably larger than the number of possible instantaneous servers, then the infinite user model is applicable. Instead of the individual calling rate, an aggregate calling rate can then be used.

The previously introduced model for the Poisson distribution represents, with minor modifications, the number of arrivals in any fixed length interval for both finite and infinite user models.

The service rate is specified for the allowed user classes. In the typical case of a homogeneous telephone user population, a single service rate applies everywhere. The service rate, commonly written as μ , is the inverse of the average service (or holding) time. In most places and at most times, the average holding time is between 2 and 4 minutes. If a single number is desired, the 3-minute calls may be taken as a reasonable average.

In the time-sampled or aggregated traffic situation, the number of calls added per sampling interval is a random variable. For a uniform population of telephone users, as mentioned earlier, the Poisson distribution specifies the number of new "add-calls" that must be offered to the network at each access switch, subject to local idle user terminals.

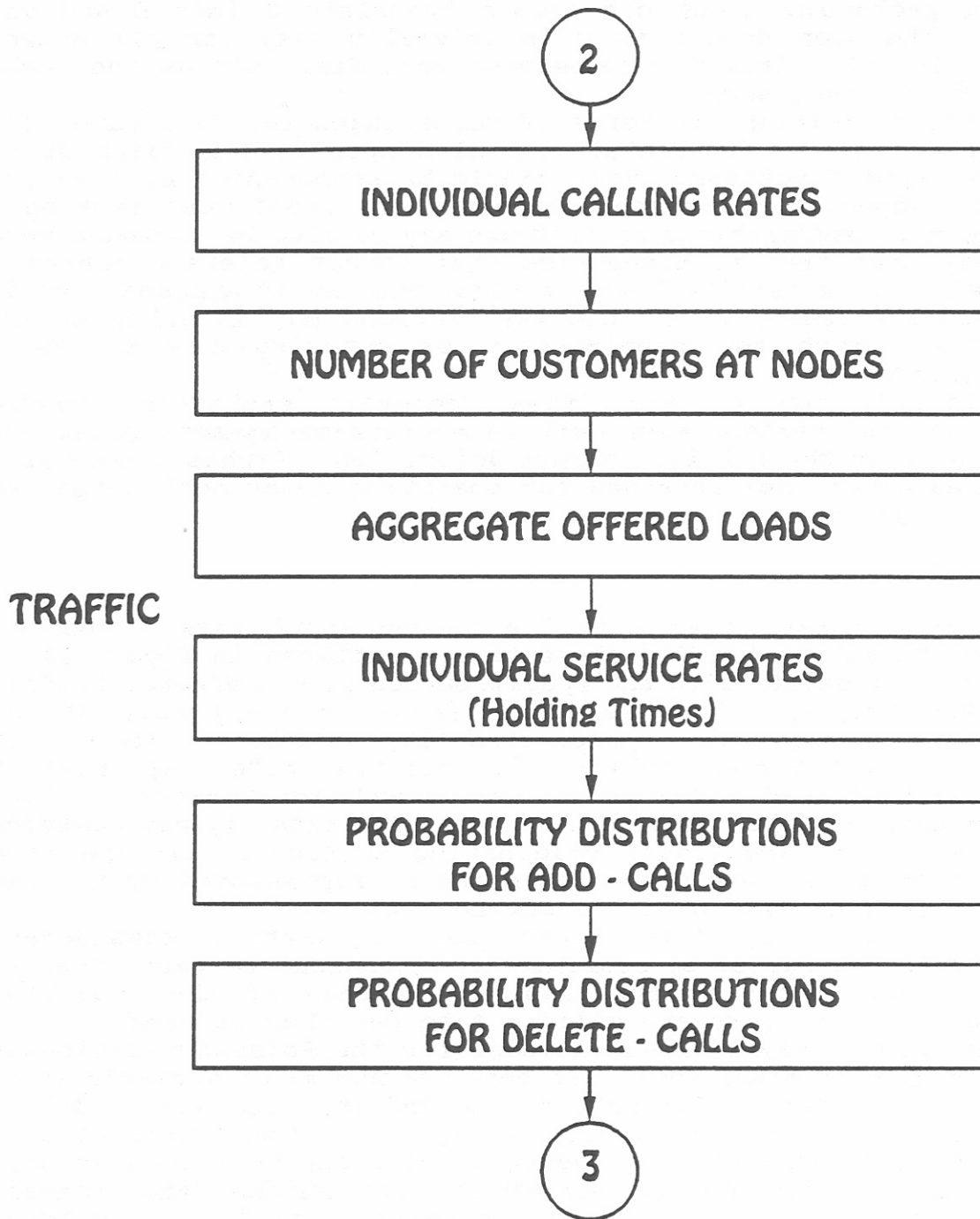


Figure 12. Input: traffic specification.

The number of calls to be deleted in a sampling interval, call them the "delete-calls," are also Poisson distributed variables. However, the delete-calls can only be drawn from the population of busy users.

Input: Stress

The third category of inputs is associated with network and traffic stress. Figure 13 illustrates the stress specification functions. These can be divided into two categories of stress: network stress and traffic stress.

Network stress or damage is specified by listing both totally and partially disabled physical facilities, such as nodes and links. These facility lists correspond to lists of reduced network capacity, modified topology, and various disabled network functions. Their effect is to redefine the set of network nodes, SN, the set of links, SL, and other pertinent network parameters in Figure 11.

Traffic stress can be defined as offered traffic overload that occurs in specified network regions (at certain switches), with specified volumes of traffic and for specified durations of time. One particular effect of traffic stress is to redefine the arrival rate λ and, perhaps to a lesser extent, the service rate μ . When dealing with crises, it seems reasonable to permit network and traffic stress events to happen more or less simultaneously.

Initialization

After the completion of input functions, Figure 10 specifies initialization as the next and final step towards simulation. Initialization includes any final adjustments needed to initiate the actual simulator run (or runs). Thus initialization consists of a set of several, more or less standard, procedures. Figure 14 lists four initialization steps.

The first task listed asks for the selection of the traffic sampling interval. The length of the interval could be kept constant and not varied for different simulation runs or for different network scenarios. But that may imply undesirable consequences for the number of call attempts (called "add-calls" in what follows) or terminated calls (called "delete-calls"). In a long time-period of statistical equilibrium, the number of add-calls and delete-calls are approximately the same for a closed service network. If the fixed sampling interval is too short, then only a few call events will occur. This will reduce the efficiency advantage of the aggregation (time-clumped) method over the discrete event-by-event simulator. On the other hand, excessively long sampling intervals will introduce a clumping error whose magnitude may be hard to estimate.

A solution to this question may be gotten from the previously introduced Poisson distribution that applies to both add-calls and delete-calls. A relatively narrow distribution, for instance one with the ratio for $\text{var } X/E(X)^2$ between 1/100 and 1/10, would imply a lower and an upper bound on the sampling time t through the inequalities

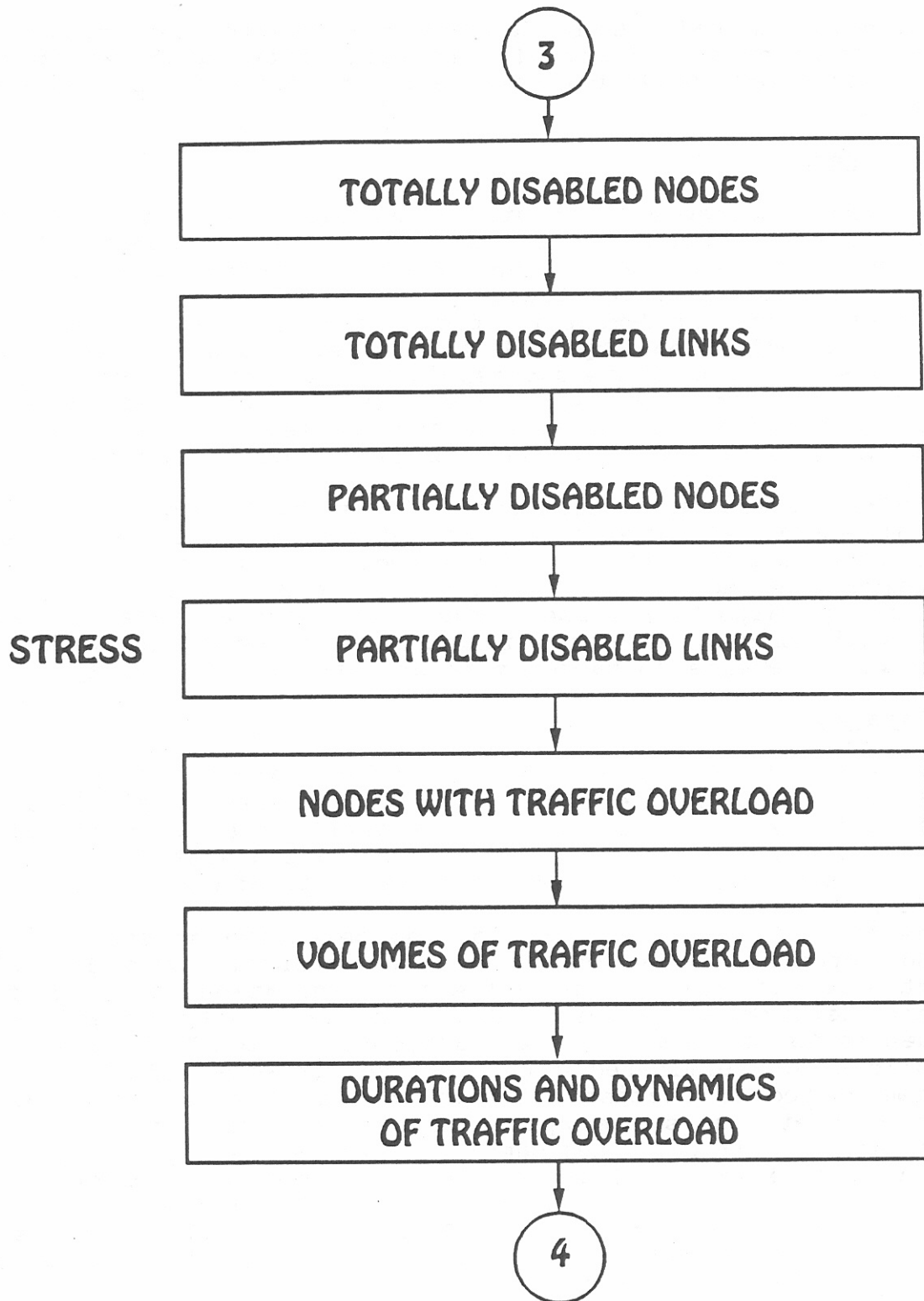


Figure 13. Input: stress specification.

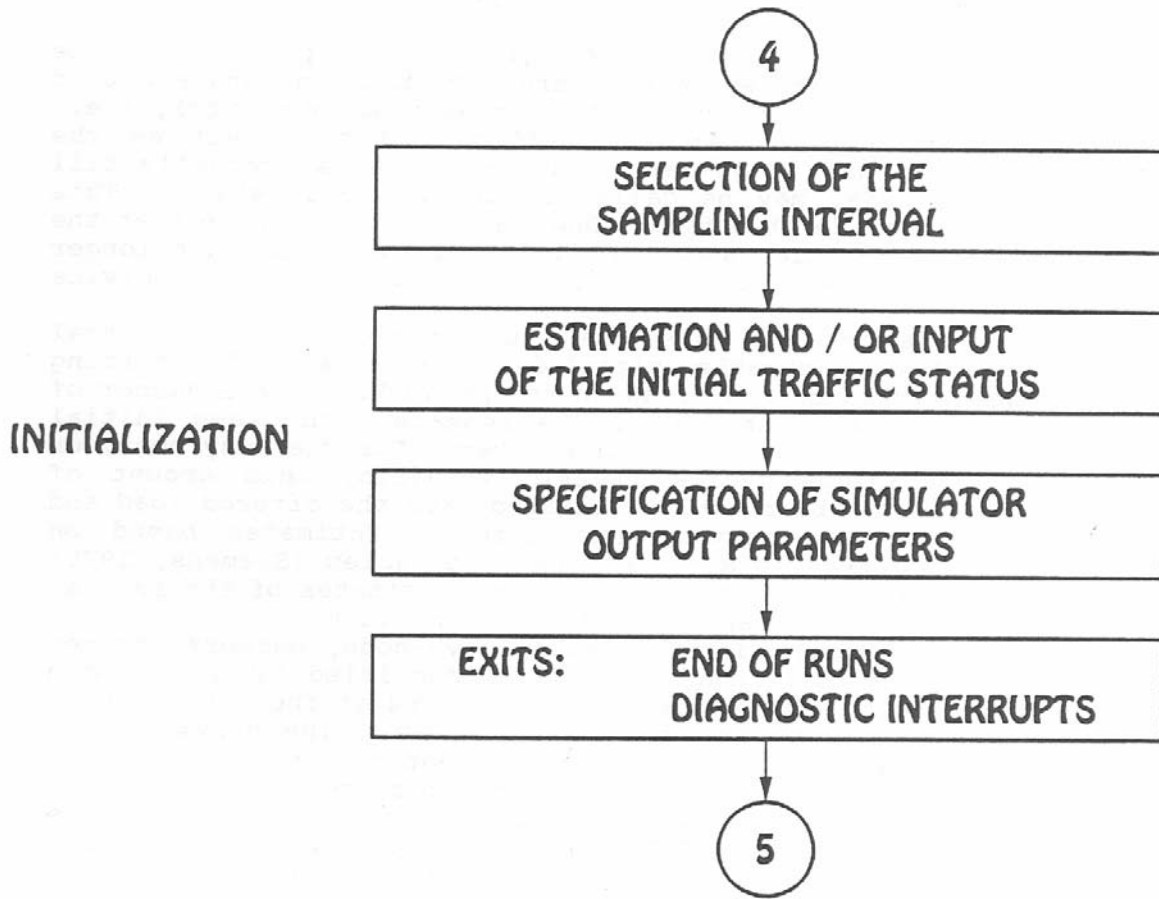


Figure 14. Initialization of simulation runs.

$$10 < N\lambda t < 100.$$

Consider the following simple example. Let the network have $N=100$ nodes (or switches). If each switch has 10,000 users, and if a user on the average makes two calls per busy hour, then one estimates $N\lambda=5.56$ offered calls per second, per switch. It follows that the simulation sampling time t should be between the bounds $2 < t < 20$ seconds. A value of $t = 10$ seconds would appear to be reasonable to start.

The second initialization item in Figure 14 pertains to the setting of initial carried network traffic status at the start of simulation runs. One can begin with a totally empty network, i.e., one with absolutely no carried traffic, and then wait as the statistical streams of add-calls and delete-calls gradually fill the network to what may be called a stable steady state. This buildup takes considerable time. One can roughly estimate that the carried traffic transients should die out in a period much longer than the sum of the typical longest interarrival plus service holding times.

Unless one is interested in the dynamics of the initial buildup itself, considerable initial delay can be saved by starting with an approximate carried load on the network. When a number of identical or similar runs are to be repeated, the same initial traffic status can be used for all of them. For the busy hour, or for any to-be-assumed voice traffic scenario, this amount of carried traffic can be roughly estimated from the offered load and the capacity of the service facilities. Estimates based on standard, Erlang type, traffic engineering tables (Siemens, 1970) can be most accurate. However, even rough estimates of the initial carried traffic status can be almost as beneficial.

As an example consider the small, five-node, network assumed in Table 6. The individual nodes are identified by n , where n ranges from 1 to 5. The total number of users at the n -th node is given by $U(n)$. It varies from 500 to 2,000. The users at the different nodes are assumed to have different calling rates, $\lambda(n)$, and service rates, $\mu(n)$. Both of these rates are given in events per minute. Thus, $\lambda(1) = 0.033$ corresponds to two call attempts per hour, while $\mu(5) = 0.333$ implies a holding time of three minutes. Utilization of the n -th node is defined as

$$\rho(n) = U(n) \lambda(n) / \mu(n).$$

It has a useful engineering interpretation for the steady-state traffic. Given that the lost (blocked) traffic is negligible, or alternatively that the number of servers is infinite, $\rho(n)$ is the estimate of the average number of users being served in the steady state. For node n , it shows the switch utilization by its $U(n)$ users. At any given time, the number of actual busy users, $x(n)$, must be an integer. The last column of Table 6 lists the rounded values for $x(n)$. Note that the final number, namely $x(5)$, is rounded upward. This is done to make the sum of $x(n)$'s even, which is necessary for two-party call realization on a circuit-switched network.

Table 6. Example of Traffic Estimation on a Small, Five-Node Network

Node n	Users $U(n)$	Calling Rate $\lambda(n)^*$	Service Rate $\mu(n)^*$	Utilization $\rho(n)$	Busy Users $x(n)$
1	2,000	.033	.255	258.8	259
2	1,000	.030	.320	93.8	94
3	1,000	.036	.275	130.9	131
4	1,000	.029	.410	70.7	71
5	500	.040	.333	60.1	61

*In units of events per minute.

During the initialization process, it may be necessary to assign node(j)-to-node(k) addresses to all calls. These are called the (j,k) calls. Table 7 continues the example started in Table 6, by illustrating one of many possible partitions of the set x(n) into a number of point-to-point calls x(j,k). Observe that the (j,k) matrix is symmetric and the sums of columns, same as the sums of rows, must equal the number of busy users:

$$\sum_k x(n,k) = \sum_j x(j,n) = x(n).$$

From a practical point of view, a very essential part of initialization is the specification of output parameters. Blocking grade of service (GOS) is the most often used output performance parameter, but it is not the only one. Blocking GOS is a ratio of the number of calls blocked by the network to the number of total offered calls. Both numbers may be statistical averages estimated for an agreed observation period, such as the busy hour. As indicated earlier in Table 5, the GOS statistic can be computed over the entire network. Or it can be applied to specific regions, subnetworks, or even individual switches in the network. If different service classes are allowed on the network for priority or other reasons, then the GOS statistics for individual classes may be important simulation output objectives. If one suspects that certain network facilities are either over- or under-utilized, then the traffic served by these facilities may have to be ascertained.

Performance parameters not directly involved in particular simulation runs can and should be ignored for practical reasons. However, if need should arise to compute them afterwards, then sufficient simulation data should be stored to permit their evaluation in the future. The definition of output parameters must be suitable either for immediate (quick-look) presentation during the simulation run or for delayed computation after the run.

The final initialization function specifies the conditions for stops and exits from an individual run or from a given series of runs. While the final exit defines the end of simulation, many other halts are associated with diagnostics and corrections necessary during the ongoing simulation process. Depending on the nature of the simulation software and the complexity of the network, many kinds of diagnostic stops and readouts can be expected, especially so in the initial test runs of the simulator.

Simulation Engine: The Add-Calls Routine

A key element of the simulation engine, see Figure 10, is the statistical add-calls routine, or ACR for short. The purpose of ACR is to add calls for each sampling interval. The number of calls added is a random variable that depends on the number of idle users and their calling rate.

Figure 15 shows an abbreviated diagram for the ACR. It is based on the assumption that the network is connected, with no isolated or separated subnetworks. The network is assumed to have

Table 7. Example of Node-to-Node Calls on the Five-Node Network

j \ k	1	2	3	4	5
1	--	53	90	65	51
2	53	--	34	2	5
3	90	34	--	3	4
4	65	2	3	--	1
5	51	5	4	1	--
Sums:	259	94	131	71	61

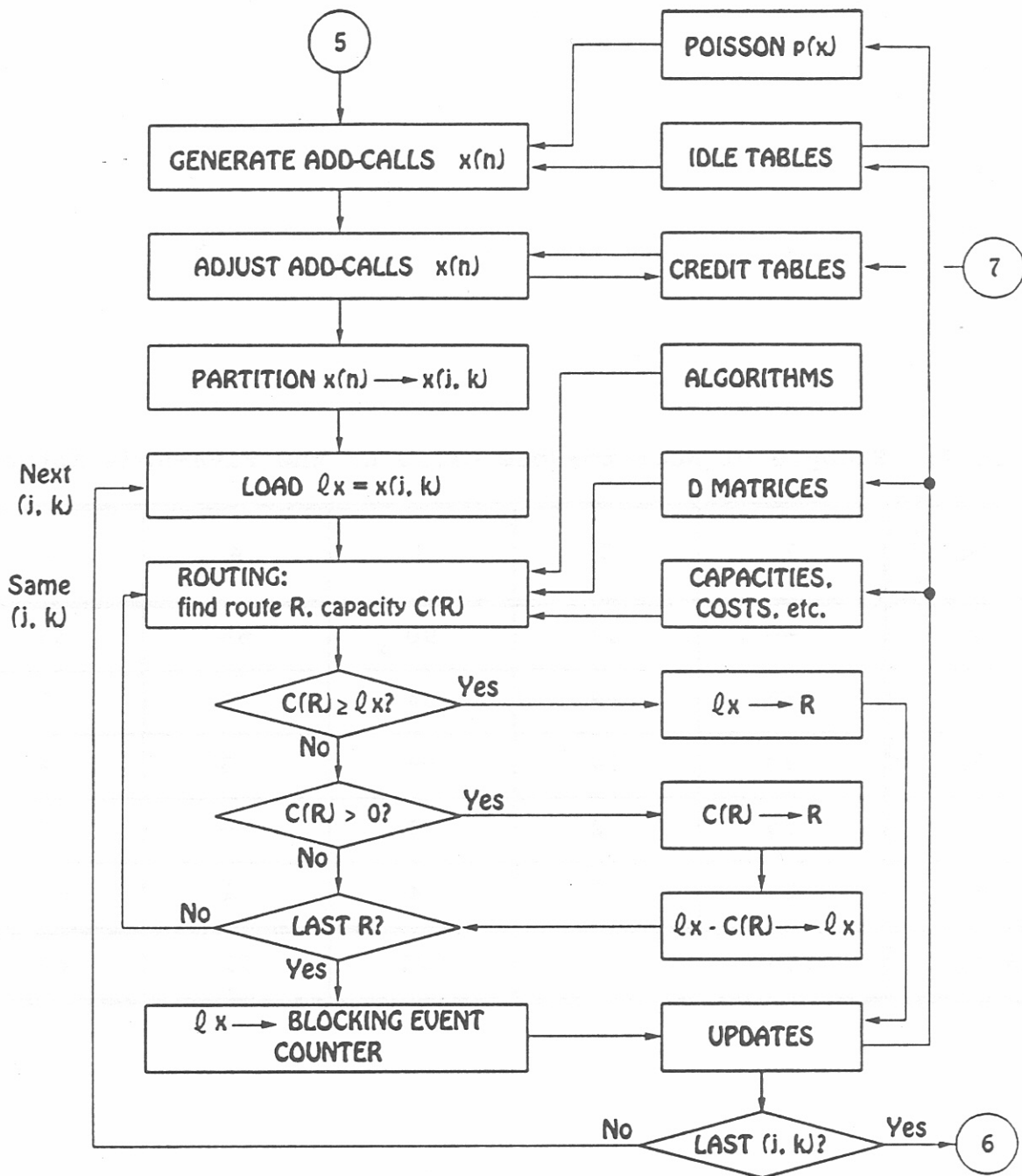


Figure 15. Add-calls routine (ACR) for a connected network.

N nodes. Index $n = 1, 2, \dots, N$ is used to identify nodes or switches.

The first step is to generate the number of calls, $x(n)$, offered per sampling interval at node n . According to the convention assumed here, both locally generated (outgoing) and remotely generated (incoming) calls are included in $x(n)$. This random number generator uses the Poisson probability distribution applied to the number of idle users at node n . Separate functional elements, called the Poisson $p(x)$ and Idle tables, are used for this purpose. If the number of users at a given node is very large, such as in excess of 10,000, and the individual calling rate is very low, one is justified in ignoring the user tables at this node by assuming an infinite user model. The Poisson model still applies to the infinite-user scenario.

Given a set of $x(n)$ at all nodes n , three adjustments are made next.

- (i) First, since every call involves two parties, the sum of all $x(n)$ must be even. If the sum happens to be odd, an extra call origination may be added to any of the N nodes.
- (ii) Second, every caller must find an opposite party at the other end. That means that the largest of the $x(n)$'s should not exceed the sum of the remaining $x(n)$'s. If $N=2$, then $x(1)$ must equal $x(2)$ and $x(1,2) = x(1)$ must hold.
- (iii) The third and final adjustment pertains to the Credit tables. It is a relatively minor issue for most telephone networks, as it corrects for a typically small fraction of previously "should have been deleted, but were not deleted" calls. The detailed reasons for this, the third, adjustment will become clear in the context of the delete-calls routine.

Next, the set of $x(n)$'s is randomly partitioned into end-to-end or node-to-node calls. If the two end nodes for a call are nodes j and k , then this call is said to be a (j,k) call. For networks with Common Channel Signaling (CCS) it does not matter which user originates a call. Similarly, the order of (j,k) or (k,j) is immaterial for node-to-node blocking GOS statistics. In what follows it is assumed that $1 \leq j < k \leq N$. There seem to be a number of partitioning methods possible. The question of which method is the fastest and most random is not resolved. However, a reasonably practical method may start with the largest $x(n)$ and distribute it over other nodes, always keeping in mind that the remaining largest source must satisfy condition (ii) above. The number of calls assigned to (j,k) is denoted as $x(j,k)$. One particularly simple method for partitioning the set of offered node loads, $x(n)$, into a set of node-to-node offered loads, $x(j,k)$, is given in Table 8.

Table 8. A Simple Routine for Partitioning the Set of $x(n)$'s into $x(j,k)$'s

1.	Start
2.	Given: Number of nodes $N \geq 2$ Offered calls $x(n)$; $n = 1, 2, \dots, N$
3.	If: $N = 2$ go to 4 $N = 3$ go to 5 $N > 3$ go to 6
4.	Set: $x(1,2) = x(1)$ Go to 7
5.	Set: $x(1,2) = [x(1) + x(2) - x(3)]/2$ $x(1,3) = [x(1) + x(3) - x(2)]/2$ $x(2,3) = [x(2) + x(3) - x(1)]/2$ Go to 7
6.	If $x(j) = \max x(n)$ and $x(k) = \min x(n)$ set: $x(j,k) = x(k)$ $x(j) = x(j) - x(k)$ $x(k) = 0$ $N = N-1$ Go to 3
7.	Exit

This partitioning method for $x(j,k)$'s requires no more than N partial sorting steps for N nodes. For large networks that is a significantly smaller task than the complete sort which, depending on algorithm, uses between $N \log_2 N$ and N^2 elementary operations. In the case of the complete sort, efficient algorithms are known both for direct sorting of the numbers or of their indices (Knuth, 1968; Press et al., 1988).

The $x(j,k)$ -set is clearly not unique. One can use an existing set to generate many others by adding to the $x(j,k)$ -topology any number of $x(n)$ -preserving loops. A simple example of such a loop is

. . . +a, -a, +a, -a, . . . , +a, -a, . . . ,

where at some point the loop doubles back on itself. Such loops can be short or long and they may cross themselves any number of times. They may also cross other loops. The only restriction is that every $x(j,k)$ remain nonnegative.

Given the set of $x(j,k)$'s, the task of routing the new calls comes next. To simplify matters, let us denote by lx the load $x(j,k)$ offered to the pair (j,k) . Then for each (j,k) a separate routing exercise must be performed to carry the offered load lx . The execution of routing tasks for all (j,k) pairs is handled by the outer loop in Figure 15. On the left hand margin of the diagram that loop is identified by Next (j,k) .

Routing can involve different strategies and different degrees of knowledge about the actual network status. Usually the strategy depends on a fixed and stored routing algorithm. But that need not always be the case. In some situations, for different kinds of stress, certain more or less cooperative searches involving a group of algorithms may be attempted. Whatever the case, a stored or generated set of algorithms is necessary.

Routing also requires knowledge about the topology of the network. Figure 15 indicates the presence of this information under the heading of D matrices. Here D stands for "distance." Distance matrix, D , includes the connectivity matrix, C , as a special case. But, perhaps more significantly, various routing schemes appear much easier to execute using the D matrix. The D matrix can contain several "distance" entries, such as those pertaining to the number of links or hops, the mileages involved, costs of usage, traffic status, engineering capacity margins, reservation of specific facilities, and others. The inclusion of costs and capacity margins leads to the "generalized distance," a concept that may find more use in the ISDN networks of the future than in present-day telephony.

Next, a route (or a set of routes) is established for each pair of nodes (j,k) . This is a sequential "search-and-test" procedure performed by the inner loop in Figure 15. On the left-hand margin this loop is shown as Same (j,k) . The routing algorithm indicates a set of possible routes, R . The capacity tables reveal the available capacity, $C(R)$, for each route, R . If a capacity $C(R)$ is found to exceed the offered load lx , load lx is assigned to route R . Updates are carried out for available link

capacity tables, as well as for the counts of idle and busy customers.

If for all routes, capacity $C(R)$ is less than load lx , the inner loop could declare a blocking event for the entire load. However, a more refined distribution of load lx over several available routes is assumed in Figure 15. If any nonzero capacity is found on a preferred route, that capacity is immediately used to reduce lx . Attempts are then made to distribute the remaining offered load over remaining allowed routes. Only after all available routes with nonzero capacity are exhausted, one assigns the unassigned load to the Blocking event counter. This counter is intended to be a convenient, temporary, file within the ACR loops. It may be cleared between successive, perhaps even completely identical simulation runs of network, traffic and stress. More permanent blocking-event counters are introduced as part of the Output routines.

If (j,k) is not the final node pair, the outer loop picks the next (j,k) . On the other hand, if (j,k) is the final pair, the Add-calls routine is finished for the sampling time in question. The simulation engine advances to the event counters needed for output and thereafter, to the Delete-calls routine (DCR).

One final comment should be made about the illustrated ACR. As outlined in Figure 15, the routine may be too oversimplified. For specific network applications, various ACR modifications may be either necessary or desirable. One particular modification example is encountered in damaged networks that are separated or divided into two or more completely disconnected subnetworks. For such networks, in addition to the familiar GOS blocking one has total service unavailability between certain node pairs. We denote this as "network separation" blocking. While it equals 100% blocking for certain node pairs, when averaged over all network traffic this separation blocking may be considerably less than 100% and thus of interest.

Figure 16 illustrates one ACR modification appropriate for simulation of separated networks. The following motivation is assumed. Let nodes j and k be in two separated subnetworks. Then the $lx=x(j,k)$ calls to be made between the two nodes are impossible. To estimate the network separation blocking in the two separated subnetworks, one needs to estimate the fractions of lx that originate from j and k , respectively.

Let switches j and k both have call-arrivals that obey the Poisson model. Let a_j and a_k be the respective call arrival intensities for the two nodes. Define

$$A = a_j / (a_j + a_k),$$

and assume that $lx=m$ is given. Then it can be shown that the number of calls generated at j and at k follow the binomial distribution. More specifically,

$$\begin{aligned} \text{Pr}(i \text{ calls at } j | m \text{ total calls}) &= \binom{m}{i} A^i (1-A)^{m-i}, \\ \text{Pr}(i \text{ calls at } k | m \text{ total calls}) &= \binom{m}{i} A^{m-i} (1-A)^i, \end{aligned}$$

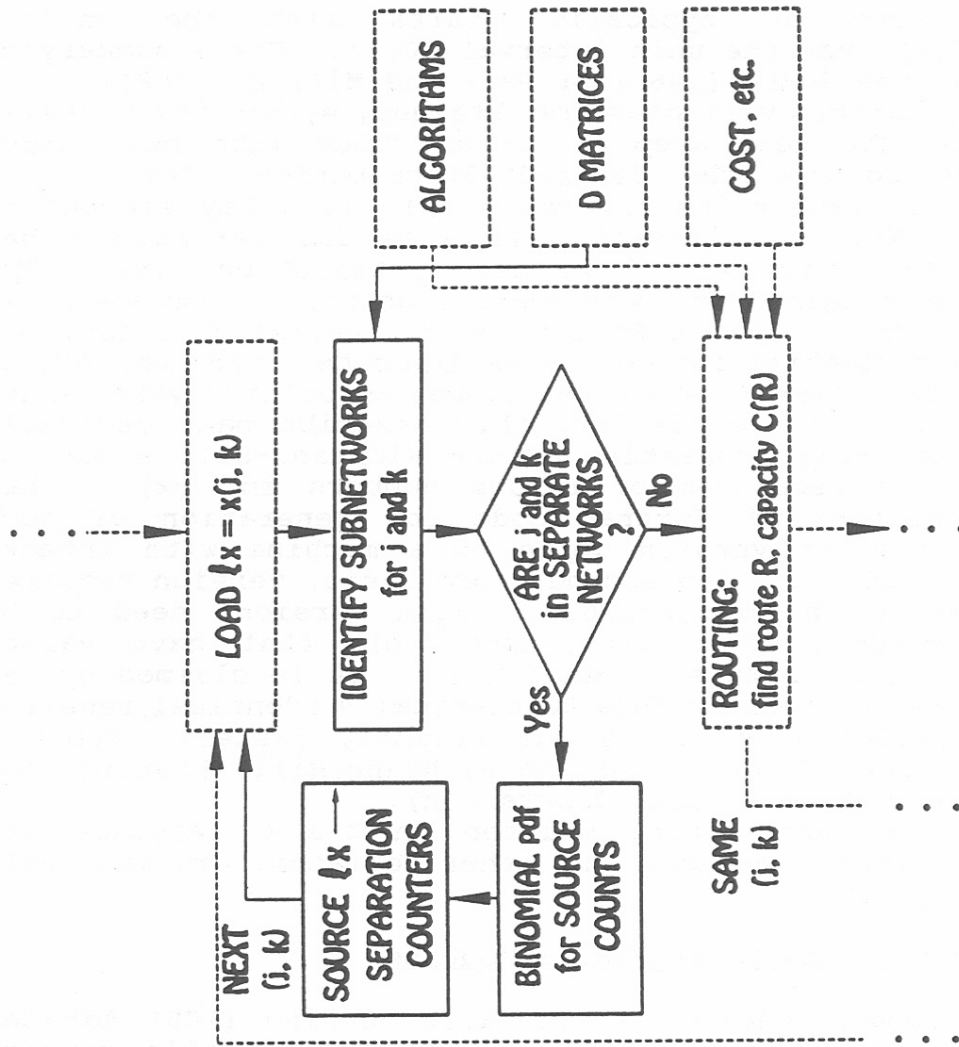


Figure 16. ACR modification for a separated network.

The role of the binomial random number generator is shown in Figure 16 to apply to j and k whenever they are found to be in separated subnetworks. The resultant, "separation blocked" attempts are stored in Separation counters for later processing.

Simulation Engine: Random Number Generation

In ACR, DCR, and elsewhere there is need for random number generation. In the previous section, for example, independent Poisson and binomial variables had to be generated. Whether the desired distribution is Poisson, binomial, Gaussian, or any other, the number generation typically starts with the uniform distribution $U(x)$ over the unit interval $[0,1]$. For a summary of useful methods, see Knuth (1968) or Park and Miller (1988).

Once these uniform variables are obtained, either individually or as a group, the next step is to map them into new random variables that possess the desired distribution, $F(x)$. The function $F(x)$, or rather its inverse $F^{-1}(x)$, is a key element in this mapping. More specifically, if a random variable x has distribution $U(x)$; then $y = F^{-1}(x)$ has distribution $F(x)$. The distribution, so obtained for y is continuous for continuous $F(x)$. When $F(x)$ is discrete, like a Poisson or a binomial distribution, nearest neighbor quantization generates discrete y from uniform x .

Coates et al., (1988) review the issues associated with random number generation. In particular, they describe many desirable (perhaps even optimal) properties of the Wichmann-Hill algorithm for generation of sequences of random numbers in $U(x)$. They provide two versions of Fortran code for generation of such variables. The first version works on a machine with integer arithmetic up to 30,323. The second, more basic, version requires integer arithmetic up to 5,212,632. Both versions need to be seeded with integers $x(0)$, $y(0)$, and $z(0)$, that have values anywhere in the range between 1 and 30,000. It is claimed by the authors that both versions produce statistically identical results, but that the second version may run slightly faster. Table 9 illustrates the second version of the Wichmann-Hill algorithm for generation of N uniform random numbers $u(n)$.

Random number generators, written in the C language and applicable to uniform, Poisson, and other deviates, are available in Press et al., (1988).

Simulation Engine: The Delete-Calls Routine

Equal in importance to the add-calls routine (ACR) for the operation of the simulation engine is the delete-calls routine (DCR). A functional diagram of the DCR is given in Figure 17. One observes that the DCR is the dual of the ACR. The basic structure of Figure 17 is nearly identical to Figure 15, however, most of the DCR variables and tables have new names. Their meaning is the converse of terms defined earlier for ACR.

During each sampling interval and at every switching node, n , the DCR attempts to remove $y(n)$ existing calls. These $y(n)$ are random variables generated from a Poisson distribution, denoted as $q(x)$. Distribution function $q(x)$ differs from the previous $p(x)$,

Table 9. The Basic Wichmann-Hill Algorithm

1. Start
2. Given: Integers
a = 171
b = 172
c = 170
Distinct primes
p = 30269
q = 30307
r = 30323
Seed integers
$1 \leq x(o) \leq 30000$
$1 \leq y(o) \leq 30000$
$1 \leq z(o) \leq 30000$
Sequence length N
3. For n = 1, 2, ..., N do:
$x(n) = ax(n - 1) \pmod{p}$
$y(n) = by(n - 1) \pmod{q}$
$z(n) = cz(n - 1) \pmod{r}$
$u(n) = \left(\frac{x(n)}{p} + \frac{y(n)}{q} + \frac{z(n)}{r} \right) \pmod{1}$
4. Exit

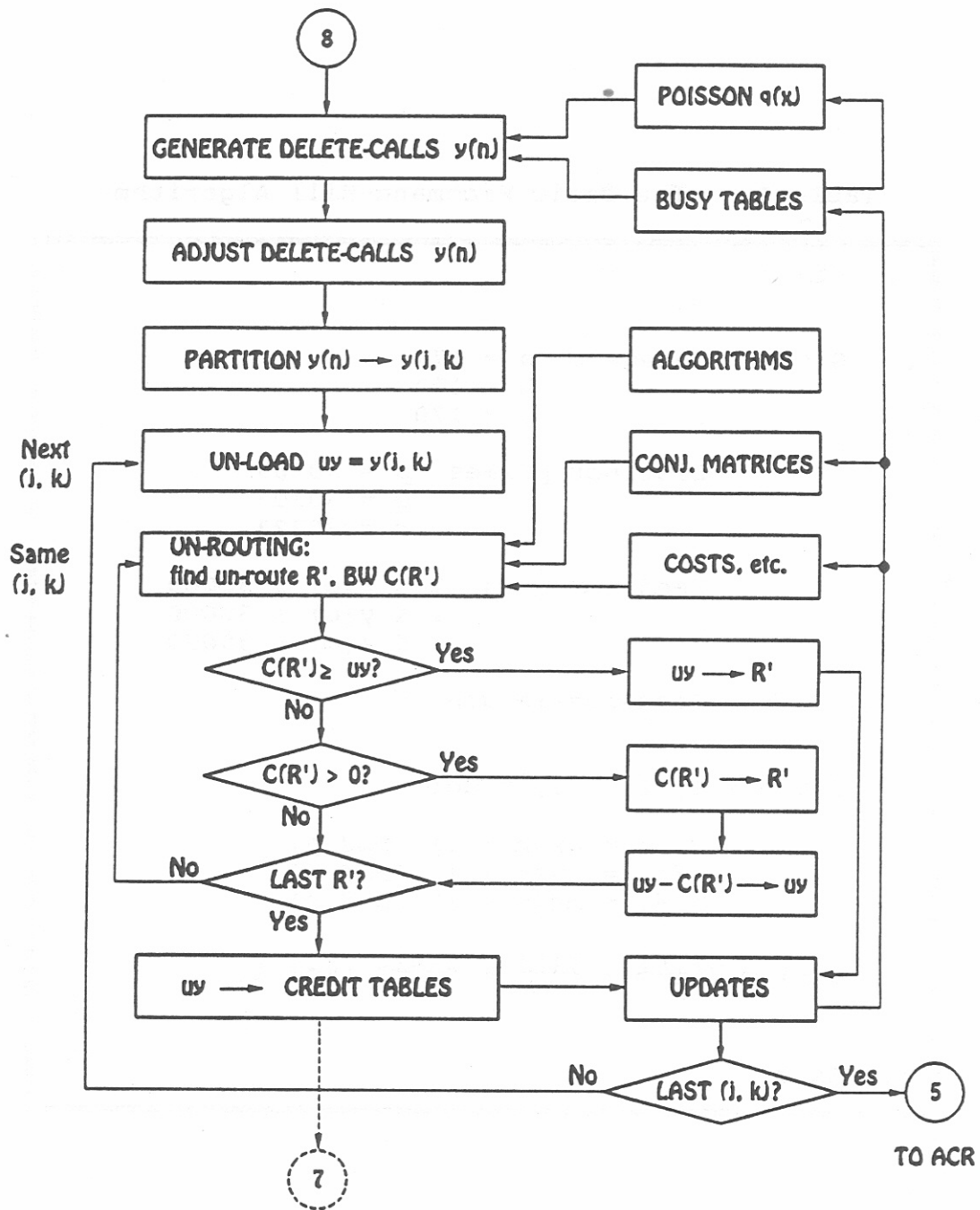


Figure 17. Delete-calls routine (DCR).

compare Figures 15 and 17. Function $q(x)$ depends on the number of busy users and their specified mean holding times, not on the number of idle users and their calling rates.

As before for $x(n)$, random variables $y(n)$ are modified (if needed) to satisfy the realizability conditions (i) and (ii) of the previous section. No credit adjustments are necessary for $y(n)$'s. Next, the set of $y(n)$'s is partitioned into point-to-point deletion numbers $y(j,k)$ and the respective deletions are attempted for all node pairs (j,k) with nonzero $y(j,k)$.

The process of exhausting all (j,k) options employs two nested loops. On the left margin of Figure 17 the two loops are identified as Next (j,k) and Same (j,k) . Since certain existing routes are to be removed here, the appropriate name for the functions within the loops may be "un-routing." Likewise, the deleted load can be called an "un-load," and so on. The algorithms for un-routing could be the same as for routing, but the distance and connectivity matrices are different. They are now based on call-occupied (busy) links. The network un-routing matrices are in this sense the conjugates of the true routing matrices. The two sets of matrices can be viewed as a single matrix, where the individual elements are pairs of numbers: unused link capacity being the first number and occupied capacity the second.

One defines, for each node pair, its un-load $uy = y(j,k)$. One next searches consecutively for un-routes R' and their capacities $C(R')$ to accommodate uy . If the search is successful, existing circuits are cleared, capacity and connectivity tables are updated, and the routine proceeds to the next (j,k) .

If the un-routing search is unsuccessful, some part of uy may remain intact and un-deleted (or un-un-routed). Such events are the GOS blocking equivalents for un-routing. They are expected to be rare events. However, to reduce the bias of accumulated uy remainders, the following stratagem may be employed. One stores the uy remainders and applies them against (i.e., subtracts from) the add-calls in the next sampling cycle. The Credit tables that collect uy 's in Figure 17 are used in this manner to adjust the add-calls in Figure 15.

After completion of the final (j,k) un-route, the DCR is finished for a particular timing cycle. The simulator steps ahead to the next ACR.

Output

This simulation example is concerned only with circuit-switched network performance related to call blocking. As already discussed, the primary performance is the network blocking GOS. When a network is connected (not separated into two disjoint networks), the blocking GOS is caused by traffic congestion. It is also the only blocking measure for such connected networks, unless one chooses to discriminate between the GOS for different regions, different originating and terminating end offices, different service classes, or at different times in a dynamically changing scenario.

When a network is divided into two or more separate subnetworks, the blocking GOS still has its intended meaning.

However, another blocking type--this one due to service unavailability or separation blocking--may be equally or more significant.

In both cases, the simulator output routines are quite similar. Temporary or scratch-pad registers of the ACR are used to count all different classifications of blocking events throughout the simulation run. At predetermined intervals, the contents of these temporary registers are transferred to the permanent registers of the output routines. Figure 18 summarizes the main output functions.

The simulator outputs are various reports, such as tables and graphs. Their content and form depend on the definition of the performance measures. Examine, for instance, the blocking GOS. It is defined as the ratio of the number of network blocked calls to all attempted calls. A trivial amount of computation is involved in this step. However, if one desires confidence levels or tests of statistical significance for the GOS numbers, then more complicated statistical processing becomes necessary.

Tables, Arrays, and Databases

Simulation of any reasonably large and realistically complex network requires processing of volumes of data. The same is true for time-sampled simulation of circuit-switched telephone networks. Depending on computer speed, software, memory, and database management systems, different arrangements of serial and parallel tasks can be implemented for the functions outlined earlier. Normally, the data to be handled are grouped in various arrays, such as lists, tables, matrices, and so forth.

The exact formats and amounts of data arrays cannot be specified here. Too much depends on network applications and simulation system implementation. However, estimates can be made for memory requirements associated with the time-sampled approach described above. In what follows, we divide the data arrays into two categories: the lists (one-dimensional arrays) and matrices (two-dimensional arrays).

Table 10 presents the ten apparently most significant lists for time-sampled simulation of circuit-switched telephone networks. However, not all of these lists are absolutely necessary for all networks. In certain scenarios some lists may be deleted, others modified or considerably abbreviated. Occasionally, some particularly expedient lists may have to be added. For instance, because of

$$ub(n) + ui(n) = u(n),$$

one of the first three lists, namely $u(n)$, $ub(n)$ and $ui(n)$, is redundant. For homogeneous infinite-user network models one sets

$$\lambda(n) = \lambda,$$

$$\mu(n) = \mu,$$

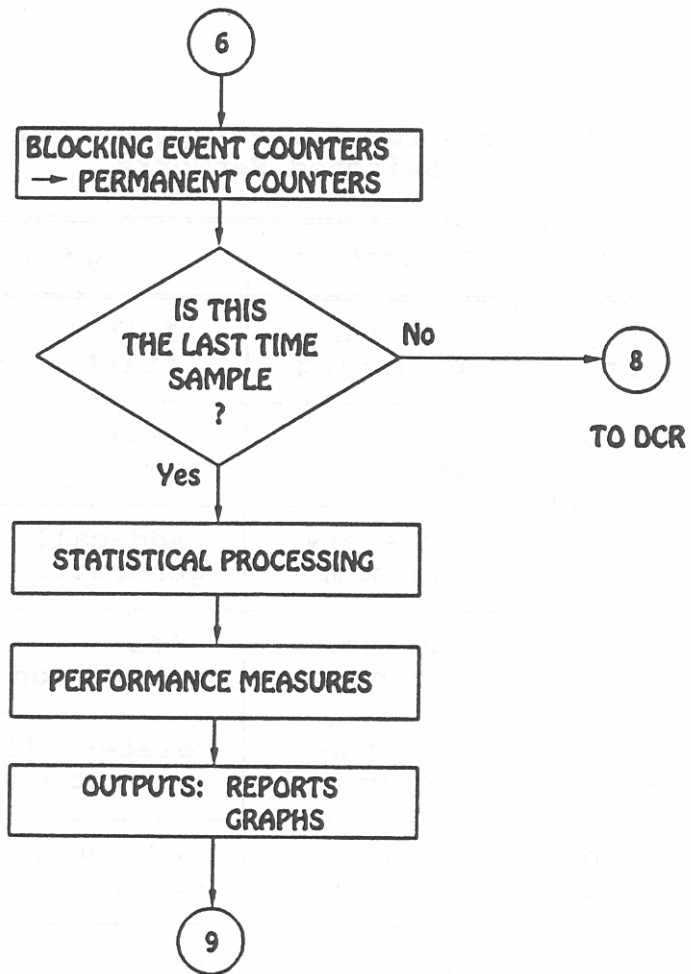


Figure 18. Output routines.

Table 10. List Arrays for Time-Sampled Simulation of Circuit-Switched Networks

Symbol	Function	Generation	Usage	Dimension
$u(n)$	Number of users	At the beginning	Initial traffic	N
$u_b(n)$	Number of busy users	In every time sample	Delete-call generation	N
$u_i(n)$	Number of idle users	In every time sample	Add-call generation	N
$\lambda(n)$	Calling rate	At the beginning	Add-call generation	$\ll N$
$\mu(n)$	Service rate	At the beginning	Delete-call generation	$\ll N$
$x(n)$	Number of add-calls	In every time sample	Add-call generation	N
$y(n)$	Number of delete-calls	In every time sample	Delete-call generation	N
$z(n)$	Deletion credits	Occasional time sample	In add- and delete-calls	$\ll N$
$b(n)$	Blocking counts	In every time sample	GOS, network blocking	$\ll N$
$s(n)$	Separation counts	In every time sample	GOS, network separation	$< N$

so that two constant numbers suffice to replace the two tables, $\lambda(n)$ and $\mu(n)$. To simplify further, one could set credits $z(n) = 0$. This would leave additional residual traffic in the network, thus increasing congestion and yielding upper bounds on the blocking grade-of-service. Finally, when dealing with connected (not separated) networks, the last list of $s(n)$'s is redundant.

One concludes from Table 10 that only about 5 to 6 lists may be essential at the same time. Except for calling rate and service rate tables, which consist of positive real numbers, the elements of all other tables are nonnegative integers. The total volume occupied by these list arrays is about $6N$. Even for thousand-node network topologies, this storage requirement appears almost negligible.

A more significant demand on the simulator capabilities, both in terms of speed and storage capabilities, appears to be the dominant need for repeated processing (i.e., generation, erasure, and modification) of certain lists. The $x(n)$ and $y(n)$ random numbers, for example, are generated in each sampling interval. Subsequent processing further generates new sets of $ub(n)$ and $ui(n)$, both in the ACR and DCR phases of the same sampling interval. It is important that the speed of the machine allows execution of these list processing tasks in the shortest possible time.

The matrix array requirement is indicated in Table 11. Eleven matrices are listed. Again, as for arrays of Table 10, some matrices are more important than others. In certain network models, specific matrices can be substantially reduced, while others can be ignored altogether. However, it is entirely possible that additional matrix arrays could also be added, if needed in certain network situations.

Since the connectivity matrix C is a trivial part of the distance matrix, D , the initial network topology is fully represented by D . For a full-duplex network, the network size need not be larger than $N(N-1)/2$. As traffic congests the network facilities, some links may become fully loaded. These links are then not available for additional traffic. The new traffic-dependent topology is depicted by a modification of D , called D_t . A distance element in D_t is always larger than or equal to the corresponding element in D . When the carried traffic is light, the two matrices can be identical. Matrix A_t , called the activity matrix, is also a distance matrix. But it considers only channels currently active or busy. These channels can be un-routed or left intact as part of DCR functions.

There are three link matrix arrays shown. Because a multi-channel link can be partly busy and partly idle at the same time, the busy and idle link tables need not be duals within the overall link tables, $l(j,k)$. Since in most practical networks the number of links is much smaller than would be needed for full connectivity, the sizes of all link tables are much less than $N(N-1)/2$. The busy and idle link tables could be changed at any time sample. However, these changes are expected to be random and rather infrequent in a simulation run.

Table 11. Matrix Arrays for Time-Sampled Simulation of Circuit-Switched Networks

Symbol	Function	Generation	Usage	Dimension
D	Distance matrix	At the beginning	Initial topology	$N(N-1)/2$
D_t	D, adjusted for traffic	In every ACR and DCR step	Network still available	$N(N-1)/2$
A_t	Activity matrix	In every ACR and DCR step	Network in use	$N(N-1)/2$
$l(j,k)$	Link tables	At the beginning	Capacity, cost	$\ll N(N-1)/2$
$lb(j,k)$	Busy Link tables	In every ACR and DCR step	Busy link capacities	$\ll N(N-1)/2$
$li(j,k)$	Idle link tables	In every ACR and DCR step	Idle link capacities	$\ll N(N-1)/2$
$x(j,k)$	Number of j-to-k calls	In every ACR step	New set of add-calls	$\ll N(N-1)/2$
$y(j,k)$	j-to-k un-calls	In every DCR step	New set of delete-calls	$\ll N(N-1)/2$
$R(j,k)$	Routing tables	Depends on algorithms	Listing of j-to-k routes	$\ll N(N-1)/2$
$C(R)$	Route capacities	Depends on algorithms	Largely for fixed routes	$\ll N(N-1)/2$
$b(j,k)$	j-to-k blockages	In every ACR step	Node-to-node blocking GOS	$\ll N(N-1)/2$

The matrices, $x(j,k)$ and $y(j,k)$, for node-to-node add-calls and delete-calls need normally be stored only for the execution of a particular ACR and DCR, respectively. If each $x(j,k)$ is routed immediately, or each $y(j,k)$ is un-routed immediately, the only reason for their longer storage could be associated with some post-mortem statistics. If the sampling interval is not too large, the sizes of both $x(j,k)$ and $y(j,k)$ will be considerably less than $N(N-1)/2$ in any single interval.

The two routing tables, $R(j,k)$ and $C(R)$, are used only when the network routing algorithms call for such. Otherwise they can be ignored. When fixed routing tables are not employed, available end-to-end capacities must be computed from individual link capacities and their occupancies.

The last row of Table 11 lists the node-to-node blocking counts. These are usually relatively rare events even for most networks under stress. Moreover, there seems to be little reason that the blocking counts cannot be immediately added up, either totally or according to some distinguishable categories of events. This is so, because one is generally most interested in network-wide or regional blocking probabilities over an extended period, such as the busy hour (BH).

4.5 Simulation Scope and Performance Targets

There are two main goals for the outlined simulation program. First, the exercise must identify those satellite network alternatives that offer the best support to a stressed (or damaged) terrestrial network. And second, the simulation must yield quantitative descriptions of the resultant service improvement. As noted earlier, this improvement can mean rapid service restoral to suddenly disconnected nodes or general GOS enhancement to stressed parts of the network.

The scope of the simulation program must naturally address many questions beyond mere identification of the optimal satellite network. The identification of factors, requisite in circuit switched network simulation, can begin with quite simple network scenarios. Consider, for instance, a simulation process that consists of four categories of runs.

In the first category (Run 1) the simulator establishes a performance baseline for the terrestrial network alone, given normal traffic load and no facility damage (i.e., the no-stress condition).

In the second category (Run 2) the simulator retains the terrestrial network, but subjects it to traffic overloads and specified facility outages (i.e., the stress condition). Comparison of Run 2 and Run 1 shows how the performance of the terrestrial network suffers under the assumed stress scenario.

In the third category (Run 3) the simulator adds the satellite network to the terrestrial network, assuming no damage of facilities and no traffic overload (i.e., no stress). Comparison of Run 3 to Run 1 shows the relative satellite-terrestrial advantage under normal operating conditions.

In the fourth category (Run 4) the simulator retains the satellite-terrestrial hybrid network, but subjects the terrestrial network to the same stress as in Run 2. Comparison of Run 4 and Run 2 answers the most important question: How much, if any, does the presence of the satellite network help under stress? Related questions pertain to the design of the satellite network, such as the optional placement of ground stations. Comparisons of Run 4 to either Run 1 or Run 3 appear to be of secondary significance. Nevertheless, they may shed some light on the complex processes that occur in stressed networks, and thus may be beneficial.

The specification of simulation runs, like the above, takes place in a multidimensional parameter space. As the previous text shows, the number of parameters can be quite large. To simplify matters, it is convenient to group mutually related parameters under common headings. One possibility is a three-dimensional or three-part grouping that segregates parameters associated with the network, the traffic, and the stress. An illustration of this viewpoint is given in Figure 19.

While this viewpoint is intuitively appealing and perhaps an easy starting step for modeling, the simple three-dimensional parameter space does have considerable shortcomings. It fails to define the multiplicity of lesser parameters that are so essential for a workable simulation model. This point is addressed next by looking first at the network axis of Figure 19.

Complex and Simple Networks

Existing real networks, nationwide, regional, or local, are far from the symmetric and homogeneous topology suggested earlier in Figure 1. Being driven by historic, economic, traffic, and growth considerations, the real networks have rather irregular appearances. An example, shown in Figure 20, is the longhaul broadband network of AT&T. In this diagram, the dark squares represent more than 80 switching nodes, while there are at least 200 interconnecting cables of different high capacity optical-fibers (Ash and Schwartz, 1990). Other nationwide topologies are quite similar. See Edwards (1991).

Specification of such irregular networks, when done by distance or connectivity matrices, is not much more complex than for symmetric or regular networks. However, the difficult part comes when one must assess stress, damage, and recovery measures in different parts of the network. Ultimately, of course, the true value of simulation is in its application to real networks, whatever their topologies. But to learn about simulation characteristics and the interpretation of results, simpler topologies may also be of considerable help.

Consider the very simple and symmetric network of Figure 21. It has $N=12$ nodes and $L=12$ terrestrial links. Exactly two links home on every node, making this a ring or loop network. The diagram also shows $X=4$ Earth stations. The Earth stations are collocated with the indicated switching nodes. These Earth stations are equally spaced on the network, so that no node is more than one terrestrial span away from its nearest Earth station.

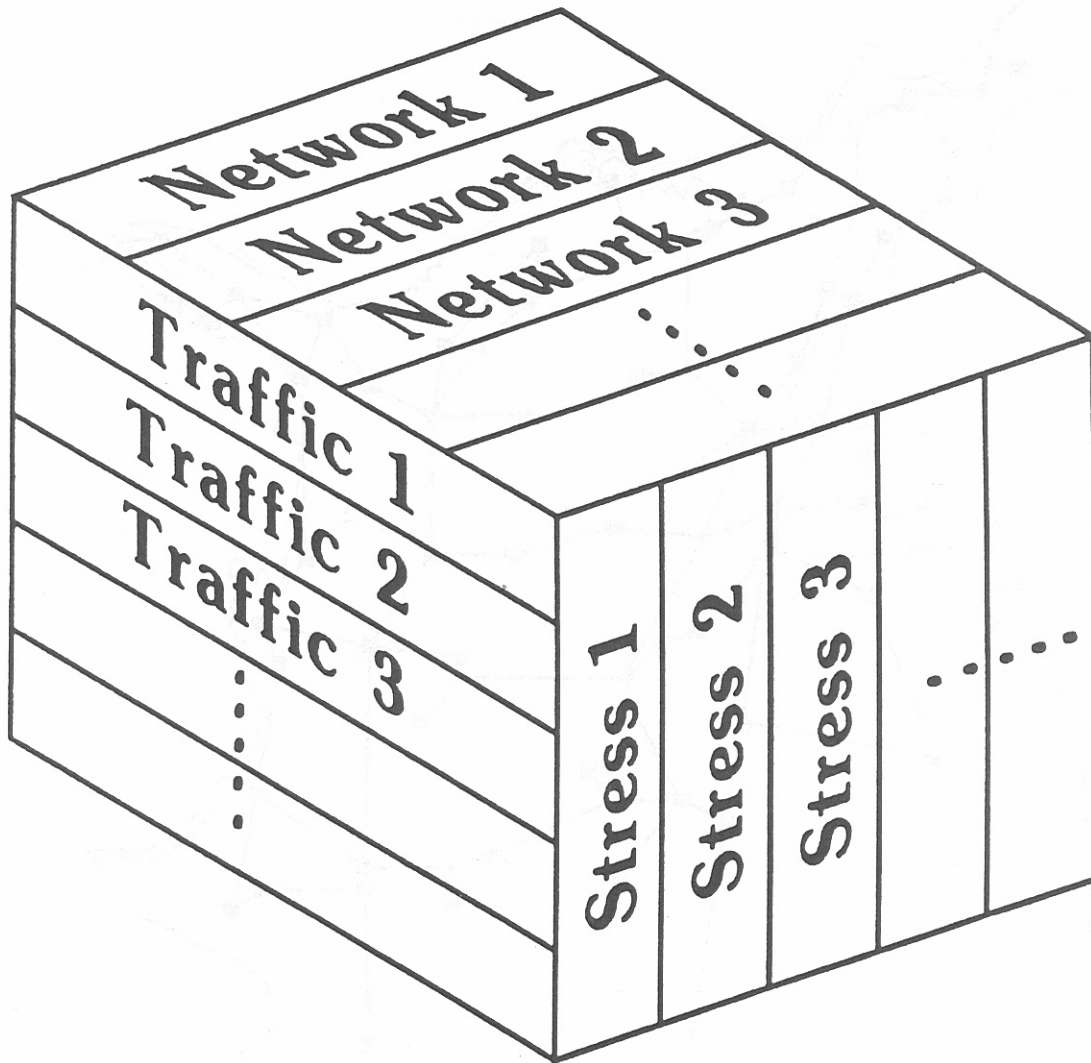


Figure 19. The parameter cube for network simulation.

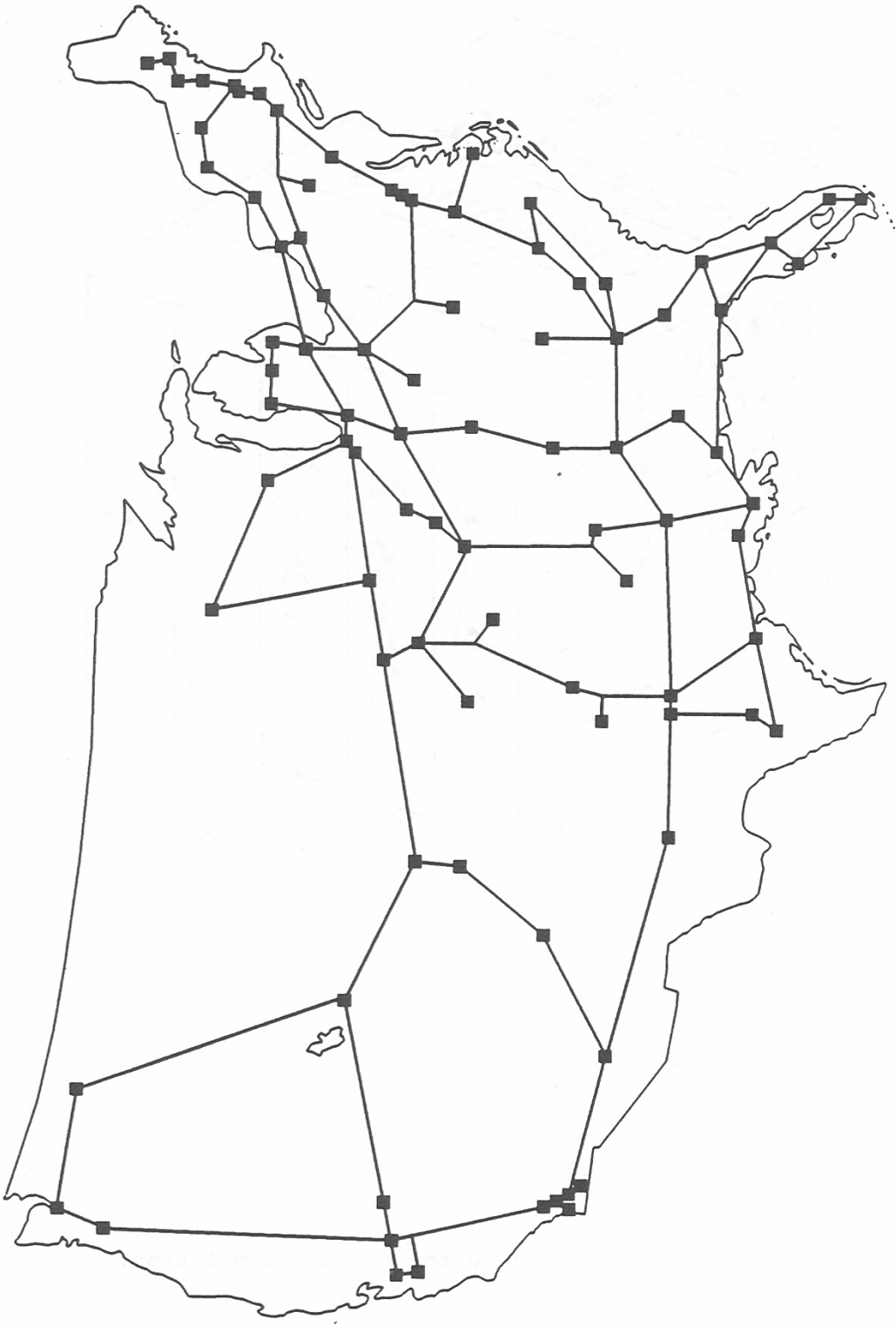


Figure 20. Illustrative network model of existing facilities.

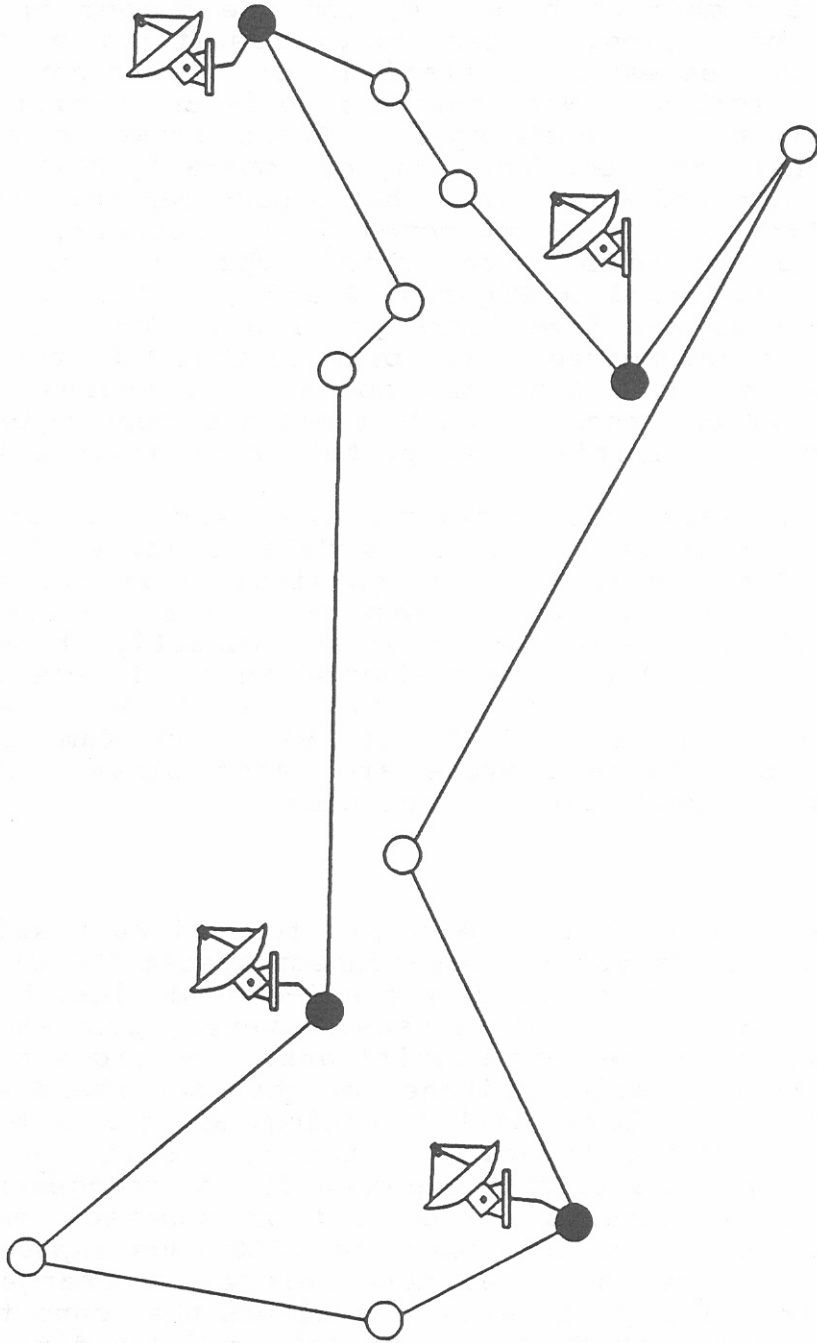


Figure 21. A symmetric test network with $N=12$, $L=12$, and $X=4$.

The nature and capacity of the space segment remains to be defined, as does the capacity of the entire terrestrial network. Given a specified offered traffic load and a stress scenario, the effects of the absence vs. presence of the satellite network on GOS can be determined by simulation. Furthermore, these results are expected to be more verifiable and therefore more meaningful because of the inherent network symmetries.

The symmetric ring topology of Figure 21 can be generalized in several ways. The number of nodes, N , and the number of Earth stations, X , can be varied. This would result in different shortest paths to the nearest Earth station. Or one can generalize the ring concept as follows. Note that a ring is any network that has exactly two links for each node. Other networks can be constructed such that each node has 3, 4, or generally $2L/N$ number of links. The maximum number of links that a node can have is $N-1$. That topology occurs in a fully connected N -node network.

Two symmetric networks with this type of increasing connectivity are illustrated in Figures 22 and 23. Figure 22 has three, while Figure 23 has five links per node. The number of Earth stations, X , is determined on the premise that X be the least value that ensures no more than one hop to the nearest Earth terminal for all network nodes. Such idealized topologies are expected to be useful in initial interpretation or diagnostics of simulator outputs.

As emphasized earlier, the existing real networks of today are usually far from symmetric. Their models require detailed specification. Furthermore, their physical structures and functional architectures continue to evolve at a rapid pace. A good example of this is the evolution of the Intelligent Network concept (Robrock, 1991). The concept started in the 1980's in the pre-divestiture Bell System. After divestiture, it has been the topic of RBOCs, as well as of AT&T studies. New families of network nodes and operations systems are major players in the Intelligent Network concept and its evolution.

Switches and DACS

A relatively common nodal element for circuit-switched networks is the digital automatic cross-connect system (DACS), also called the dynamic cross-connect, or the add-drop device, by some (Ash and Schwartz, 1990). Its future uses, however, promise to be far more common and therefore more significant. The cross-connect devices can supplement existing switches or they can stand alone. Instead of performing individual call switching, as is done by most switches, the DACS performs capacity switching. Switching at a lower speed, perhaps not faster than one capacity rearrangement per 10 minutes, the cross-connects either add or subtract channel capacities to different circuits. When the DACS adds capacity to a node-to-node pair where there was none before, it changes the logical network connectivity. Likewise, it alters the connectivity when it removes all capacity from linking certain nodes directly.

For customers interested in private data networks, DACS offers new alternatives through the Fractional T1 (F-T1) services. An

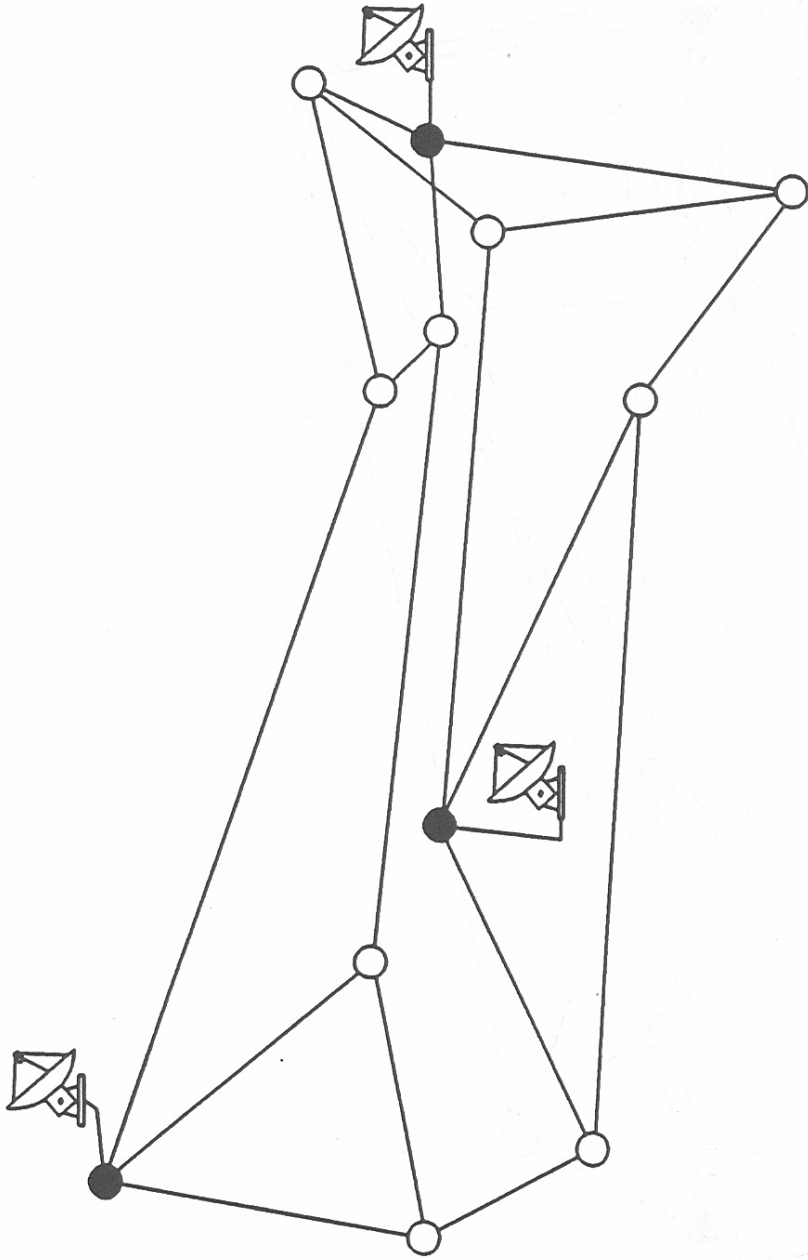


Figure 22. A symmetric test network with $N=12$, $L=18$, and $X=3$.

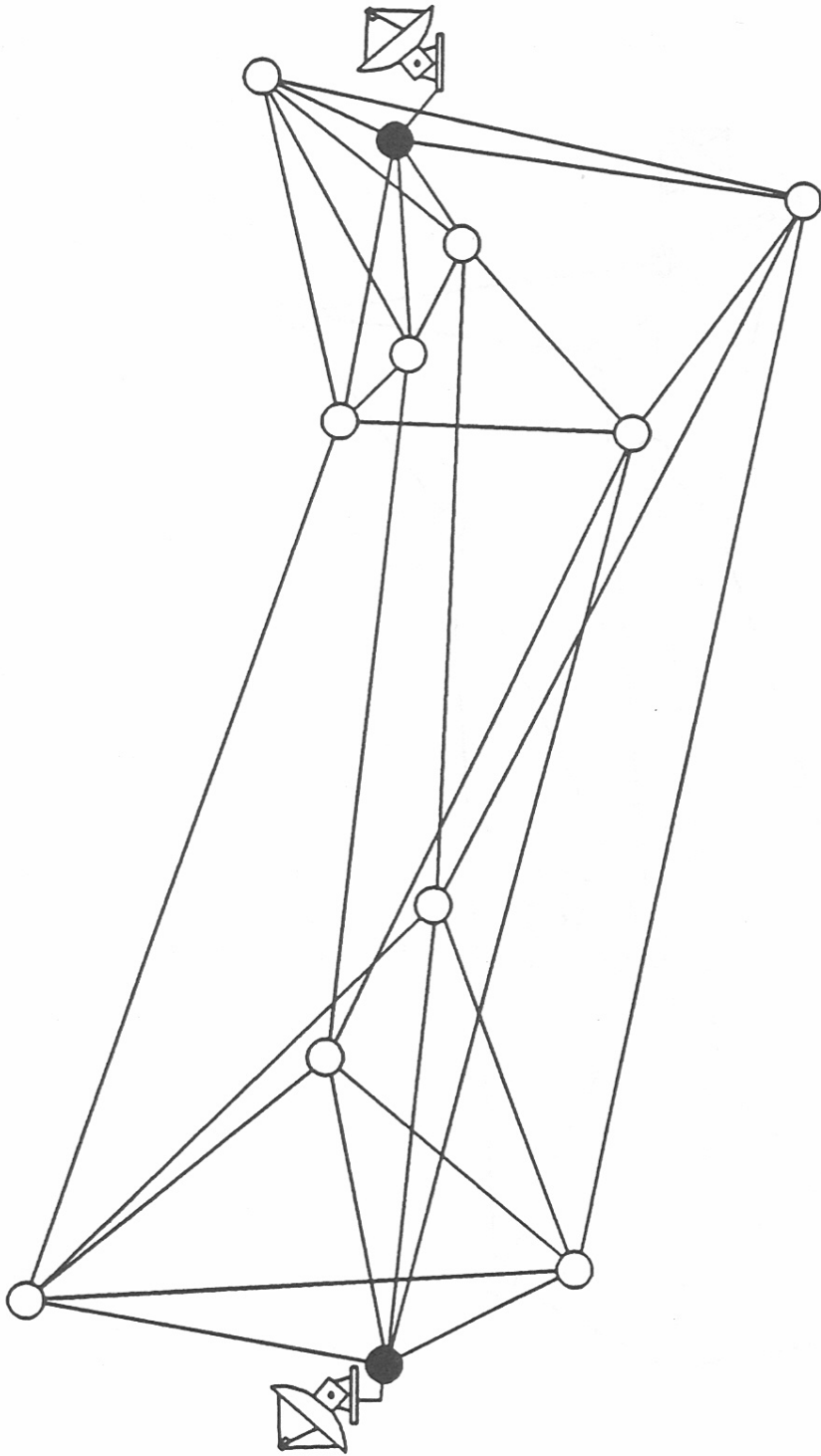


Figure 23. A symmetric test network with $N=12$, $L=30$, and $X=2$.

example of F-T1 service is the AT&T's Accunet Spectrum of Digital Services or ASDS (Edwards, 1991).

Traffic switching, a large part of which is routing, must be clearly differentiated from capacity switching. In the case of circuit-switched traffic, routing finds the circuit paths and helps establish connections for individual point-to-point calls. The typical procedure for routing is illustrated in Figure 24. When a new call is initiated, its source and destination addresses identify the end points for the desired route. First, the process tries to find a direct route. If a direct route is found, that route is assigned to the call. The network proceeds to establish the actual connection, while the routing process is released. It can turn its attention to the next arriving call.

On the other hand, if there is no direct path available, the mechanism seeks alternate routes. Many algorithms have been proposed for alternative routing. In Figure 24, the scheme of VIA routing is indicated. The VIA routing is based on the premise that for every source-destination node pair there exist unique, well-defined, VIA switches or nodes that are one link distant from both terminal nodes. In the current VIA scheme, as advocated by AT&T, only a single VIA switch is to be permitted per route. If a VIA route is found, the route is seized and the task is completed. On the other hand, if no acceptable VIA route is identified, the call is declared to be "blocked."

Alternate routing with the VIA algorithm is efficient from the network's point of view, because it strikes a compromise between offering some alternative routing relief, while avoiding the pitfall of many excessively long, resource consuming, alternate routes. There is, however, a question whether the VIA routing, which is considered optimal today, will remain as the preferred scheme in the future. Other methods may become more advantageous as technology matures (Girard and Bell, 1989; Mitra and Seery, 1991).

Capacity switching uses the DACS facilities and is based on different principles. The concept of capacity switching is illustrated in Figure 25. The key things to observe are:

- (a) Congestion of any one route is relative to the congestion of other routes. Thus, it makes little sense to add DACS capacity to route A, if that capacity must be taken away from another more congested route B.
- (b) DACS switching involves delays. Because, if a group of trunks is to be switched, but it carries live traffic, then one should wait till all calls are cleared before capacity switching. This will cause delays and some network inefficiencies.

It follows that a capacity switching controller, as shown in Figure 25, must monitor the congestion of many circuits. Based on this global congestion status, the controller can establish thresholds when to switch and when not to switch the DACS capacities from one route to another. Also the inherent DACS

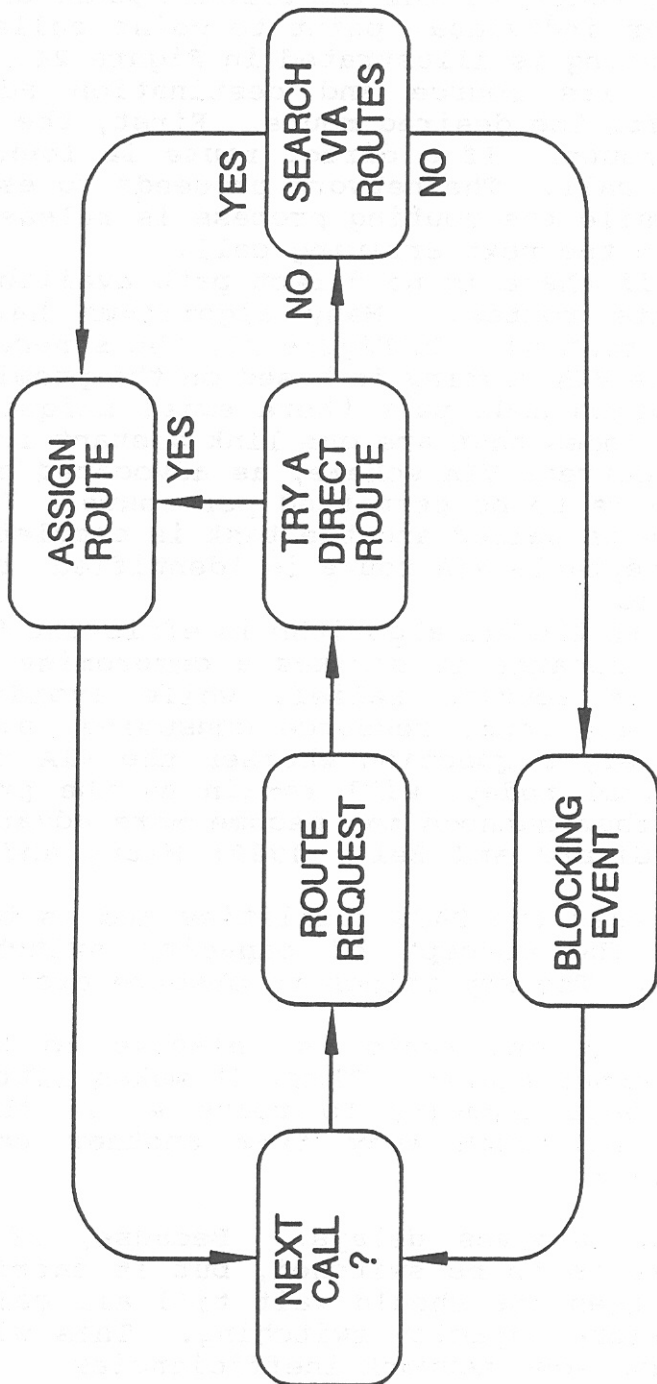


Figure 24. Preferred alternative routing scheme of today.

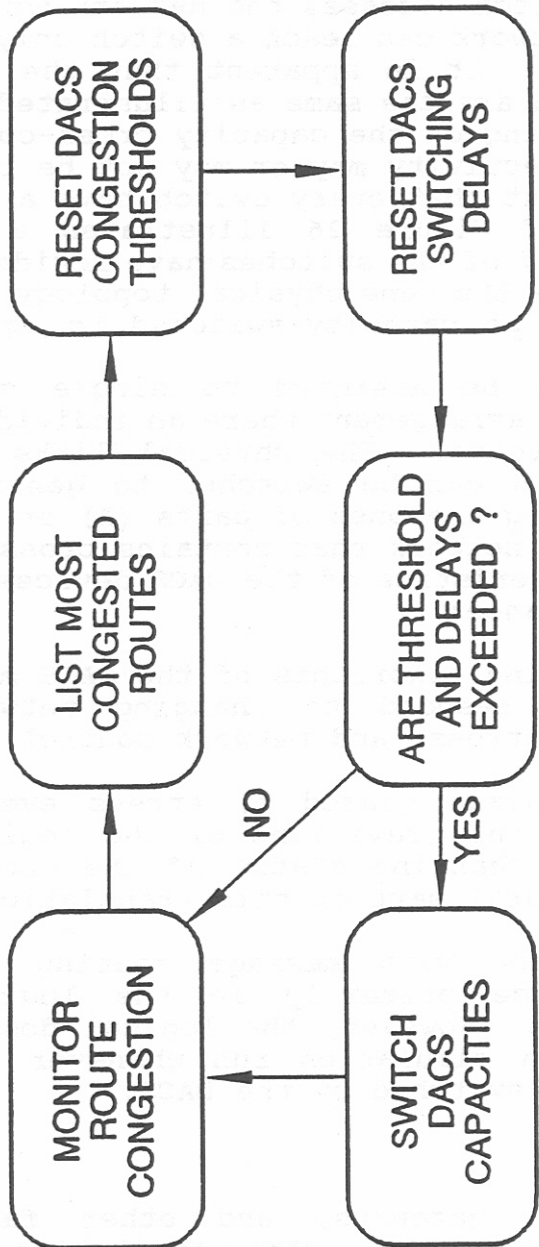


Figure 25. The principle of DACS capacity switching.

delays must be taken into account. Only when the appropriate thresholds are exceeded, does the actual switching of DACS capacity take place.

Other than the caveats (a) and (b) above, cross-connect devices are versatile tools for network applications. A simple illustration of their versatility is shown in Figure 26. The DACS or cross-connects are indicated by crossed squares. The switches are circles. In part (A) of the figure, every switch has its own dedicated DACS. A switch accesses the network through its own DACS. Conversely, the network can reach a switch only via the corresponding colocated DACS. It is apparent that the physical links of part (A) of Figure 26 are the same as illustrated earlier in Figure 6. However, depending on the capacity cross-connection of the DACS, the logical connectivity may not be the same.

There is no need to insist that every switch have a separate resident DACS. Part (B) of Figure 26 illustrates a network implementation, where only half of the switches have resident DACS. Nevertheless, this network has the same physical topology, whereas the logical connectivity can be capacity-switched to agree with part (B) of Figure 6.

Cross-connects need not be assigned to single switches. Part (C) of Figure 26 shows an arrangement where an individual DACS serves a group of three switches. The physical links are now different. However, the DACS can be switched to generate any logical connectivity, including the ones of parts (A) or (B).

When modeling a switched network that contains cross-connect devices, the technical characteristics of the DACS devices must be taken into account. For instance:

- * The capacity switching algorithms of the DACS must be implemented to respond to changing network topology, traffic, stress, and network control.
- * Physical network damage caused by stress events must be translated into revisions of the logical connectivity. The changing status of the cross-connects is an integral part of this translation.
- * Algorithms associated with message routing and switching take place primarily in the logical connectivity domain. However, the logical domain may change during a simulation run whenever the link capacities are switched by the DACS.

Network Specification

Switches, DACS, network gateways, and other facilities constitute the nodes of a given network. Other facilities, such as different channels or transmission media, constitute the previously described links. For a simulation model, their topological deployment and other characteristics must be specified.

The scope of any so general a specification can be quite extensive. A skeletal framework for network specification is initiated in Figure 27. Note that this framework follows a

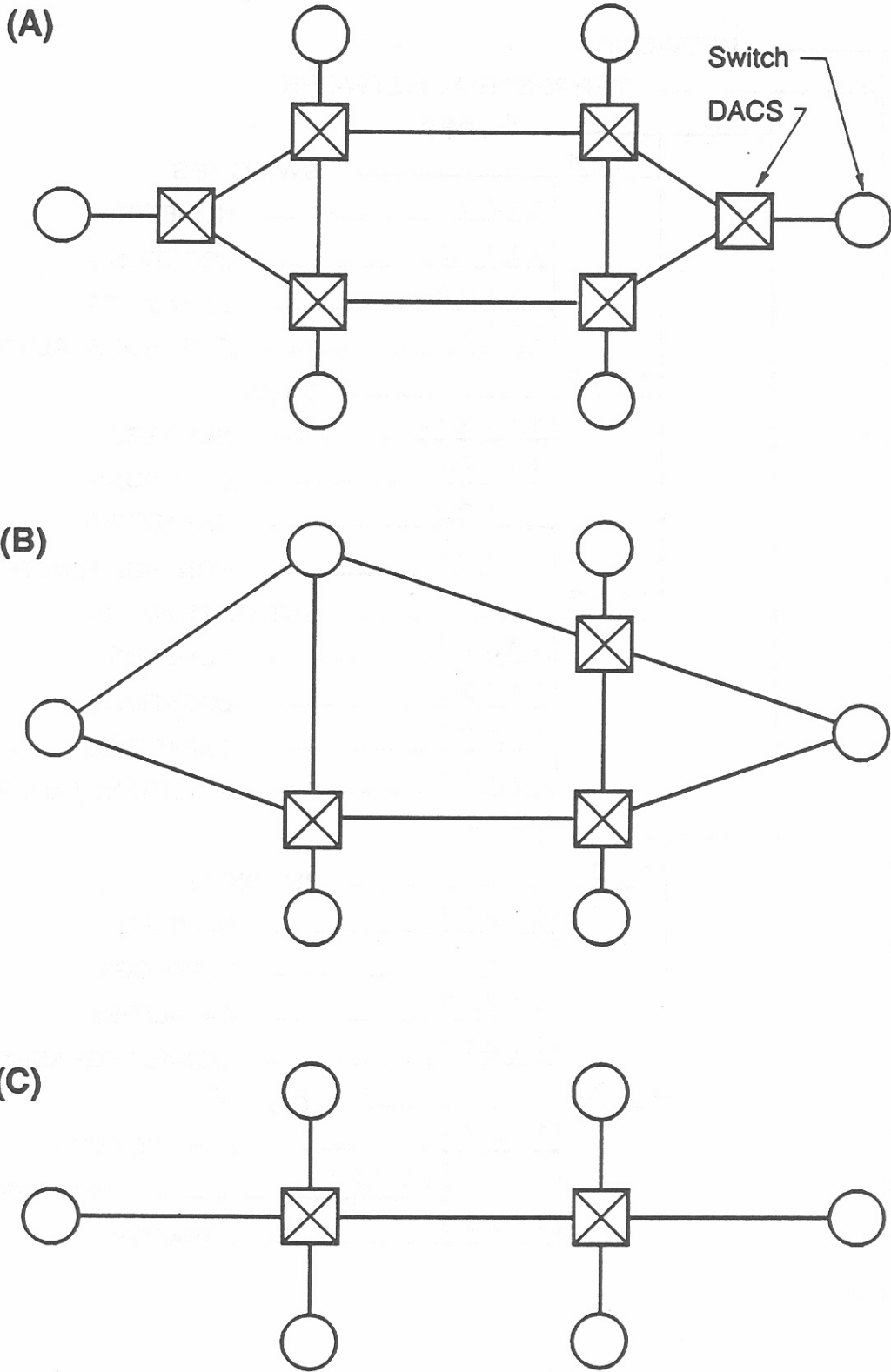
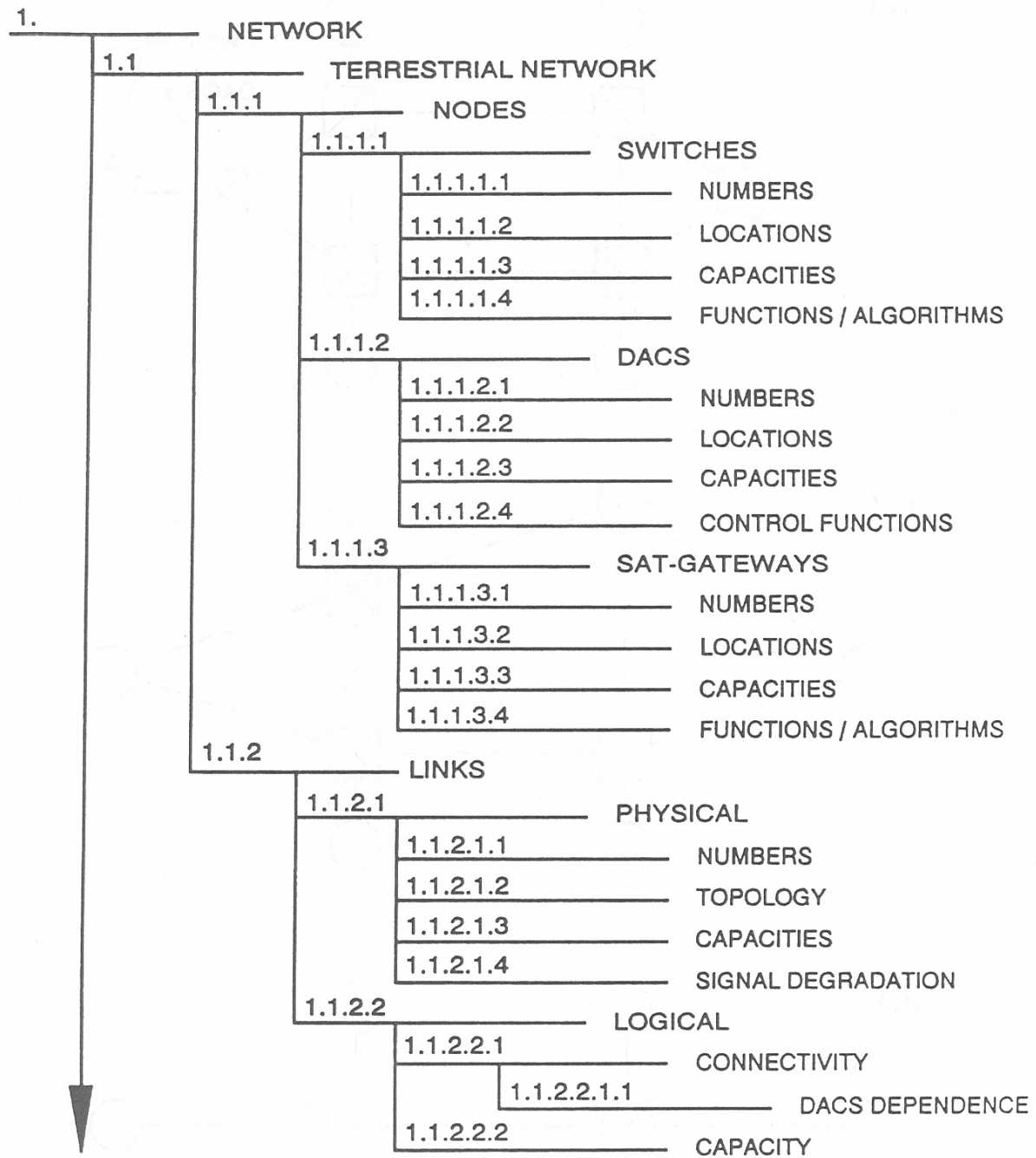


Figure 26. Different deployment of cross-connect devices can yield the same logical connectivity.



TO 1.2 and 1.3

Figure 27. Specification of the simulation model:
Part 1.1, the terrestrial network.

tree-like structure. The individual branches are ordered and numbered. This orderly arrangement is intended for ease of editing, such as additions, deletions, and various modifications. However, its use for cross-referencing is more difficult. A certain topic may be particularly germane to two or more widely separated branches of the tree. As it stands, unfortunately, the tree model is incapable of showing this relationship in a simple and direct way. Additional cross-reference maps must be used to reflect this interrelationship.

Figure 27 represents the terrestrial network part (item 1.1) of the network (item 1) to be modeled. Other parts are to be given on subsequent figures. The two dominant parts of the terrestrial network are its nodes (item 1.1.1) and links (item 1.1.2). Both nodes and links are further divided, as shown. If specification details are needed beyond the depth of a given tree, they can be easily added by expanding the numbering scheme. If, on the other hand, sufficient detail is reached, branches can be terminated. At that point it is essential that each element be numerically or logically defined to suit the simulator software.

Figure 28 specifies the satellite network (item 1.2) of the total network. The primary intent here is to simulate the satellite network in conjunction with a larger terrestrial network that contains many nodes and links. However, if simulation objectives were to change, separate simulations must also be possible for both stand-alone satellite and stand-alone terrestrial networks. The satellite network is shown to have four major elements: the number of satellites (item 1.2.1), their orbit types (item 1.2.2), the on-board communications assets (1.2.3), and the ground segment (item 1.2.4).

The ground segment, consisting of Earth stations (item 1.2.4.1) and gateways to terrestrial networks (item 1.2.4.2), has a dual counterpart in the terrestrial network, called the satellite gateways (item 1.1.1.3) in Figure 27 earlier. These two elements have many things in common and their separation with some "interface" may or may not be easily possible in all network situations. This is one illustration of the need for, and the problem of, cross-referencing of the tree structure employed here.

Figure 29 presents the third major component of network specification, namely the network control system (item 1.3). Network control or real-time management may employ its own, special purpose, control or signaling network (item 1.3.1), like the CCS of the AT&T long-haul network. This signaling network is further divided into links (item 1.3.1.1) and signal transfer points (STP), the latter being the switching and processing nodes of the signaling network (item 1.3.1.2). The signaling network connects widely dispersed service nodes, both of the traffic network and of the signaling network, to one or more network control centers (item 1.3.2).

The ultimate network management and decision-making power resides in these control centers. The centers can be of different types (item 1.3.2.2), and their functions (item 1.3.2.3) can be varied. To execute their tasks, the network control centers utilize various support systems (item 1.3.2.4), such as typically

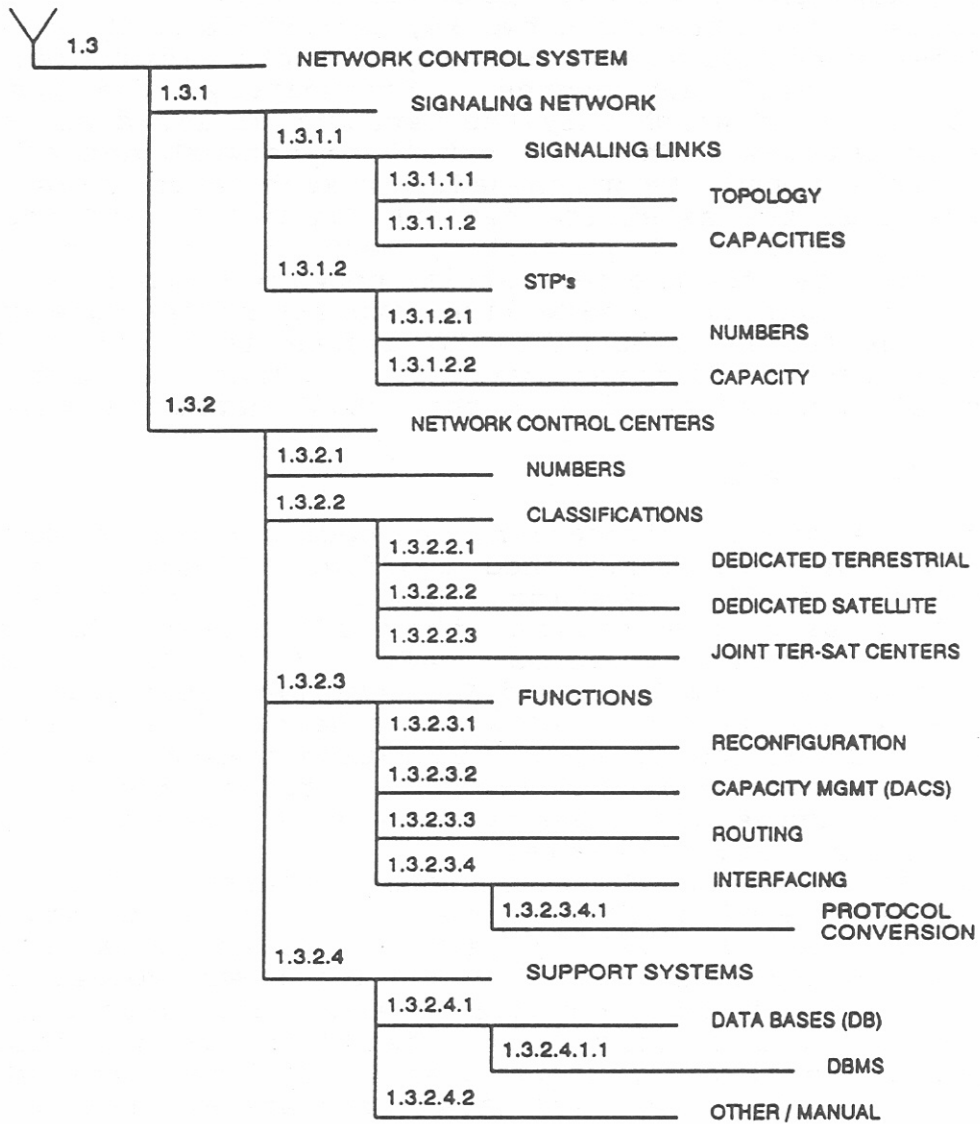


Figure 29. Specification of the simulation model:
Part 1.3, the network control system.

quite large, but rapidly accessible, data bases (item 1.3.2.4.1) and their management systems (item 1.3.2.4.1).

Simulation of stressed networks, terrestrial and/or satellite, may address scenarios where the damage affects various parts of the network to a different degree. Fortunately, the most common scenario is one in which only the terrestrial links and nodes are subject to outages. Then the satellite network and all network control systems would be undamaged. In such cases it is justified to assume that the satellite network (item 1.2) and the control systems (item 1.3) work perfectly, despite the different stress levels postulated for the terrestrial network (item 1.1). And even when both terrestrial and satellite networks suffer some stress, it may nevertheless be reasonable to assume that all control and management assets function perfectly. That is tantamount to ignoring all control details in the actual network simulation.

Traffic Specification

After network specification, the second major parameter to be defined is the circuit-switched traffic. Figure 30 provides a framework for traffic modeling (item 2). To incorporate future phases of the simulation program, the traffic is divided into three parts: the circuit switched traffic (item 2.1), the packet switched traffic (item 2.2), and ISDN traffic (item 2.3). Only the circuit-switched traffic is considered here, as the other traffic classes are delegated to later simulation phases. It is assumed that the circuit-switched traffic carries mostly telephone services. Nevertheless, some percentage of circuit-switched data messages may be permitted here.

The circuit-switched traffic model is further divided into two parts: the offered traffic (item 2.1.1), which constitutes the most significant part, and the carried traffic (item 2.1.2). The offered circuit-switched traffic can represent voice, data, and other services. While the number of end-users (item 2.1.1.2) and the total offered traffic load (item 2.1.1.3) are basic gross descriptors, equally important seem to be the detailed traffic profile (item 2.1.1.4) and the statistical traits of the constituent substreams (item 2.1.1.5). Familiar examples are the two exponential distributions that model the interarrival and holding times (see items 2.1.1.5.1 and 2.1.1.5.2, respectively) for voice traffic. For circuit switched data traffic other statistical representations may be necessary (Jagerman, 1984).

Generally, not all offered traffic is successfully carried by the network. Thus, the offered traffic can be divided into two parts: the blocked traffic and the actually carried traffic. At certain facilities, the carried traffic (item 2.1.2) can have statistical properties quite distinct from the offered traffic. Moreover, as in the case of overflow traffic (item 2.1.2.1), the statistics of the carried traffic depend on the routing algorithms and network (capacity) management regimes in force. Different methods and tools may be needed to characterize the overflow traffic in the simulator.

The traffic blocked by the network is not explicitly identified under item 2.1. The reason for this is the premise that

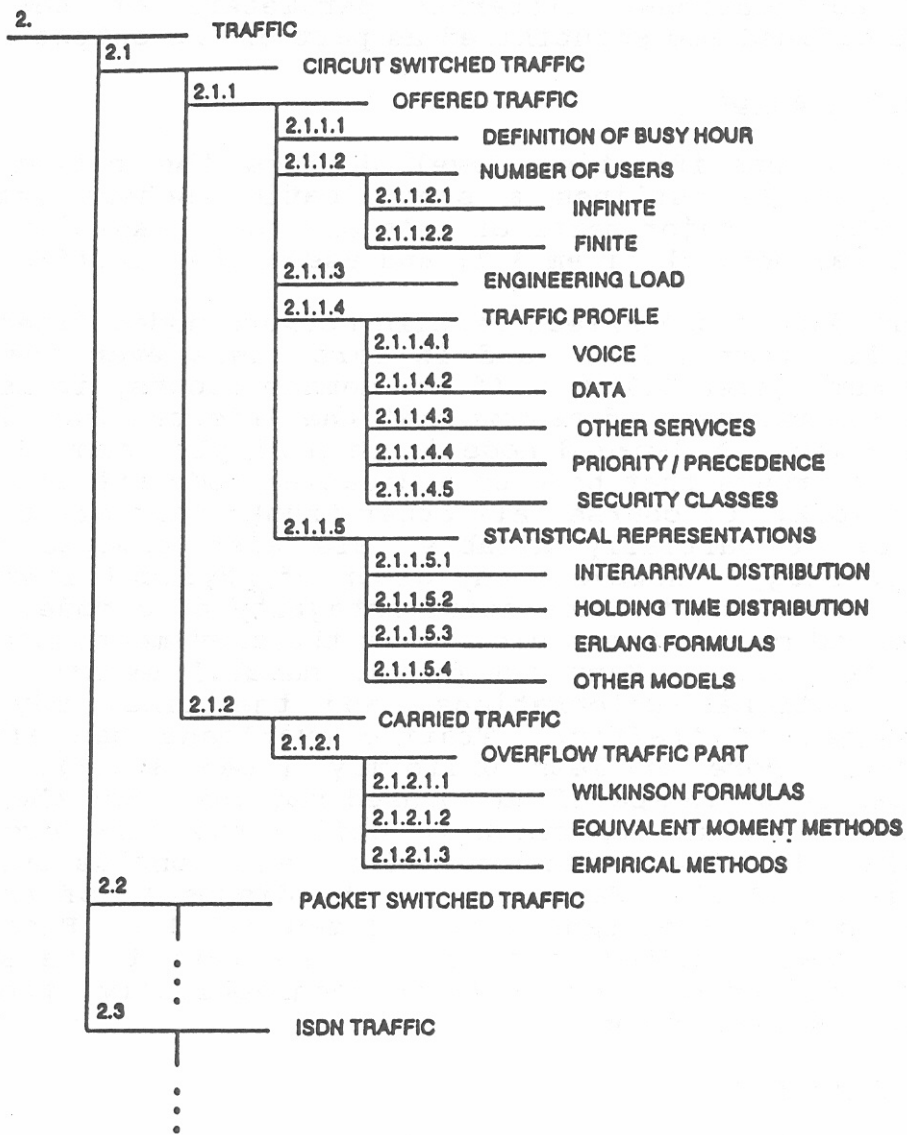


Figure 30. Specification of the simulation model:
Part 2, traffic.

blocked traffic constitutes the main output objective (or target) for circuit-switched networks. Depending on the scope of simulation applications, different parameters of the blocked traffic are defined and scrutinized as part of the output routines.

Stress Specification

The third specification level defines the network stress level. Figure 31 outlines a specification scheme for stress (item 3). The two major parts of this are the damage inflicted on the terrestrial network (item 3.1) and associated traffic overload (item 3.2).

Network damage can affect certain network nodes (item 3.1.1), certain links (item 3.1.2), and in rare cases even the network control systems (item 3.1.3). If any damage occurs, it is easiest for simulation purposes to assume that the affected facilities are totally disabled. A damaged node is then simply removed from the network. The trunks that home on a disabled node are also removed from the network. Of course, all other trunks that are themselves either fully or partially disabled are also removed from the simulated topology. However, any group of physical links can be cut without degrading the potential integrity of a node.

The second major stress element is the abovementioned traffic overload (item 3.2). Many accidents, natural events, military activities, national celebrations, and the like, may trigger abnormal surges of traffic. Traffic overloads can affect the entire network more or less uniformly (item 3.2.1). Or the overloads can occur in specified focused regions. For the study of surge dynamics, it is important to define the time profile for traffic, that is, the onset, duration, and conclusion, of the overload (item 3.2.2). Handling, or the disposal, of the stress traffic is another important factor (item 3.2.3). Finally, the volumes of crisis traffic (item 3.2.4) may have to be specified from several points of view, such as geographically, temporally, and by user classification.

Network Performance

The most frequently discussed performance parameters for simulation of circuit-switched networks are associated with blocking grade of service (GOS). The GOS parameters are used by teletraffic engineers to assess the grade or quality of service to end-users under different circumstances. The second, less popular, set of performance parameters is concerned with network facility utilization. Utilization parameters are used by network designers and managers to identify facilities that are either over- or under-utilized. Network changes are based on observed or simulated utilization numbers. A brief summary of the two performance-parameter sets follows next.

For a given "service domain," the blocking GOS is defined as the ratio of the number of blocked calls to the number of call attempts. Blocking is an act by the network. Therefore, by convention, one does not speak of a "network blocking event" when the destination terminal is either busy or fails to answer. More

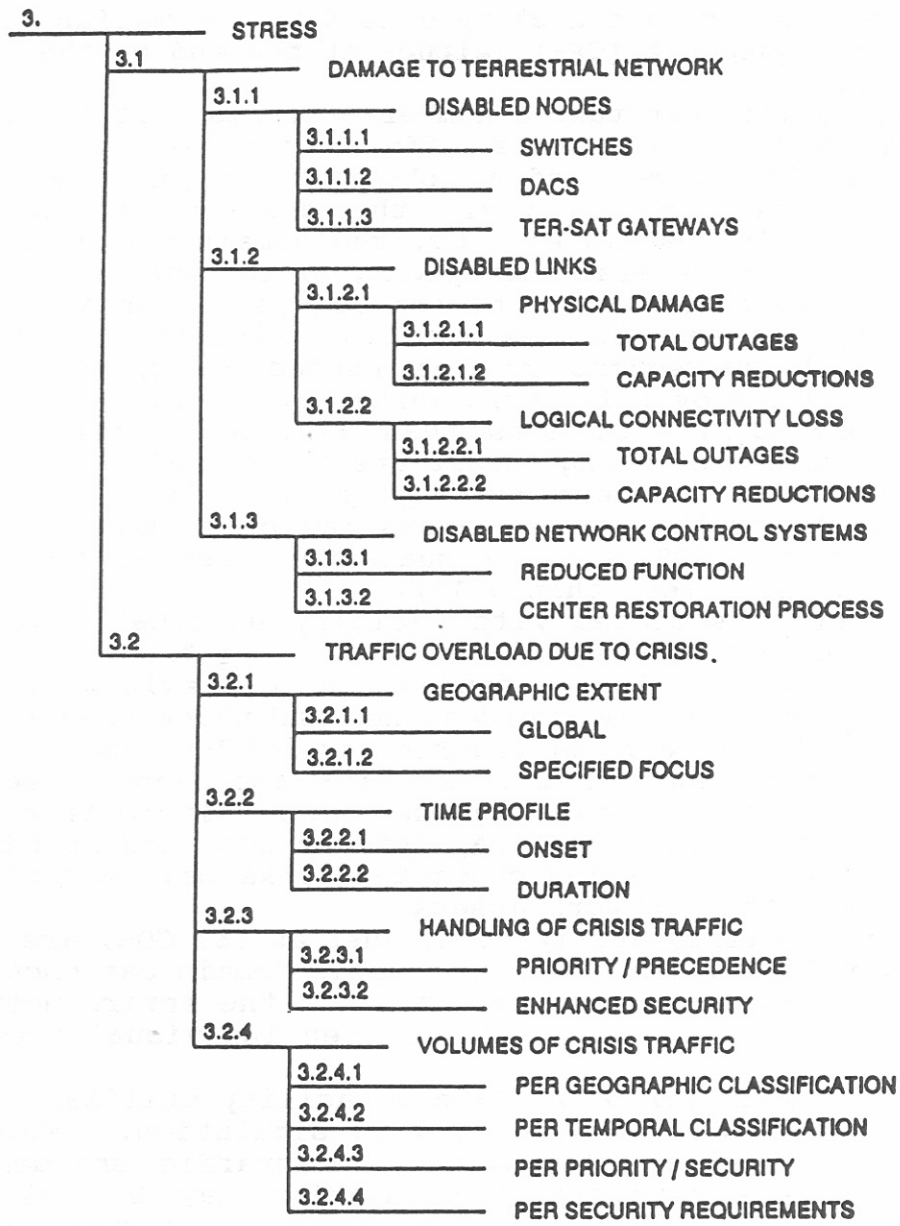


Figure 31. Specification of the simulation model: Part 3, stress.

generally, network blocking cannot be caused by any malfunction of customer premises equipment (CPE), either at the end of the called or calling party.

The service domain can take a number of forms. It may be the entire network, with steady-state offered traffic, with fixed traffic handling algorithms, and an observation (or simulation) interval that is tailored to yield the desired accuracy for probability of blocking estimates. Or, the domain can consist of specific subregions, or several subregions, of the network. At the small end, the service domain can be reduced to one or two nodes, and at its very extreme--to one or two end-users. The GOS estimates for such subregions may be needed in those network applications, where the overall performance may be quite good (such as a blocking probability of less than 2%), but where certain locations are suspected of being exceptions to the rule. These may be isolated areas with unacceptably high probability of blocking (such as over 50%), or the service areas may be of such national importance that their GOS must be guaranteed better than some critical level (perhaps less than 0.1%).

The parameters associated with facility utilization can be divided into at least two categories. First, the traffic volumes handled by nodes (such as local, tandem, or toll switches, DACS, etc.) indicate which nodes are overdesigned and which need further upgrades. Second, the Erlangs of traffic carried by links or trunk groups represent more than individual link occupancy. Because congested links lead to alternate routing, while alternate routing (depending on routing algorithms) may lead to long and roundabout paths for call completion, a few congested links may lead to very inefficient use of other network assets.

Estimates of facility utilization, just as for GOS, are based on a specific service domain. The geographic domain can encompass all the facilities (of a certain category) in the entire network. Or specific regions can be selected, or even individual nodes or trunks.

The service domain for both GOS and facility utilization can introduce a significant time dimension in simulation. When the network undergoes physical, functional, or traffic changes, as expected under stress conditions, simulation may be asked to estimate the dynamics of network performance. If the network change is a single step function, then the estimation task falls into one of three phases. See Figure 32.

Before the stress event at time $t=t_0$, the network is in a steady state. For sufficiently large T (the numerical value of T remains to be determined for specific scenarios) another steady state eventually occurs for all $t>t_0+T$. This second steady state is generally different from the first one. In the case of network damage or a traffic surge (see part (a) of Figure 32), the post-event GOS (see part (b)) and facility utilization (see part (c)) are shown to be noticeably increased. In this case, the increases represent performances that are degraded relative to the pre-event performances.

Between the two steady states, the network performance exhibits transient or dynamic behavior. In parts (b) and (c) of Figure 32 this transient behavior is shown to be approximately

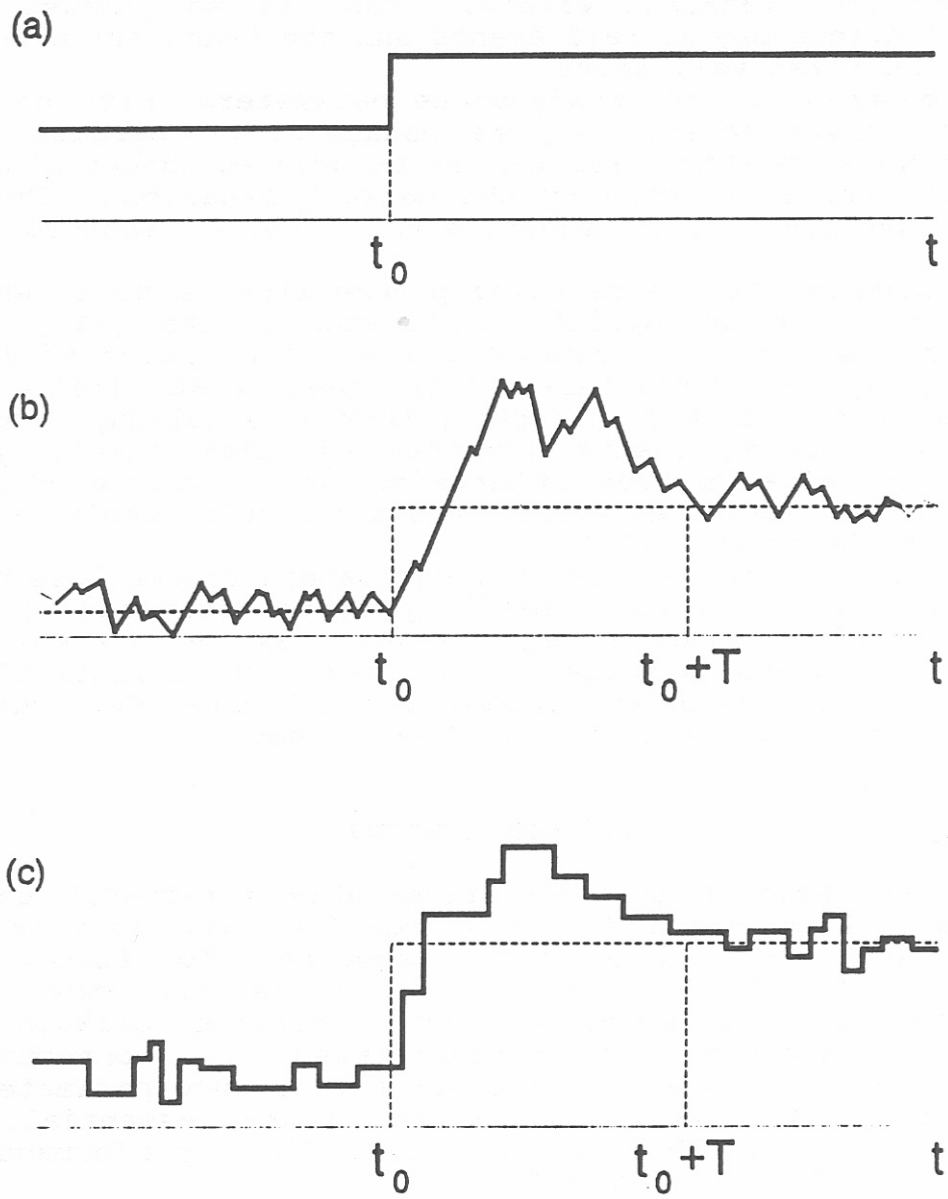


Figure 32. Network dynamics: (a) stress, (b) GOS, and (c) facility utilization.

limited to the time window $t_0 < t < t_0 + T$. From the statistical point of view, estimates and their confidence limits may be difficult to obtain for the transient window. That is so because certain congestion events may be rare events and the transient windows are likely to be relatively short.

The observation of steady-state parameters, both before and after the crisis onset, is more manageable. Because in both instances the simulation time and the associated number of observed samples are not restricted by the network dynamics. There are, however, certain basic experimental design factors to be considered.

For example, if the blocking probability is to be estimated together with its appropriate confidence limits (95%, 99%, or other), then a number of statistical sampling rules will apply. Event-counting or Monte-Carlo techniques that use either a prescribed number of call attempts (binomial sampling), prescribed numbers of blocking events (inverse binomial sampling), both bounded sample size and bounded blocking events (truncated binomial sampling), etc., have been studied and their efficiencies are known (Crow, 1974; Jeruchim, 1984).

The fact that these classical estimators become less and less efficient as the blocking probability to be estimated tends to a very small value (Guida et al., 1988), has led to specialized experimental designs (Kiemele, 1990). Pertinent illustrations are found in applications of the Extreme Value Theory (EVT) which were discussed earlier in Section 3.3 of this report.

5. CONCLUSIONS

It seems evident that any reasonably structured satellite backup facilities can offer some benefits for service outage prevention and, when outages occur, some help for faster service restoration. It is also clear that there are many complex alternatives for implementing such satellite backups. The resultant advantages and disadvantages of connectivities, capacities, recovery speed, and other performance parameters need to be ascertained. Likewise, and perhaps more essential, may be the need to estimate the associated tradeoffs of performance gains versus implementation costs.

However, cost appears to be such a difficult and moving target, that its analysis is premature at this time. The intent here is to define an approach that first addresses the technical performance characteristics of the satellite-terrestrial hybrid. Only after the technical options have been defined, their parameters delimited, and their implementation plus performance brought into a clearer focus, can one hope to demonstrate meaningful cost numbers.

This report reviews the often-acknowledged claim that the existing terrestrial U.S. networks and their services are vulnerable to many potential threats. Adaptive satellite networks are proposed as possible backup systems. Yet, their definition and design is far from clear. The problem appears to be too complex for a meaningful, direct, and credible analytical solution. At the

same time, there exist a number of modeling and simulation methodologies and computer tools. The numbers and capabilities of these simulation tools are increasing rapidly. Consequently, this report advocates the use of computer modeling and simulation systems to evaluate the potential characteristics, both advantages and disadvantages, of combined terrestrial-satellite networks.

As an illustration of a candidate approach, Section 4 discusses simulation of large, circuit-switched networks. The single most significant performance parameter for such networks is the blocking grade-of-service (GOS). To assess the GOS performance of large terrestrial and/or satellite networks in a shortened simulation time, a temporal aggregation method is proposed. Implementation of the main simulator functions is shown in a number of functional diagrams. In this method, network time is quantized in fixed length intervals. Random numbers of call arrivals and call deletions are generated for each sample interval by processing and routing a relatively large number of call events as a group. Sets of calls that cannot be successfully routed are declared as blocked. They contribute to the GOS output statistics.

Thus, the method illustrated here cannot be called an event-by-event simulation. Instead, it is a group-by-group simulation method, where the term "group" refers to the random set or group of calls handled together. The size of the average random group determines the effective speed advantage of the proposed time aggregation methodology.

6. REFERENCES

- Abramowitz, M., and I. A. Stegun (1964), Handbook of Mathematical Functions, National Bureau of Standards Applied Mathematics Series No. 55 (U.S. Government Printing Office, Washington, DC).
- Agarwal, Y. K. (1989), An algorithm for designing survivable networks, AT&T Tech. Journal 68 (May/June), pp. 64-76.
- Akimaru, H., A. Kagechika, and H. Takahashi (1986), Mean and variance of overflow traffic for time dependent inputs, IEEE Trans. on Commun. COM-34, pp. 238-243.
- Ash, G. R., and S. D. Schwartz (1990), Traffic control architectures for integrated broadband networks, Int. J. of Digital and Analog Commun. Systems 3, pp. 167-176.
- Austin, G.P., B.S. Bosik, and C.J. Capece (1989), The universal port concept, AT&T Tech. Journal 68 (March/April), pp. 14-22.
- Balaban, P., K. S. Shanmugan, and B. W. Stuck (1984), Computer-aided modeling, analysis, and design of communication systems: Introduction and issue overview, IEEE J. Selected Areas in Commun. SAC-2, pp. 1-8.

- Bell, T. E. (1990), Technical challenges to a decentralized phone system, *IEEE Spectrum* 27 (September), pp. 32-37.
- Bell, T. E., and G. Zorpette (1991), Phone cable cut slashes airline, finance activity, *The Institute - A News Supplement for IEEE Spectrum* (March).
- Benes, V. E. (1957), A sufficient set of statistics for a simple telephone exchange model, *BSTJ* 36, pp. 939-973.
- Berberana, I. (1990), Application of extreme value theory to the analysis of a network simulation, *Proc. of Annual Simulation Symposium*, pp. 105-121.
- Bharath-Kumar, K., and P. Kermani (1984), Performance Evaluation Tool (PET): An analysis tool for computer communication networks, *IEEE J. Selected Areas in Commun.* SAC-2, pp. 220-226.
- Biersack, E. W. (1990), Annotated bibliography on network interconnection, *IEEE J. Selected Areas in Commun.* SAC-8, pp. 22-41.
- Boensch, C. J., and R. J. Sogegian (1989), Portable telecommunications for National Security Preparedness, *Signal* (July), pp. 51-55.
- Braun, A. (1989), Telephone service continues through earthquake, *Telecommunications* (December), p. 44.
- Bridges, J. B., and S. B. Sen (1990), Taking a fresh look at traffic engineering, *Bellcore Exchange* (January/February), pp. 14-19.
- Brush, G., and N. Marlow (1990), Assuring the dependability of telecommunications networks and services, *IEEE Network Mag.* 4 (January), pp. 29-34.
- Butto, M., I. Pilloni, and C. Scarati (1989), POISSON: Procedure for the optimal insertion of a switching satellite in an operative network, *CSELT Technical Reports (Italy)* XVIII (June), pp. 213-218.
- Cain, J. B., J. W. Nieto, M. D. Noakes, and E. L. Althouse (1989), A class of adaptive routing and link assignment algorithms for large-scale networks with dynamic routing, *Conference Record of MILCOM '89 (Boston, MA)*, pp. 38.2.1-38.2.6.
- Cassandras, C. G., and S. G. Strickland (1988), Perturbation analytic methodologies for design and optimization of communication networks, *IEEE J. Selected Areas in Commun.* SAC-6, pp. 158-171.

- CCITT (1989a), Integrated Services Digital Network (ISDN) - Overall network aspects and functions, user-network interface, Blue Book, Vol. III, Fascicle III.8, Recommendations I.310-I.470, Geneva, Switzerland.
- CCITT (1989b), Data communication networks: Open Systems Interconnection (OSI) - Model and notation, service definition, Blue Book, Vol. VIII, Fascicle VIII.4, Recommendations X.200-X.219, Geneva, Switzerland.
- CCITT (1989c), User demand, Blue Book, Vol. II, Fascicle II.3, Recommendation E.711, Geneva, Switzerland.
- CCITT (1989d), Functional specification and description language (SDL), Blue Book, Vol. X, Fascicle X.1, Recommendation Z.100, Geneva, Switzerland.
- Chang, F. (1989), Routing-sequence optimization for circuit switched networks, AT&T Tech. Journal 68, pp. 57-63.
- Chiarawongse, J., M. M. Srinivasan, T. J. Teorey (1988), Performance analysis of a large interconnected network by decomposition techniques, IEEE Network Mag. 2 (July), pp. 19-27.
- Coates, R. F. W., G. J. Janacek, and K. V. Lever (1988), Monte Carlo simulation and random number generation, IEEE J. Selected Areas in Commun. SAC-6, pp. 58-66.
- Cochran, W. G., and G. M. Cox (1957), Experimental Designs (John Wiley & Sons, New York, NY).
- Crow, E. L. (1974), Confidence limits for digital error rates, OT Report 74-51, NTIS Order No. COM 751 0793.
- Cruz, G. C., R. S. Hisiger, and R. S. Wolff (1989), Strategic telecommunications network planning in the context of emerging technologies, architectures, and services, IEEE J. Selected Areas in Commun. SAC-7, pp. 1198-1206.
- Doner, J. A. (1988), GENESIM, IEEE J. Selected Areas in Commun. SAC-6, pp. 172-179.
- Dunn, D. A., and M. G. Johnson (1989), Demand for data communications, IEEE Network Mag. 3 (May), pp. 8-12.
- Edwards, M. (1991), Get T1's clout at a fraction of the cost, Communications News (April), pp. 22-25.
- FCC (1988), Assignment of orbital locations to space stations in the domestic fixed-satellite service, Memorandum Opinion and Order of the Federal Communications Commission, Washington, DC.

- Frost, V. S., W. W. Larue, and K. S. Shanmugan (1988), Efficient techniques for the simulation of computer communications networks, IEEE J. Selected Areas in Commun. SAC-6, pp. 146-157.
- Garzia, R. F., and M. R. Garzia (1990), Network Modeling, Simulation, and Analysis (Marcel Dekker, Inc., New York, NY).
- Garzia, M. R. (1990), A study of network adaptive routing. Published as part of Garzia, R. F., and M. R. Garzia, Network Modeling, Simulation, and Analysis (Marcel Dekker, Inc., New York, NY), pp. 211-235.
- Garzia, M. R., and C. M. Lockhart (1989), Nonhierarchical communications networks: An application of compartmental modeling, IEEE Trans. on Commun. COM-37, pp. 555-564.
- Garzia, M. R., and C. M. Lockhart (1990), Modeling network dynamics. Published as part of Garzia, R. F., and M. R. Garzia, Network Modeling, Simulation, and Analysis (Marcel Dekker, Inc., New York, NY), pp. 237-294.
- Gifford, L. F. (1987), Adaptive routing and traffic control in damaged circuit switched networks, Conference Record of MILCOM '87, Washington, DC, pp. 1.2.1-1.2.6.
- Gimpelson, L. A., and J. H. Weber (1964), UNISIM - A simulation program for communications networks, Proc. of Fall Joint Computer Conference, pp. 233-249.
- Girard, A., and M. A. Bell (1989), Blocking evaluation for networks with residual capacity adaptive routing, IEEE Trans. on Commun. COM-37, pp. 1372-1380.
- Groenbaek, I. (1986), Conversion between the TCP and ISO transport protocols as a method of achieving interoperability between data communications systems, IEEE J. Selected Areas in Commun. SAC-4, pp. 288-296.
- Grover, W. D., B. Venables, J. H. Sandham, and A. F. Milne (1990), Performance studies of a self-healing network protocol in Telecom Canada long-haul network. T1M1.3 Working Group document from Alberta Telecomm. Research Centre (Edmonton, Alberta, Canada).
- Guida, M., D. Iovino, and M. Longo (1988), Comparative performance analysis of some extrapolative estimators of probability tails, IEEE J. Selected Areas in Commun. SAC-6, pp. 76-84.
- Gumbel, E. J. (1958), Statistics of Extremes (Columbia University Press, New York, NY).
- Hamming, R. W. (1962), Numerical Methods for Scientists and Engineers (McGraw-Hill Book Co., New York, NY).

- Harbison, S. P., and G. L. Steele (1987), C: A Reference Manual (Prentice-Hall, Englewood Cliffs, NJ).
- Hardy, G. H., J. E. Littlewood, and G. Polya (1964), Inequalities (Cambridge at the University Press, Cambridge, UK).
- Ilyas, M., and H. T. Mouftah (1985), Performance evaluation of computer communications networks, IEEE Commun. Mag. 23 (April), pp. 18-29.
- Jagerman, D. L. (1975), Nonstationary blocking in telephone traffic, BSTJ 54, pp. 625-661.
- Jagerman, D. L. (1984), Methods in traffic calculations, AT&T Bell Labs. Tech. Journal 63, pp. 1283-1310.
- Jeruchim, M. C. (1976), On the estimation of error probability using generalized extreme-value theory, IEEE Trans. Information Theory IT-22, pp. 108-110.
- Jeruchim, M. C. (1984), Techniques for estimating the bit error rate in the simulation of digital communication systems, IEEE J. Selected Areas in Commun. SAC-2, pp. 153-170.
- Joel, A. E. (1982), A History of Engineering and Science in the Bell System: Switching Technology 1925 - 1975 (Bell Telephone Laboratories, Inc., Murray Hill, NJ).
- Kato, S., M. Morikura, S. Kubota, K. Enomoto, and M. Umehira (1990), TDMA equipment for DYANET, NTT Review (Japan) 2, No. 3, pp. 47-54.
- Katz, S. S. (1967), Statistical performance analysis of a switched communications network, Fifth International Teletraffic Congress, New York, NY, pp. 566-575.
- Kaudel, F. J. (1989), Proposal for a framework of network survivability. TlQ1 submission from Northern Telecom. Inc., Orangeburg, NY (October).
- Kearns, T. J., and M. C. Mellon (1990), The role of ISDN signaling in global networks, IEEE Commun. Mag. 28 (July), pp. 36-43.
- Kempthorne, O. (1952), The Design and Analysis of Experiments (John Wiley & Sons, New York, NY).
- Kernighan, B., and D. Ritchie (1978), The C Programming Language (Prentice-Hall, Englewood Cliffs, NJ).
- Kiemele, M. J. (1990), Integrated network performance analysis. Published as part of Garzia, R. F., and M. R. Garzia, Network Modeling, Simulation, and Analysis (Marcel Dekker, Inc., New York, NY), pp. 309-328.

- Kinzie, R. W. (1989), Outer space - A good place to keep your communications, Via Satellite (March), pp. 39-41.
- Kleijnen, J. P. C. (1979), The role of statistical methodology in simulation. Published in Methodology in Systems Modelling and Simulation, B. P. Zeigler, M. S. Elzas, G. J. Klir, and T. I. Oren, Eds. (North-Holland Press, Amsterdam, Netherlands).
- Kleinrock, L. (1975), Queueing Systems, Volume I: Theory (John Wiley & Sons, New York, NY).
- Knuth, D. E. (1968), The Art of Computer Programming, Vol. 1 Fundamental Algorithms (Addison-Wesley, Reading, MA).
- Kochan, S. G. (1983), Programming in C (Hayden Book Co., Hasbrouck Heights, NJ).
- Koval, D.O., H. K. Kua, and K. H. Cha (1986), Computer modelling and reliable evaluation of network operational paths, Seventh Conference on Electronic Systems, pp. 530-537.
- Kurose, J. F., and H. T. Mouftah (1988), Computer-aided modeling, analysis, and design of communication networks, IEEE J. Selected Areas in Commun. SAC-6, pp. 130-145.
- Linfield, R. F. (1990), Congestion-reduction and service-restoration strategies for telecommunication networks, NTIA Report 90-257, NTIS Order No. PB90-207838/AS.
- Lippmann, R. P. (1982), Steady state performance of survivable routing procedures for circuit-switched mixed-media networks, Technical Report 633, MIT Lincoln Laboratory, Lexington, MA.
- Lockhart, C. M. (1990), Design and performance analysis of survivable networks. Published as part of Garzia, R. F., and M. R. Garzia, Network Modeling, Simulation, and Analysis (Marcel Dekker, Inc., New York, NY), pp. 185-209.
- Mathis, V. K. (1989), GENSIM: An interactive discrete event simulator for telephone networks, Advances in AI and Simulation, Society for Computer Simulation, pp. 252-257.
- Merschtina, B. (1987), Packet switching for high-performance data transmission, AT&T Technology 2 (1), pp. 40-47.
- Mills, R. (1987), Statistical Analysis of Steady State Simulation Output Data with SIMSCRIPT II.5, CACI Products Co. Publications, La Jolla, CA.
- Mitra, D., and J. B. Seery (1991), Comparative evaluation of randomized and dynamic routing strategies for circuit-switched networks, IEEE Trans. on Commun. COM-39, pp. 102-116.

- Modarressi, A. R., and R. A. Skoog (1990), Signaling System No. 7: A tutorial, IEEE Commun. Mag. 28 (July), pp. 19-36.
- Moretti, R. (1990), SDL and object oriented design: A way for producing quality software, CSELT Tech. Reports, Torino, Italy, XVIII (April), pp-131-134.
- Morihiro, Y., S. Okasaka, H. Shiota, and S. Ueno (1990), A new satellite communication system for public telecommunications networks - DYANET, NTT Review (Japan) 2, No. 3, pp. 12-19.
- Mouftah, H. T., and S. Bhatia (1984), Design tradeoffs for local access systems in computer networks, IEEE J. Selected Areas in Commun. SAC-2, pp. 264-277.
- Mouftah, H. T., and K. S. Shanmugan (1987), Computer aided techniques for communications systems engineering, IEEE Commun. Mag. 25 (July), pp. 48-54.
- Mouldin, R., S. Adams, G. Harrick, B. Demeyer, and J. Hardy (1989), A new path metric for survivable circuit switched routing, Conference Record of MILCOM '89, Boston, MA, pp. 38.5.1-38.5.5.
- Naderi, F. M., and S. J. Campanella (1988), NASA's advanced communications technology satellite (ACTS), Proc. of AIAA 12th International Commun. Satellite Conf., Arlington, VA, pp. 204-224.
- Nakashima, H., B. Nishimoto, M. Nakayama, and Y. Arai (1990), Satellite channel control unit for DYANET, NTT Review (Japan) 2, No. 3, pp. 27-34.
- NCS (1988a), PSN and NETS network level EMP effects evaluations: Sensitivity of Northern Telecom switches, Automatic Electric switches and fiber optic transmission facilities, National Communications System, Washington, DC, March. [Contains proprietary AT&T data.]
- NCS (1988b), The effects of high-altitude electromagnetic pulse (HEMP) on telecommunications assets, Tech. Information Bulletin 88-3, National Communications Systems, Washington, DC, June.
- NCS (1989a), Traffic congestion analysis on a circuit-switched telecommunications network, National Communications System, Washington, DC, January.
- NCS (1989b), Network connectivity analysis model (NCAM) user's manual, National Communications System, Washington, DC, (March).

- Nesenbergs, M. (1989), Stand-alone terrestrial and satellite networks for nationwide interoperation of broadband networks, NTIA Report 89-253, NTIS Order No. PB90-172404/AS.
- Newport, K. T., and P. K. Varshney (1989), On the design of performance-constrained survivable networks, Conference Record of MILCOM '89, Boston, MA, pp. 38.1.1-38.1.7.
- NRC (1986), The Policy Planning Environment for National Security Environment, Final Report to the National Communications System by the Committee on National Security Telecommunications Policy Planning Environment, National Research Council (National Academy Press, Washington, DC).
- NRC (1989), Growing Vulnerability of the Public Switched Networks: Implications for National Security Emergency Preparedness (National Academy Press, Washington, DC).
- Ohnuki, M., K. Nagayama, and Y. Inada (1990), DYANET satellite communication transit switching system, NTT Review (Japan) 2, No. 3, pp. 20-26.
- O'Reilly, P. J. P., and J. L. Hammond (1984), An efficient simulation technique for performance studies of CSMA/CD local networks, IEEE J. Selected Areas in Commun. SAC-2, pp. 238-249.
- Palmer, L. C., and L. W. White (1990), Demand assignment in the ACTS LBR system, IEEE Trans. on Commun. 38, pp. 684-692.
- Park, S. K., and K. W. Miller (1988), Random number generators: Good ones are hard to find, Commun. ACM 31, pp. 1192-1201.
- Patacchini, A., and F. M. Galante (1987), The roles of satellites in the ISDN. Published as part of New Systems and Services in Telecommunications, III, Cantraine, G., and J. Destine, Eds. (Elsevier Science Publishers, North Holland).
- Pawlikowski, K. (1990), steady-state simulation of queueing processes: A survey of problems and solutions, ACM Computing Surveys 22, pp. 123-170.
- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1988), Numerical Recipes in C (Cambridge University Press, Cambridge, UK).
- Ratz, H. C. (1988), Extreme value engineering for local network traffic, IEEE Trans. on Commun. COM-36, pp. 1302-1308.
- Richters, J. S., and C. A. Dvorak (1988), Framework for defining the quality of communications services, IEEE Commun. Mag. 26 (October), pp. 17-23.

- Robrock, R. B. (1991), The intelligent network--Changing the face of telecommunications, Proc. IEEE 79, pp. 7-20.
- Sanchez, J. J. (1988), SKYNET satellite service - New choices for data networking, AT&T Technology 3 (4), pp. 16-23.
- Saracco, R., and P. A. J. Tilanus (1987), CCITT SDL: Overview of the language and its applications, Computer Networks and ISDN Applications (Elsevier Science Publishers, North Holland), pp. 65-74.
- Sauer, C. H., E. A. McNair, and J. F. Kurose (1984), Queueing network simulations of computer communications, IEEE J. Selected Areas in Commun. SAC-2, pp. 203-220.
- Schroeder, M. A., and K. T. Newport (1989), Enhanced network survivability through balanced resource criticality, Conference Record of MILCOM '89, Boston, MA, pp. 38.4.1-38.4.5.
- Schruben, L. W. (1987), Using simulation to solve problems: A tutorial on the analysis of simulation output, Proc. of ACM/IEEE 1987 Winter Simulation Conference, Atlanta, GA, pp. 40-42.
- Shanmugan, K. S., P. Titchener, and W. Newman (1989), Simulation-based CAAD tools for communication and signal processing systems, Proc. of ICC '89, Boston, MA.
- Shannon, R. E. (1981), Test for the verification and validation of computer simulation models, Proc. of ACM/IEEE 1981 Winter Simulation Conference, Atlanta, GA, pp. 573-577.
- Siemens (1970), Telephone Traffic Theory Tables and Charts (Siemens AG, Muenchen, Germany), Second Edition, Part 1.
- Simulation (1987), Catalog of simulation software, Simulation 49, pp. 165-181.
- Spirn, J. R., J. Chien, and W. Hawe (1984), Bursty traffic local network modeling, IEEE J. Selected Areas in Commun. SAC-2, pp. 250-258.
- St. Jacques, M., and D. Stevens (1989), Simulation of telephone traffic for a real-time network control expert system, Advances in AI and Simulation, Society for Computer Simulation, pp. 243-249.
- SunGard (1989), Satellite test proves out T1 channel for long-haul disaster recovery, Network Management (September), p. 30.
- Sunshine, C. A. (1990), Network interconnection and gateways, IEEE J. Selected Areas in Commun. SAC-8, pp. 4-11.

- Svobodova, L. (1989), Implementing OSI systems, IEEE J. Selected Areas in Commun. SAC-7, pp. 1115-1130.
- Timko, J. W. (1987), AT&T systems architecture, AT&T Technology 2 (3), pp. 4-13.
- Waite, M., D. Martin, and S. Prata (1983), Unix Primer Plus (Howard Sams & Co, Indianapolis, IN).
- Weinstein, S. B. (1973), Theory and application of some classical and generalized asymptotic distributions of extreme values, IEEE Trans. on Information Theory IT-19, pp. 148-154.
- Wilks, S. S. (1963), Mathematical statistics (John Wiley & Sons, New York, NY).
- Wilkinson, R. I. (1971), Some comparisons of load and loss data with current teletraffic theory, BSTJ. 50, pp. 2807-2834.
- Williams, R.H. (1991), Iridium offers contact to any point on Earth, Signal 45 (February), pp. 95-97.
- Wilson, J. R., and A. A. B. Pritsker (1978), Evaluation of startup policies in simulation experiments, SIMULATION 31, pp. 79-89.
- Wolf, J. J., and B. Ghosh (1988), Simulation and analysis of Very Large Area Networks (VLAN) using an information flow model, IEEE Network Mag. 2 (July), pp. 5-18.
- Wright, D.L., J.R. Balombin, and P.Y. Sohn (1990), Advanced communications technology satellite (ACTS) and potential system applications, Proc. IEEE 78, pp. 1165-1175.
- Wu, T. H., D. J. Kolar, and R. H. Cardwell (1988), Survivable network architectures for broad-band fiber optic networks: Model and performance comparison, IEEE J. on Lightwave Technology LT-6, pp. 1698-1709.
- Yokoi, T., and K. Kodaira (1989), Grade of service in the ISDN era, IEEE Commun. Mag. 27 (April), pp. 46-50.
- Yunus, M. N. (1987), Approximation for mean of overflow traffic with discrete time-dependent input, Proc. IEEE 75, pp. 1536-1537.
- Zorpette, G. (1989), Keeping the phone lines open, IEEE Spectrum 26 (June), pp. 32-36.

BIBLIOGRAPHIC DATA SHEET

	1. PUBLICATION NO. NTIA Report 91-281	2. Gov't Accession No.	3. Recipient's Accession No.
4. TITLE AND SUBTITLE Simulation of Hybrid Terrestrial-Satellite Networks for Service Restoral and Performance Efficiency		5. Publication Date November 1991	6. Performing Organization Code NTIA/ITS.N1
7. AUTHOR(S) Martin Nesenbergs		9. Project/Task/Work Unit No.	
8. PERFORMING ORGANIZATION NAME AND ADDRESS National Telecommunications and Information Admin. Institute for Telecommunication Sciences 325 Broadway Boulder, CO 80303-3328		10. Contract/Grant No.	
11. Sponsoring Organization Name and Address National Telecommunications and Information Admin. Herbert C. Hoover Building 14th and Constitution Avenue, NW Washington, DC 20230		12. Type of Report and Period Covered	
14. SUPPLEMENTARY NOTES		13.	
15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.) <p>Motivated by recognized vulnerabilities of the terrestrial public networks, this report addresses the question whether an appropriate introduction of advanced satellite systems would or would not benefit the telecommunication services for the currently existing terrestrial infrastructure. What the satellite subnetwork should be, and what performance gains are to be realized, are two key issues. Given voice, data or integrated services traffic, the survivability and restoral effectiveness of different network configurations is likely to vary considerably for different crisis scenarios. It is concluded that answers to these and other complex, performance-related, questions can only be gotten by means of computer <u>modeling and simulation</u>. Today there seem to be sufficient simulation tools available for the task. The report reviews the overall plan and simulation objectives for circuit-switched networks. From the many proposed methodologies, the discrete event and temporal aggregation methods are emphasized. The importance of simulator inputs is demonstrated through needs for relatively detailed specifications of the terrestrial and satellite networks, their interfaces, the offered traffic, stress (i.e., network damage and traffic overload), and service performance measures required in the simulation output.</p> <p>Key words: advanced satellites; blocking grade-of-service; circuit-switched networks; modeling; network damage; performance; satellite-terrestrial hybrid; service restoral; simulation; stress; temporal aggregation; traffic</p>			
17. AVAILABILITY STATEMENT <input checked="" type="checkbox"/> UNLIMITED. <input type="checkbox"/> FOR OFFICIAL DISTRIBUTION.		18. Security Class. (This report) Unclassified	20. Number of pages 119
		19. Security Class. (This page) Unclassified	21. Price: