

Interim Federal Standard 1033 Reference Manual

N.B. Seitz



U.S. DEPARTMENT OF COMMERCE
Philip M. Klutznick, Secretary

Henry Geller, Assistant Secretary
for Communications and Information

December 1980

PREFACE

This report is the first of a planned two-report series designed to assist users in understanding and applying Interim Federal Standard 1033, "Telecommunications: Digital Communication Performance Parameters." This first volume, the Interim Federal Standard 1033 Reference Manual, outlines potential benefits of the standard, summarizes its objectives and content, and provides a tutorial "essay" on the meaning and importance of each standard parameter. Its sequel, the Interim Federal Standard 1033 Application Manual, will provide guidelines for applying the standard in user requirements analysis, service performance specification, and service selection; and will illustrate the use of these guidelines in a representative system development example.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
ABSTRACT	1
1. INTRODUCTION	1
1.1 Background	1
1.2 Purpose and Scope of Report	4
2. THE BENEFITS	5
2.1 Introduction	5
2.2 Existing Procurement Practices	6
2.3 Sources of Inefficiency	8
2.4 FED STD 1033 Benefits	14
3. FED STD 1033 EXPLAINED	18
3.1 Introduction	18
3.2 Key Technical Problems	18
3.3 FED STD 1033 Approach	24
4. UNDERSTANDING THE PARAMETERS	53
4.1 Introduction	53
4.2 Access Parameters	55
4.3 User Information Transfer Parameters	67
4.4 Disengagement Parameters	97
4.5 Secondary Parameters	102
4.6 Ancillary Parameters	107
4.7 Summary	116
5. ANNOTATED REFERENCES	118

LIST OF FIGURES

	Page
Figure 2-1. Service specification form.	11
Figure 2-2. Existing vs functional procurement.	15
Figure 3-1. System dependence.	19
Figure 3-2. Detailed parameter definition.	21
Figure 3-3. User dependence.	23
Figure 3-4. Parameter development overview.	25
Figure 3-5. User/system interfaces.	27
Figure 3-6. Reference event concept.	29
Figure 3-7. Aggregate user concept.	33
Figure 3-8. Block transfer function definition.	37
Figure 3-9. Access outcome definition.	40
Figure 3-10. Block transfer outcomes.	41
Figure 3-11. Disengagement outcomes.	43
Figure 3-12. Outcome summary.	44
Figure 3-13. Access parameters.	45
Figure 3-14. Secondary parameter development.	48
Figure 3-15. Ancillary parameter development - access example.	51
Figure 3-16. FED STD 1033 parameters.	54
Figure 4-1. Truncation of the access time distribution.	58
Figure 4-2. Bit and block transfer rate definitions.	86
Figure 4-3. Relationship between rate efficiency and error control block length.	95
Figure 4-4. Impact of user input pattern on user message transfer time fraction.	113
Figure 4-5. Example service performance specification form.	117

INTERIM FEDERAL STANDARD 1033 REFERENCE MANUAL

Neal B. Seitz*

There is a growing need within the Federal government for a user-oriented, system-independent, functional means of specifying data communication performance. A recently published Federal Standard, Interim Federal Standard 1033, defines a set of standard performance parameters designed to meet that need. This report is basically an explanation and elaboration of that standard. The report first outlines the need for the standard, and the potential benefits of its use, from the viewpoint of the end user, the communication supplier, and the communication manager. The report then summarizes the objectives and content of the standard in informal, nontechnical terms. Finally, the report examines the meaning and importance of each standard parameter in a series of tutorial parameter "essays." Typical parameter values are presented, and design implications are discussed.

1. INTRODUCTION

1.1 Background

On March 8, 1979, the Federal Telecommunications Standards Committee (FTSC) voted to approve publication of a new Federal Standard: Interim Federal Standard 1033, "Digital Communication Performance Parameters". The purpose of the new standard is stated in its opening paragraph:

"to improve Federal government procurement of digital telecommunication systems and services by providing user-oriented, system-independent means of specifying communication performance."

The essence of the FED STD 1033 approach is summed up in the phrase "user-oriented, system-independent". The FED STD 1033 parameters focus on user performance concerns rather than engineering design considerations; they describe end-to-end services rather than particular system facilities; and they apply to all systems, irrespective of transmission medium, network topology, or control protocol. These standard parameters will improve Federal data communication procurement by providing a common framework for functional specification and top-down design; and will promote competition and innovation in the data communications industry by simplifying performance comparison.

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

The FTSC approval of FED STD 1033 culminated over 5 years of development, coordination, and review efforts involving more than a dozen Federal agencies. The National Telecommunications and Information Administration's Institute for Telecommunication Sciences (NTIA/ITS) chaired the FTSC's Telecommunication Performance Standards subcommittee, and had overall responsibility for developing the standard (Seitz and McManamon, 1978). The National Communications System (NCS) organized and chaired the FTSC, and played a key role in coordinating the standard within the Federal government and with industry. The National Bureau of Standards contributed importantly to the subcommittee work, and provided early contacts with non-FTSC standards organizations, notably the American National Standards Institute (ANSI). The General Services Administration provided guidance on Federal implementation of the standard and was responsible for its formal publication (GSA, 1979). Other Federal contributors included the National Aeronautics and Space Administration, the Library of Congress, and the Defense Communications Engineering Center.

The FED STD 1033 development effort also benefited substantially from various nongovernment standards activities. Over a dozen industry organizations participated in the FCC's specialized common carrier Quality and Reliability inquiry (FCC, 1975); and their recommendations were summarized in an ITS report which provided a technical foundation for the standard (Seitz and McManamon, 1976). The FTSC maintained close coordination with CCITT and ISO groups addressing performance issues during development of the standard (e.g., CCITT, 1977; ISO, 1978); and the draft standard was aligned with international recommendations on a number of topics. NCS solicited direct public comments on the standard on two separate occasions (NCS, 1977; NCS, 1978); and these comments had a substantial impact on the final document.

Perhaps the most important non-Federal contribution to the standard was that of ANSI's Data Communication Performance Task Group, X3S35. That group has been working in standard performance assessment for over 15 years, and has produced two protocol-based ANSI performance standards (ANSI, 1974; ANSI, 1980). Both standards were valuable precedents for FED STD 1033, and the Task Group's detailed review of the draft Federal Standard substantially improved its clarity. Task Group X3S35 and the FTSC's Telecommunication Performance Standards subcommittee essentially united in 1979, with the objective of adapting and refining FED STD 1033 for proposal as an American National Standard. That work is currently under way.

In approving FED STD 1033, the FTSC also adopted an important qualifying provision - that the standard be designated as an interim Federal Standard. The "interim" designation identifies a Federal Standard as an evolving standard,

typically undergoing initial application trials; and makes its use by Federal agencies optional. Non-interim federal standards typically are more established, and are mandatory for use by all Federal agencies. Interim federal standards are normally converted to mandatory federal standards after the necessary application trials (and any necessary revisions) are completed.

A number of factors led the FTSC to conclude that the interim step was necessary in the case of FED STD 1033. The first was the relative novelty of the FED STD 1033 approach. The standard places the user/system interfaces well outside the traditional DTE/DCE¹ boundaries; defines the performance parameters in terms of general "reference events" rather than system-specific interface signals; and treats the user and the system as co-responsible entities who jointly determine overall communication performance. These departures from convention clearly suggested the need for a user familiarization period.

A second reason for the interim decision was the substantial impact the standard will have on existing Federal procurement practices. FED STD 1033 requires (and enables) a functional approach to communication procurement, in which the end user needs are specified without presupposing any particular system design. As discussed in Section 2, such an approach is far from realization in many Federal organizations today. The interim period will allow time for necessary procedural changes to occur gradually, in a natural, evolutionary way.

A final reason for the interim decision was the fact that the standard gave at least some reviewers an impression of substantial complexity. The following comment (from an otherwise favorable Department of Defense reviewer) illustrates one such reaction:

"The standard and supporting documents are judged to be a highly sophisticated technical approach to a difficult problem, which is to treat all manner of digital communication systems under the same standard umbrella. The standard represents the latest in technical thought on the subject and as such is not intended for the uninformed reader. Much effort must be devoted to understanding it."

Although many reviewers did not share this opinion, the fact that a significant minority did reinforced the view that a gradual implementation of the standard would be advisable.

The question of complexity deserves a brief separate discussion here. There is nothing inherently complex about the performance parameters specified in FED STD 1033; they express basic performance concerns which are readily understandable,

¹Data Terminal Equipment/Data Circuit-Terminating Equipment.

and vitally important, to data communications users. The impression of complexity the standard conveys to some is a result of the rather rigorous way the parameters are defined. Such rigor was necessary, in the absence of a completed measurement standard, to ensure the comparability of measured performance parameter values. Two effective steps are now being taken to eliminate this problem:

1. NTIA/ITS is developing a performance measurement standard which will incorporate the mathematical details of the performance parameter definitions into a standard, machine-independent computer program. This standard program will transform observed performance data into parameter values in a totally uniform way, to ensure comparability; and will eliminate the need for mathematical formality in FED STD 1033. The new standard, Federal Standard 1043, will be titled "Digital Communication Performance Measurement Methods."
2. On the strength of the upcoming measurement standard, the joint ANSI/FTSC Task Group is recasting the FED STD 1033 parameter definitions in a more informal, narrative style. The developing ANSI standard, designated X3S35/125, is titled "User-Oriented Data Communication Performance Parameters."

Interim Federal Standard 1033 was officially published by GSA's Federal Supply Service on August 29, 1979. Since that time, NTIA and NCS have received over 600 separate requests for the standard, from government and industry organizations in the U.S. and more than a dozen foreign countries; and a number of initial applications of the standard are under way. It is likely that when the joint ANSI/FTSC Task Group completes its adaptation of FED STD 1033, the U.S. government will adopt the ANSI version as a mandatory Federal Standard, to be used in all Federal data communication procurements. Feedback from initial applications of the interim Federal Standard will be extremely useful to the joint Task Group in shaping its ultimate successor.

1.2 Purpose and Scope of Report

The purpose of this report is to encourage and facilitate initial use of Interim Federal Standard 1033 by providing an informal, non-technical presentation of its objectives and content. Earlier reports on 1033 were directed to standards developers, and addressed technical issues associated with parameter definition and measurement. This report is directed to standards users, and presents the standard from a more practical, user-oriented perspective.

The report is divided into three major sections. Section 2 outlines the need for the standard and the potential benefits of its use. Section 3 summarizes the objectives and content of the standard. Section 4 provides a tutorial "essay"

on the meaning and importance of each standard parameter. An annotated bibliography of technical reports and papers dealing with performance assessment issues is provided in Section 5. Guidelines for applying the standard in actual communication procurements will be provided in a planned sequel to this report, the Interim Federal Standard 1033 Application Manual.

The explanatory, user-oriented nature of this report necessarily imposes certain limitations. The report takes the FED STD 1033 parameters as a given starting point, and provides little discussion of how or why they were selected. The parameter selection process is described thoroughly in Seitz and McManamon (1978). The report defines the standard parameters in an informal, narrative style, with mathematical details and "fine points" intentionally omitted. More rigorous parameter definitions are provided in the standard itself. Finally, the report does not address the complex subject of performance measurement, or related issues such as sampling strategy. These topics will be addressed in the measurement standard (FED STD 1043) and in later reports. A technical basis for this work has been established in a series of ITS reports by Crow (1974, 1978, 1979) and Crow and Miles (1977).

2. THE BENEFITS

2.1 Introduction

Why use Federal Standard 1033? The question is a serious one, particularly since adherence to the standard will require a significant change in procedures and thinking for many potential users. This section answers that question in terms of tangible benefits that will accrue to end users, suppliers, and communication managers who use FED STD 1033 as a functional framework for future communication procurements. The section is divided into three major subsections. The first outlines existing Federal data communication procurement practices; the second identifies current sources of inefficiency in these practices; and the third describes the benefits of the functional procurement methods proposed herein. Current inefficiencies and potential benefits are discussed from three points of view: end user, supplier, and communication manager. Although the discussion focuses on Federal procurement, similar problems and opportunities exist in many non-Federal organizations as well.

2.2 Existing Procurement Practices

A useful description of existing Federal data communication procurement practices requires that we distinguish three general categories of participants: end users, communication suppliers, and communication managers. An end user is an individual or entity that produces or "consumes" information transmitted over a telecommunication system. Typical end users of data communication service are a human terminal operator and a remote computer application program in a tele-processing network. We employ the term end user (or "user") rather broadly here, to denote either groups of users or individuals.

The term communication supplier here denotes any nongovernment organization that provides telecommunication services or equipment to the Federal government. Traditionally, the major suppliers of telecommunication services have been the common carriers. The equipment suppliers include suppliers of interconnect equipment (e.g., modems, data terminals, PBX's); and suppliers of primary equipment (e.g., microwave and satellite terminals, antennas). Inasmuch as all communication equipment exists to provide communication services, it can be said that all communication suppliers ultimately contribute to the provision of services.

The third category of participant in Federal communication procurement is the communication manager. The communication manager essentially serves as a broker, or middleman, between a user who requires communication services and a supplier or group of suppliers who provide them. There are at least three levels of communication managers in the Federal government: (1) organizational managers, responsible for a particular installation or activity (e.g., the Department of Commerce Laboratories in Boulder); (2) departmental managers, responsible for an overall Federal agency (e.g., the Department of Commerce); and (3) administrative managers, responsible for coordinating multi-agency procurements. The latter function is performed for most nonmilitary agencies by GSA's Automated Data and Telecommunications Service (ADTS), and for the Department of Defense by the Defense Communications Agency (DCA). At the organizational level, the communication management function typically is performed by a single individual; at the administrative level, the function is performed by a large Federal organization with a multi-million dollar budget.²

These characterizations provide a basis for describing the key steps in a typical Federal data communication procurement. For the sake of concreteness, we

²As an example, GSA/ADTS employs 2,500 people and buys \$500 million worth of data communications equipment annually.

presume a situation in which data communication service is needed to interconnect the users of a number of geographically remote time-sharing computers owned by a single, nonmilitary Federal organization. The procurement would be conducted, according to GSA's Federal Property Management Regulations (FPMR's), as follows (GSA, 1978):³

1. The need for a data communication service is perceived by a group of users (e.g., programmers, administrators, ADP managers) at the various computing sites. As an example, frequent conferences and close working relationships between individuals at different work sites might require the exchange of large amounts of digital information, in real-time, between sites. Or, two or more remote computers might serve as backups to each other, to maintain data processing service continuity during periods of equipment outage or unusually heavy usage.
2. Representatives of the various user communities organize a "data communications study group" which conducts a "data communications study." Quoting from Subchapter F, Part 101-36 of the FPMR's,

"The data communications study includes a detailed analysis of the proposed data processing system and the environment within which it will operate and a determination as to the feasibility and economy of data communications under the circumstances. Also, such a study indicates the additional equipment and the type and number of communications lines which are estimated to be required; the impact on the format of data and data banks, codes to be used and programming required; and, most significantly, the important elements of cost."
3. The study group produces, as its output, "a written report detailing the data communications system which most economically and effectively satisfies the requirements of the proposed data processing system."
4. The report is transmitted to a higher authority (with responsibility for both the individual computers and the proposed communications extension) for analysis and validation. The FPMR's state that "such analysis should consider the extent to which the proposed data communications system will satisfy the requirements of the proposed data processing system and should include a cost/benefit study to determine whether to include or exclude data communications from the proposed data processing system."
5. Upon completion of this review, the organizational communication manager prepares a series of telecommunication service request forms (GSA 2936-B) detailing the required circuits and facilities; and transmits these forms, through the appropriate departmental communication manager, to GSA.

³The discussion disregards the agency process of funding acquisition.

6. GSA/ADTS communication management personnel review the service request forms and approve or disapprove them within 20 workdays. In this connection, the FPMR's state that "if no action is taken by GSA within the 20 workdays after receipt of a request from an agency ... the agency concerned may proceed as if, in fact, approval had been granted."
7. Upon receipt of GSA approval, the organizational communication manager submits an order for the necessary communication equipment and services to the appropriate government procurement officials; and the normal process of supplier bidding, bid evaluation, contract award, delivery, and installation proceeds from there.

The DoD procurement process differs from that described above in various details, but the basic plan is similar. Requirements for new services are perceived and initially reviewed within the using commands. Each military department has one or more validation offices, equivalent to the departmental communication managers, that are responsible for assessing new requirements. Validated requirements are submitted to the Defense Commercial Communications Office of DCA, which accomplishes the actual procurement of services from nongovernment suppliers (GAO, 1977).

2.3 Sources of Inefficiency

Any system specification process involves two fundamental steps:

1. Specifying what the system must do, in terms of a set of required functions and associated performance levels;
2. Specifying how the system will achieve these objectives, in terms of specific components, interconnections, and operations.

The first step is alternatively called a "user requirements analysis" or "system requirements analysis." Properly conducted, it is a careful examination of the user function the system must support; it determines the quantitative impact of system performance on user effectiveness, and thereby defines the objectives of system design. As an example, communication delay would have vastly different impacts on the user functions of inventory accounting and missile fire control; and the user requirements in the two cases would differ correspondingly. The output of the requirements analysis step is called a functional specification to emphasize the fact that it defines "what", but not "how".⁴

The second step in the system specification process is the detailed system design. In this step, the analyst postulates various ways a system could be constructed to meet the user requirements; and evaluates each relative to the

⁴Such a specification also identifies relevant cost and operational constraints.

defined constraints. As an example, a given end-to-end delay requirement might be satisfied by a dedicated communication link with a relatively low information transfer rate; a polled message switching network with a higher rate; or, conceivably, by a totally nontelecommunications solution such as express mail. The output of the system design step is called a design specification (or fabrication specification) to emphasize its focus on "how." In sum, the functional specification defines the service; the design specification defines the system that provides that service.

A major deficiency in current Federal data communication procurement practices is that they do not clearly distinguish between these two basic specification steps. The "data communications study" described in the FPMR's includes both the user requirements analysis and the system design; its output, the study report, documents the results of both. No intermediate functional specification is required or, in fact, even suggested.

If anything, the current FPMR's encourage agencies to mix the specification of user requirements with system design. As an example, one of eight "factors to be examined" in a data communication study is described as follows: "Accuracy required and the necessity for employment of error detection and correction techniques." Accuracy is a user performance requirement; error detection and correction techniques are particular methods of achieving a stated accuracy requirement. The FPMR's list the following factors to be identified in the data communications study report:

1. Type of service.
2. Line requirements.
3. Hardware and software requirements.
4. Error detection/correction techniques required.
5. Time- or event-dependent statement of additions to the communications capability required for the handling of expected increases in workload or in demands on the system.
6. Space requirements.
7. Detailed statement of estimated one-time and recurring costs.
8. Cryptographic security requirements.
9. Noncryptographic security requirements.

10. Rationale why data communication via electronic means is required as opposed to mail or courier service.
11. Access security.
12. Other facilities in the area which can accommodate the requirement within their existing or enhanced capabilities.
13. Review of existing communications capability.

Factors 1-4, 7, 12, and 13 are outputs of a system design, i.e., a study of "how". The other seven factors are outputs of a user requirements analysis, i.e., a study of "what".

Combining user requirements analysis and system design in a single "data communications study" often has the effect of forcing the user to assume the primary responsibility for system design. Although the study group can, in principle, include communication managers and other consultants as well as users, it is the user who needs the service; and necessity motivates participation. At the same time, current GSA regulations do not allow the study group to stop at the functional specification stage. As noted above, the FPMR's stipulate that the study group must produce a detailed design specification, and obtain GSA/ADTS approval of it, prior to procurement. Figure 2-1 identifies the categories of information such a group would submit to GSA at Step 5 of the procurement process described above; in essence, what is required is a list of specific services and equipment needed to implement a completed system design.

Forcing the user into a design role has a number of major disadvantages. These can be described from the point of view of the end user, the supplier, or the communication manager. The typical end user views data communication as transport service to be used, like the mail, in moving information from one place to another. He has little interest in how information transport is physically accomplished; his concerns are with its ultimate speed, accuracy, and reliability, and with the ultimate cost of the service.

Most users would be very willing to stop at the functional specification stage if they could be sure their requirements would be met in a reasonable manner. Communication system design to them is a complex, time-consuming, treacherous process which diverts valuable time and resources from their real mission. They feel, justifiably, that they should not have to understand the design of a system in order to use it.

Suppliers approach the data communications market place as sellers rather than as buyers, but their dissatisfaction with user design is no less strong. To a

PAGE OF _____
 IRCN 0074-GSA-OT

Data, Facsimile, and Record Telecommunications Services Serving Computers

SECTION I - GENERAL INFORMATION

1. Transaction: Add Change
 2. Submission Type: FPMR Approval Request Notification
 3. Organization Code: _____
 4. System ID: _____
 5. Operation Date: Y R M O
 6. Disconnect Date: Y R M O
 7. Submission Date: Y M D
 8. Will the requested facilities be used for the storage, retrieval, processing, or transfer of personal information? If so, please attach appropriate certification as prescribed in Federal privacy regulations. Yes No

SECTION II - COMPUTER LOCATION INFORMATION

9. Location ID: _____
 10. Operating Activity (Optional): _____
 11. Street Address: _____
 12. City: _____
 13. State: _____
 14. ZIP Code: _____

SECTION III - CIRCUIT AND MODEM INFORMATION

Reference Line No.	Quantity	CIRCUIT INFORMATION						MODEM INFORMATION					
		Common Carrier Code or Name	Restoration Priority	Type of Circuit	Full or Half Duplex (F or H)	Conditioning	Local or Intercity (L or I)	Manufacturer Code or Name	Model No.	Purchased or Leased (P or L)	Cost Per Modem in Dollars (Optional)	Transmission Speed in Bits Per Second (BPS)	
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)

15. Remarks (if necessary, continue on reverse)

16A. REQUESTOR'S SIGNATURE

16B. REQUESTOR'S NAME AND TITLE (Type or print)

16C. TELEPHONE NO.

17A. CONTACT'S NAME

17B. TELEPHONE NO.

16D. AGENCY AND OFFICE

Figure 2-1. Service specification form.

supplier, the existence of a preconceived user design in a procurement specification means two things: (1) his opportunity to choose the most efficient, economical method of meeting the user need has been usurped; and (2) he may have been precluded from seeking the user's business altogether if the products or services of a competitor have been specified in preference to his own. User-developed design specifications are frequently incomplete, ambiguous, and misleading; and such specifications inevitably increase supplier uncertainty, inefficiency, and financial risk. Poorly prepared specifications can also directly reduce supplier revenue, both by delaying the installation of new services and by hastening their abandonment. In short, user design efforts are often most unwelcome to the data communication supplier.

As noted earlier in this section, the Federal communication manager essentially operates as a broker between a user who requires communication services and one or more suppliers who provide them. Federal communication managers are charged with two general objectives: (1) to meet the communication needs of their user clients at the lowest possible cost, and (2) to implement Federal policy guidelines in such areas as reliance on the private sector (OMB, 1979); efficient spectrum utilization (NTIA, 1980); and Federal Telecommunications System interoperability (National Security Council, 1979).

Both communication management responsibilities are frustrated by the current concentration of system design responsibility in Federal user organizations. Since the typical user has relatively little communication expertise, his "designs" tend to be inefficient, costly, brute-force approaches - e.g., dedicated lines. Faced with the complex task of designing a communication system, users tend to skip the requirements phase altogether, with the result that the procured system may have little relationship to actual needs. Even in cases where users are capable of efficient design, they are not in a position to assess the feasibility of "common user" solutions, since they are not normally aware of the requirements of other Federal users. Once a user organization has committed itself to a particular system design, that design becomes the basis of discussion with the communication manager, with the result that little consideration is given to other design alternatives or applicable Federal policy guidelines. In extreme cases, the communication manager is reduced to a "paper pusher" and "front man" with no real authority or responsibility for anything.

Users placed in a design role often seek direct assistance from communication suppliers. While most suppliers are undoubtedly conscientious, they have little

incentive to suggest alternatives provided by their competitors; and they may be unaware of pertinent Federal policy guidelines. Suppliers placed in a communication management role also have a strong tendency to overdesign: first, because they are uncertain about the real user need; and second, because they receive more revenue from more elaborate services. Like the users, suppliers often are not in a position to aggregate requirements among different Federal organizations. Premature user consultations with suppliers often result in the user's requirements being very simply defined: "whatever the last salesman out the door said he needed."

The following excerpts from a recent General Accounting Office report substantiate these claims with respect to defense communications (GAO, 1977):

- "Users, rather than a centralized authority having cognizance of Defense-wide needs, decide upon the method of satisfying their individual requirements."
- Requests for new services often lack sufficient explanatory information to permit adequate consideration of alternative means of satisfying the requirements. Validation officers we talked with indicated that users' requests for dedicated services are seldom questioned if the user can provide the funds."
- "Defense Commercial Communications Office officials contend that they are not technically or quantitatively staffed to prepare designs or provide alternative solutions to service from those proposed by the carrier. Air Force Communications Service officials informed us that their review of the carrier's proposal was limited to determining its technical adequacy."
- "Based on our review of about 500 leased dedicated circuits costing over \$5.6 million annually, we found about 450 circuits, costing nearly \$4.9 million annually, which are candidates for either elimination, reconfiguration for more economical service, or integration into switched common-user networks."

This situation raises some fundamental questions. Data communication procurements are going on at a rate approaching \$2 billion annually in the Federal government. If these procurements are, in fact, poorly conceived in a large number of cases, how much potential improvement in Federal user productivity is being sacrificed? How much is industry innovation being inhibited? How much Federal tax money is being wasted on inappropriate system configurations?

These questions are difficult to answer in quantitative terms, but the impacts clearly are substantial in all areas. In commenting on ITS projections of potential savings in the latter area, a prestigious National Research Council study panel made the following observation (NRC, 1977):

"The impact of a good data network selection method for the General Services Administration (GSA) to meet the users' needs at the least cost is understated at best. The impact would not be just the quoted 5% of 6% of the government data bill ... A larger goal to shoot for is avoiding the wasteful procurement of a system significantly more costly than necessary to fill the requirement. If we conservatively assume 20% could be saved by an efficient method to select the right service or system, the impact could be \$400 million per year"

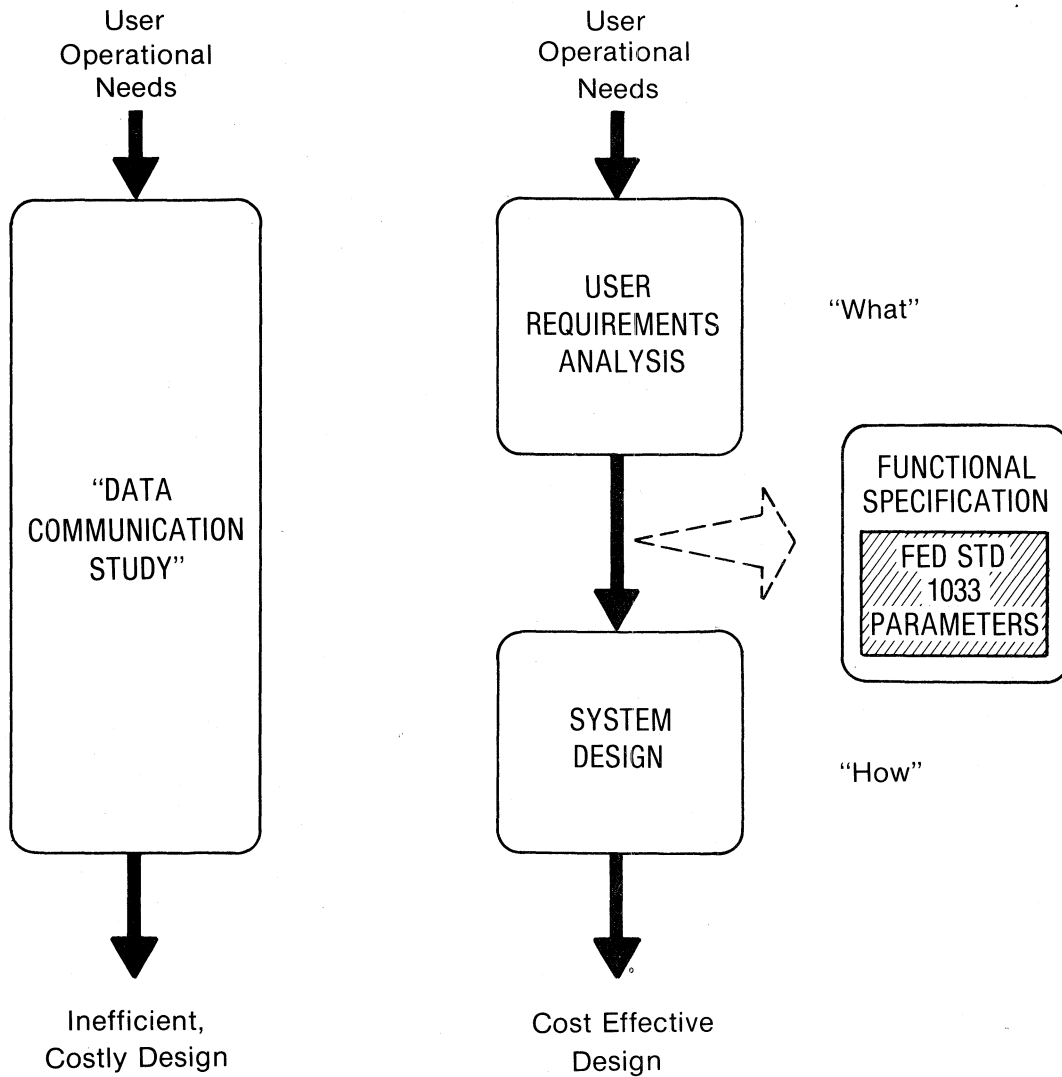
Federal data communication expenditures are currently growing at a rate in excess of 20% per year, and are expected to continue to do so at least through the mid-1980's.

2.4 FED STD 1033 Benefits

Why is the distinction between user requirements analysis and system design so unclear in current Federal procurement regulations? Why has it been so difficult to realize the benefits of a functional approach to data communication procurement? A key problem has been the absence of a suitable framework for functional performance specification - in essence, a "common language" for relating the performance needs of end users with the capabilities of supplier systems and services.

Figure 2-2 compares the existing and functional procurement approaches and illustrates the crucial role of the functional specification in achieving the latter. The existing approach, Figure 2-2a, mixes user requirements analysis and system design in a single, undifferentiated "data communication study." This approach encourages a mixing of "what" with "how", and results in poor delineation of the user requirements and an inefficient, costly system design. The functional approach, Figure 2-2b, clearly distinguishes the user requirements analysis and the system design as separate studies; and produces a clear, complete description of the user requirements and a more cost effective system design. The key element that makes this separation possible is the functional specification: a single statement of both the user requirements and the minimum system objectives in terms of system-independent, functional performance descriptors. Federal Standard 1033 provides such a set of functional performance descriptors.

Just how will FED STD 1033 benefit Federal data communication procurement? Again, the question can be addressed from the point of view of the end user, the



a. Existing Approach

b. Functional Approach

Figure 2-2. Existing vs functional procurement.

supplier, or the communication manager.⁵ The standard will benefit Federal end users in two obvious ways:

1. By relieving them of a burdensome responsibility for communication system design. Users will be enabled and encouraged to regard data communications as a transport service - their natural inclination in the first place.
2. By allowing them to define their data communication needs more precisely. The 1033 parameters will enable users to pinpoint the specific impacts of communication performance on their own operations, thereby minimizing procurement uncertainty and the risk of costly mistakes.

An equally beneficial, but less obvious, user application of FED STD 1033 will be in assessing potential uses of data communications in improving government productivity. Federal employees frequently encounter situations where it seems clear that a necessary government function is "information limited" - i.e., the function could be performed much more efficiently, with a substantial cost savings to Federal taxpayers, if accurate, timely input information were available. Nevertheless, it can be very difficult to quantify the potential savings in monetary terms; and Federal budget planners quite properly expect such evidence if a major commitment of resources is required. Federal Standard 1033 provides an ideal framework for assessing such situations because it focuses analysis on the user need without presupposing a particular design solution - telecommunications or otherwise. The effect of information delay, inaccuracy, or unavailability on an information-dependent function can be determined without specifying a transport method for that information; and the FED STD 1033 parameters enable the analyst to do just that.

Federal Standard 1033 will benefit data communication suppliers in three tangible ways:

1. By enlarging their participation in the design of Federal data communication systems. Suppliers will receive both the direct stimulus of new "turn-key" business, and the indirect stimulus of enhanced design responsibility.
2. By simplifying the process of describing their system or service offerings. Suppliers will be able to develop a single basic performance specification applicable to all potential Federal uses.

⁵The discussion assumes supplier design responsibility. Similar benefits accrue when Federal communication managers are responsible for design as long as a system-independent functional specification is developed.

3. By maximizing their opportunity to compete for Federal business. Expanded use of functional specifications in Federal procurement will prevent arbitrary exclusion of qualified suppliers.⁶

Use of the standard will also benefit data communication suppliers in an indirect way, by rewarding successful innovation and the marketing of cost effective products and services.

Federal Standard 1033 will assist Federal communication managers in discharging both their user service and their policy implementation responsibilities. To meet a user's data communication needs, a communication manager clearly must know what those needs are - and few managers enjoy that knowledge under the present conditions. Implementing Federal policy guidelines requires a certain authority over Federal procurement decisions - and again, few Federal communication managers actually have that authority today. Successful implementation of the standard will provide specific benefits in at least three phases of communication management:

1. Requirements Specification. The standard will improve the quality of user requirements specifications, and will facilitate their development through the cataloging of similar prior applications.
2. Service Acquisition. The standard will simplify the matching of end user requirements with offered systems and services. The NRC estimate cited earlier provides an indication of the total potential improvement here.
3. Service Assessment. The standard will provide a basis for agreement between users and suppliers on the quality and reliability of delivered services, and will facilitate direct comparison of service alternatives.

All of these benefits are a result of the standard's "common denominator" property.

An "ideal" situation with respect to Federal procurement of data communication systems and services might be briefly summarized as follows:

1. Federal User Organizations would understand and accept the need to describe their communication requirements in a system-independent manner. Users would specify performance requirements in functional terms, without reference to particular communication facilities or services. Individual parameter values would be determined on the basis of their impact on the user process being served; as an example, the Bit Error Probability requirement for a digital air traffic control system would be determined by considering the impact of bit errors on air traffic control effectiveness.

⁶This is true of equipment suppliers as well as service suppliers, since subsystems can also be specified in functional terms.

2. Industry Suppliers of communication systems and services would be willing to specify their performance in uniform functional terms, and would be appropriately compensated for their effort in doing so. Available facilities and services would be catalogued, along with major Federal applications, in a central communications catalog which would facilitate the design and procurement process.
3. Federal Communication Managers would assist end users in defining realistic communication requirements; and would have the authority and the resources to select the best available means of meeting these needs. Where appropriate, they would aggregate independent user requirements to be met by a single common user system. They would, in sum, effectively manage Federal communication procurement. A major improvement in the cost effectiveness of Federal data communication systems would result.

Federal Standard 1033 will contribute substantially to the realization of this ideal by providing a system-independent, functional framework for future communication procurements.

3. FED STD 1033 EXPLAINED

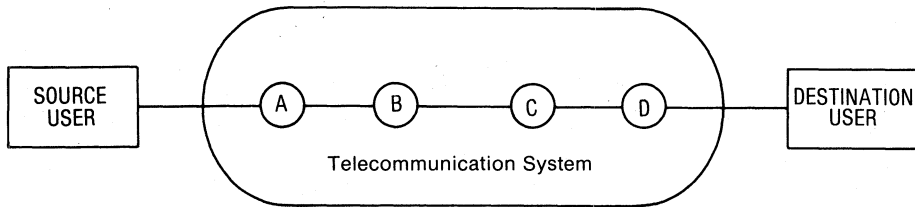
3.1 Introduction

This section summarizes the objectives and content of interim Federal Standard 1033. The section is divided into two major subsections. The first describes three key technical problems which influenced development of the standard. The second summarizes the overall FED STD 1033 approach. Refer to the standard and its supporting reports for further detail on each topic.

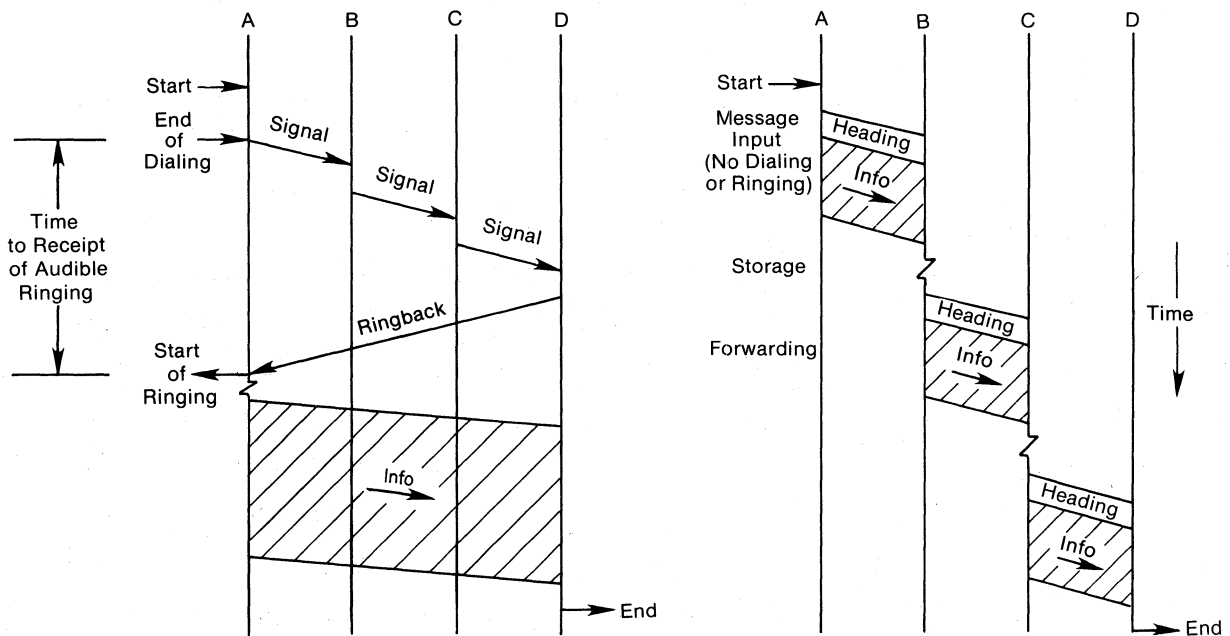
3.2 Key Technical Problems

Development of FED STD 1033 required the solution of three key technical problems. The first of these was the problem of system dependence. A survey of candidate parameters revealed that the great majority were defined such that they could only be applied to systems with particular topology or protocol features. This is undesirable because it prevents use of the parameters in comparing systems that provide the same ultimate service by means of different detailed designs.

A good insight into the problem of system dependence can be obtained by considering the differences between traditional circuit-switched and message-switched transactions, as shown in Figure 3-1. (Note that time proceeds down the page in this and similar diagrams). In a circuit-switched transaction, service is provided by setting up an end-to-end path, or circuit, from source to destination



a. Network Topology.



b. Circuit-Switched Transaction.

c. Message-Switched Transaction.

Figure 3-1. System dependence.

prior to the start of user information transfer. The individual links which comprise the end-to-end circuit are all allocated to that particular user pair for the duration of the transaction, independent of usage; and all links are used concurrently during transfer. A familiar example is a normal voice telephone call.

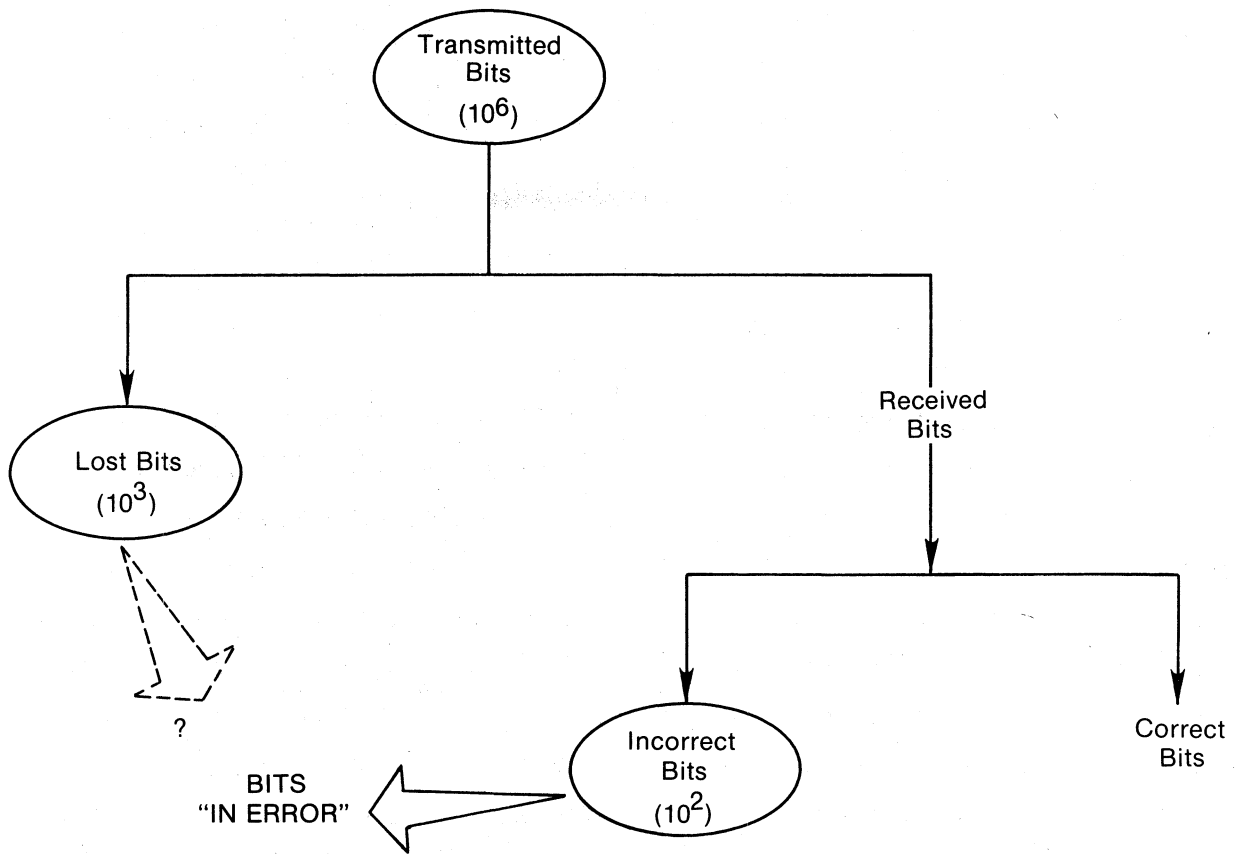
In a message-switched transaction, no end-to-end circuit is set up prior to the start of user information transfer. Instead, the user message is forwarded through the network link by link; and the entire message is stored for some period of time at each intermediate node. Individual links are allocated to a particular user pair only during actual forwarding of their message. At all other times, the link may support other users. DoD's AUTODIN I is a classic example of a message-switching system. Another example is GSA's Automatic Record System.⁷

One parameter which is commonly used in expressing the performance of circuit-switched systems is the Time to Receipt of Audible Ringing - the elapsed time from the end of dialing to the start of ringing (Fig. 3-1). No counterpart to this parameter is possible in the case of message-switched systems. The function of switching is performed by interpretation of the message heading at each node, rather than by pre-transmission signaling; and the concepts of dialing and ringing are thus irrelevant. A user wishing to compare the performance of circuit-switched and message-switched services cannot do so in terms of Time to Receipt of Audible Ringing; other, system-independent descriptors of performance are required. The absence of such system-independent performance descriptors has been a major difficulty with performance comparison heretofore.

The second key problem encountered in developing FED STD 1033 was the problem of detailed parameter definition. In most cases, traditional narrative definitions of performance parameters are not precise enough to ensure their uniform application to comparable service offerings. The result, of course, is a potential for inefficiency and error in the process of matching service offerings with end user needs.

As an example of the parameter definition problem, consider the familiar accuracy parameter Bit Error Probability (Figure 3-2). A typical narrative defines this parameter as "bits in error per bits transmitted", but makes no mention of whether (or how) bits lost in transmission should be counted. Two obvious choices, both consistent with the narrative definition, would be (1) to count lost bits

⁷ Packet-switching is similar to message-switching, except that the messages are divided into smaller units, called packets, which are forwarded through the network separately.



$$\left\{ \begin{array}{l} \text{Bit Error} \\ \text{Probability} \end{array} \right\} = \left\{ \begin{array}{l} \text{Bits "In Error"} \\ \text{per Bits Transmitted} \end{array} \right\}$$

Choice 1 — Lost Bits Counted:

$$\left\{ \begin{array}{l} \text{Bit Error} \\ \text{Probability} \end{array} \right\} = \frac{10^2 + 10^3}{10^6} = 1.1 \times 10^{-3}$$

(Disparity
in Values)

Choice 2 — Lost Bits Not Counted:

$$\left\{ \begin{array}{l} \text{Bit Error} \\ \text{Probability} \end{array} \right\} = \frac{10^2}{10^6} = 10^{-4}$$

Figure 3-2. Detailed parameter definition.

with incorrect bits in calculating Bit Error Probability; and (2) to consider only received incorrect bits in calculating Bit Error Probability.

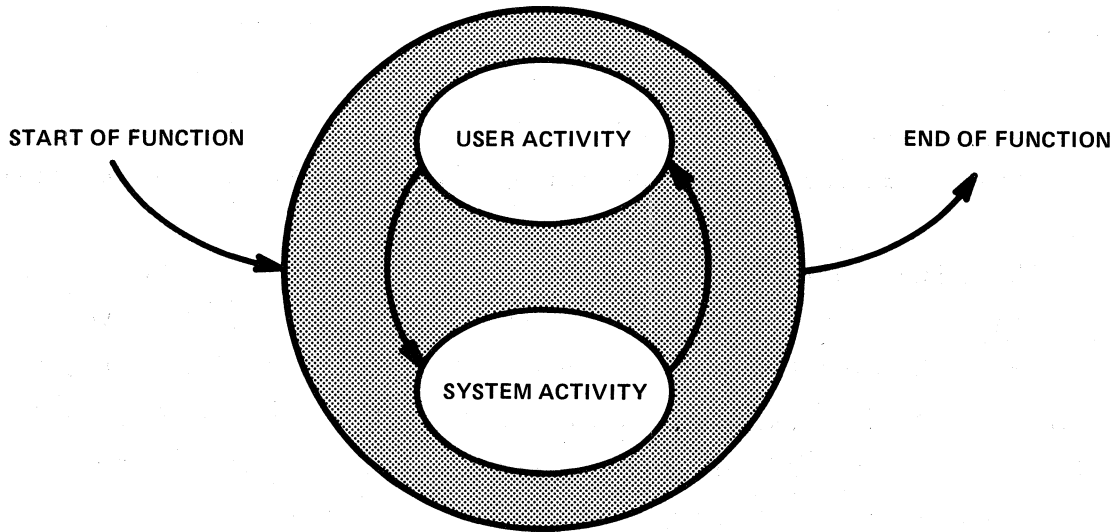
This ambiguity can have a substantial effect on measured parameter values. Assume that of a million (10^6) bits transmitted during a test, a thousand (10^3) are lost and a hundred (10^2) are inverted in transmission. The measured Bit Error Probability values under the two choices are 1.1×10^{-3} and 10^{-4} -- an order of magnitude error in interpreting the meaning of a narrative parameter definition.⁸ Other examples could be cited, but the problem is clear: defining words in terms of other words is an endless process which inevitably leaves room for misinterpretation.

The third key problem encountered in developing FED STD 1033 was the problem of user dependence. In most cases, the communication process involves a sequence of interactions between the users and the system; and overall communication performance depends, then, on user performance as well as system performance. There is an obvious problem in employing user dependent parameters in specifying required system performance: the carrier or other supplier normally has no control over user performance, and hence cannot ensure that user dependent parameter values will be met. Nevertheless, many of the parameters which best describe communication performance from the end user point of view are user dependent (e.g., "throughput").

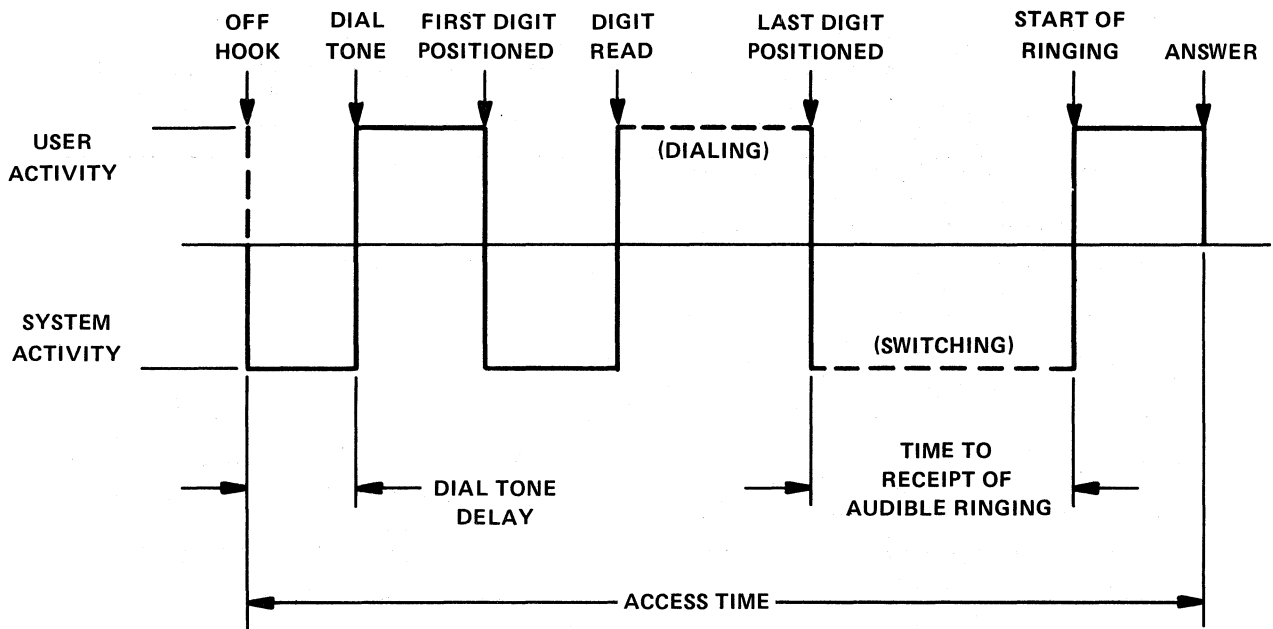
As a simple illustration of the user dependence problem, consider the position of a user who wishes to place a voice call over the public switched network (Figure 3-3). As he initiates the call, his major concern is with how soon conversation can begin, i.e., the total delay between his off-hook action and the called party's answer. The performance parameter Access Time describes exactly this delay; but its values depend not only on the system's speed in signaling and switching, but on the user's speed in dialing and answering.

The telephone companies have traditionally avoided this problem by focusing on parameters which describe unilateral system performance, e.g., Dial Tone Delay and Time to Receipt of Audible Ringing. Unfortunately, such parameters have two major disadvantages from the user point of view: (1) they are system dependent, as noted above; and (2) they do not reflect differences in the "functional burden" placed on the user by otherwise equivalent services. As an example, neither parameter would reflect, in terms of better performance values, the significant

⁸Data loss does occur in many systems, as a result (e.g.) of signal fading and retransmission protocol failures. See Section 4.3.4 for further discussion.



a. Alternating Activities within a User Dependent Function



b. Voice Telephone Access Illustration

Figure 3-3. User dependence.

advantage of abbreviated dialing over conventional rotary dialing. A few recent telephone company studies have recognized this limitation, and stress the need to describe the influence of user delay on overall end-to-end performance (e.g., see Duffy and Mercer, 1978). Nevertheless, a precise quantitative framework for expressing this influence has not been proposed heretofore.

3.3 FED STD 1033 Approach

Figure 3-4 summarizes the overall approach used in developing performance parameters for interim Federal Standard (Seitz and Bodson, 1980). The parameter development process consisted of four major steps:

1. Model Development. Existing and proposed data communication services were surveyed and certain universal performance characteristics shared by all were identified. These characteristics were consolidated in a simple, user-oriented model which provided a system-independent basis for the performance parameter definitions.
2. Function Definition. Five primary communication functions were selected and defined in terms of model reference events. These functions (access, bit transfer, block transfer, message transfer, and disengagement) provided a specific focus for the performance description effort.
3. Outcome Definition. Each primary function was analyzed to determine the possible outcomes an individual "trial performance" might encounter. Possible outcomes were grouped into three general outcome categories: successful performance, incorrect performance, and nonperformance. These categories correspond to the three general performance concerns (or "criteria") most frequently expressed by end users: efficiency (or "speed"), accuracy, and reliability.
4. Parameter Selection. Each primary function was considered relative to each performance outcome in matrix fashion; and one or more specific parameters were selected to represent performance relative to each function/outcome pair. Parameters were selected on the basis of expressed user interest, and consisted of probabilities, waiting times, time rates, and rate efficiencies. The matrix approach ensured that no significant aspect of communication performance would be overlooked in the parameter selection process.

The following paragraphs describe the results of these steps in more detail.

3.3.1 Model Development

In order to describe communication performance as seen by the end user, it is necessary to develop a "user's-eye view" of the communication process itself. What is the nature of the interface between an end user and a telecommunication system, and how is information transferred across such interfaces? How can the process of telecommunication be described in a way that is meaningful and familiar

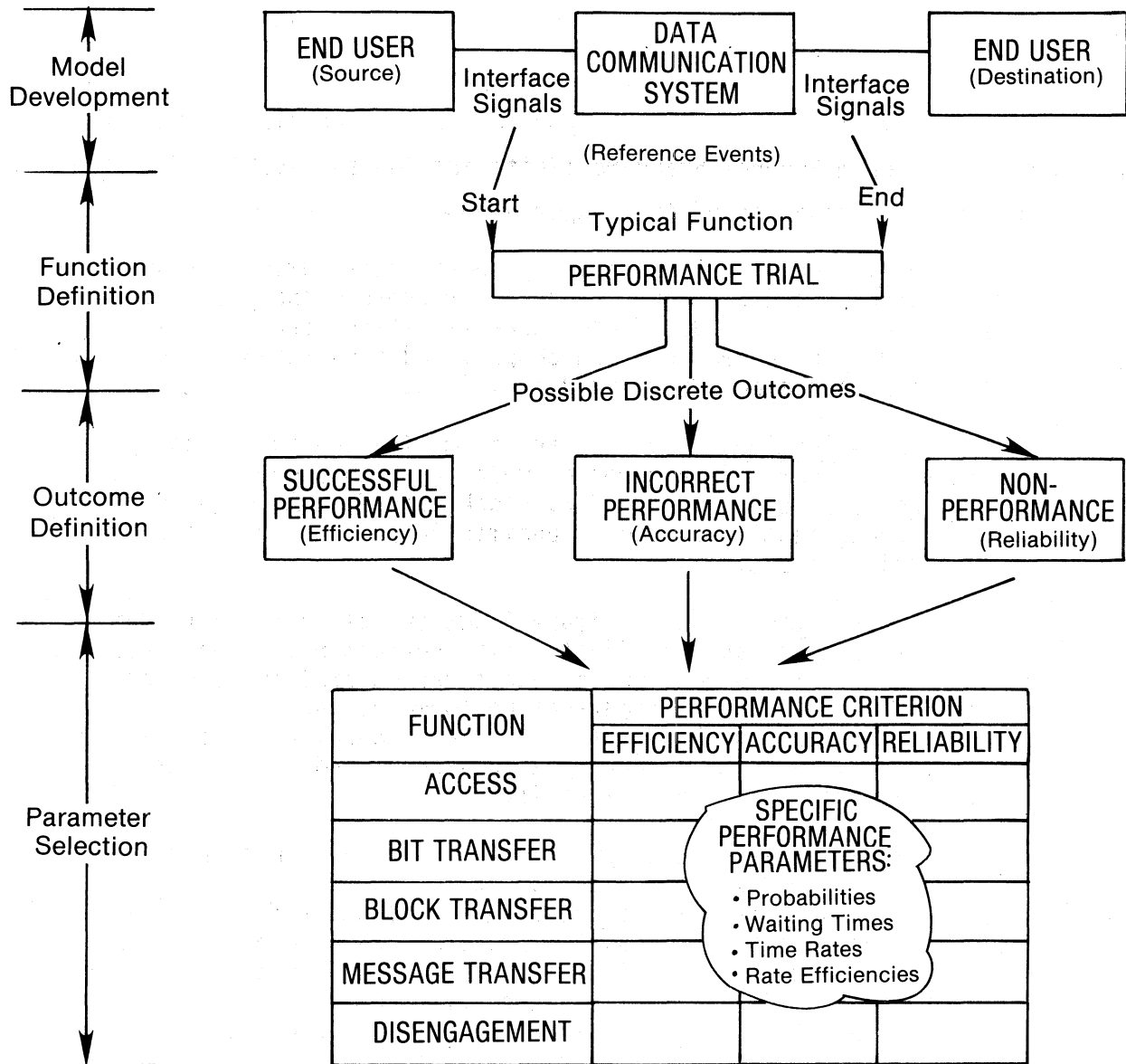


Figure 3-4. Parameter development overview.

to the end user, and yet not restricted to a particular type of interface or a particular interaction sequence? How should the performance parameter definitions be related to such a description? These are questions which FED STD 1033 answers with the aid of a user-oriented telecommunication process model.

The model defines the end user of a telecommunication system or service as one or more of the following types of entities (Fig. 3-5):

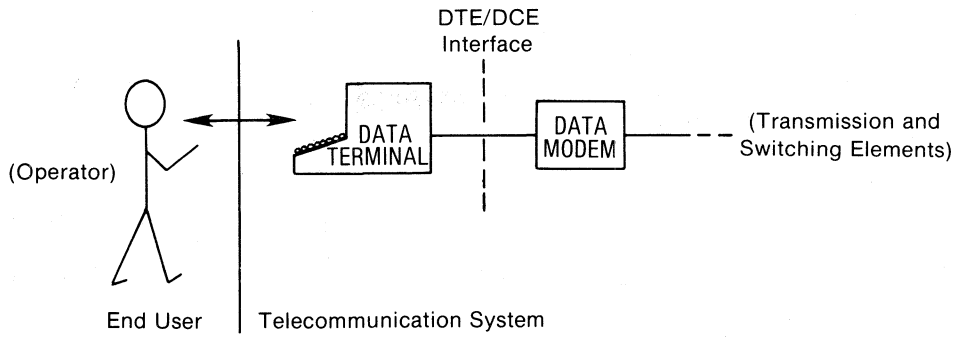
1. A human terminal operator.
2. An unattended device medium (such as punched cards).
3. A computer application program performing data processing functions unrelated to telecommunication.

In some cases, more than one type of entity supports the overall user function: for example, a terminal operator providing control inputs and a punched paper tape providing information storage at the same data communication station.

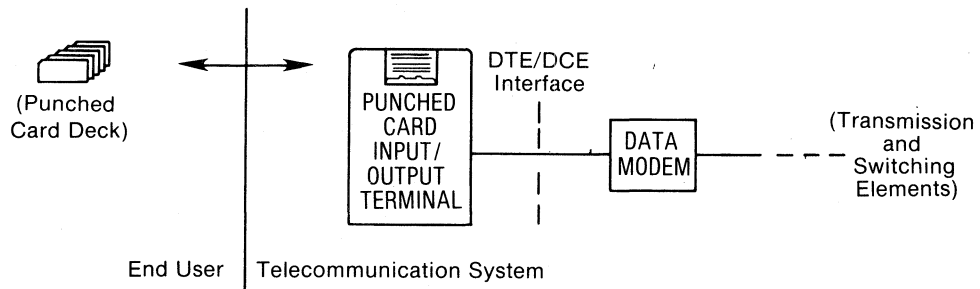
In most traditional data communication configurations, the end user is an operator or medium, and there is a distinct physical unit, the data terminal, which converts information from user-readable form into "transmittable" form and vice versa. In all such cases, the telecommunication system is defined to include the data terminal and all elements of the information transfer channel on its line side. The user/system interface then corresponds to the physical interface between the operator or medium and the terminal (Figure 3-5a,b).

The development of computer communications and teleprocessing has given rise to configurations in which the end user is an application program within a digital computer. In most such cases, there is a separate program within the same computer (often called a "telecommunications access method") which functions as a first point of contact for application programs requiring telecommunication service. In all cases where such a program can be clearly identified, the telecommunication system is defined to include the access method and all functional and physical elements of the information transfer channel on its line side. The user/system interface then corresponds to the functional interface between the application program and the access method or its equivalent. One such interface, applicable to the so-called Open Systems Interconnection (OSI) model being developed by the International Standards Organization, is shown in Figure 3-5c (ISO, 1979).

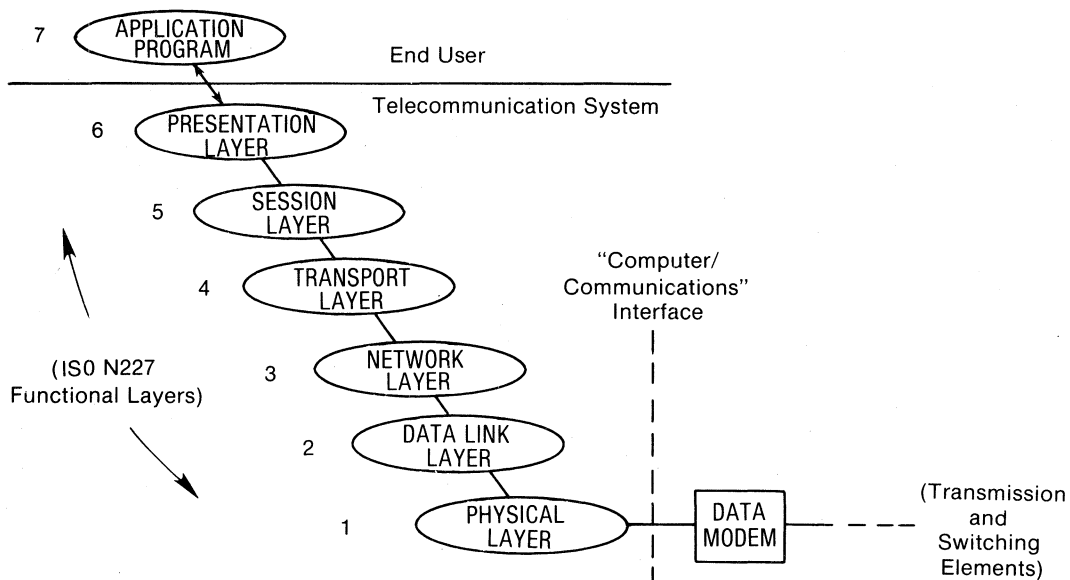
The above definitions place the end user interfaces well outside the traditional DTE/DCE or "computer/communications" boundaries. This viewpoint is essential in a user-oriented standard, since modern terminals and high-level protocols



a. Operator/Terminal Interface.



b. Medium/Terminal Interface.



c. Application Program/Access Method Interface.

Figure 3-5. User/system interfaces.

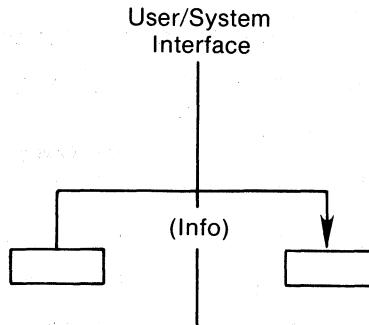
perform communication functions (such as error control, flow control, and virtual circuit establishment) which have a profound effect on end-to-end performance. One modern data communication network whose end user interfaces are defined in this way is IBM's Systems Network Architecture (McFadyen, 1976).

There are some computer communication and teleprocessing configurations in which it is difficult to draw a clear functional boundary between the "application program" and the "telecommunications access method." Guidelines for establishing appropriate user/system functional interfaces in such configurations are provided in Seitz and McManamon (1978).

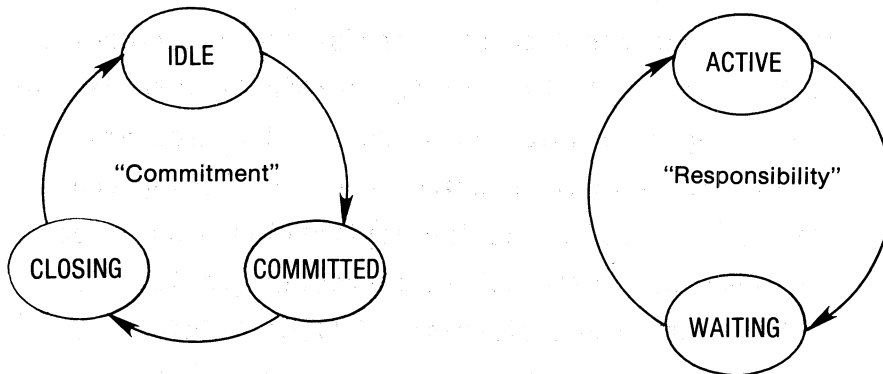
Information can be transferred across user/system interfaces in a variety of ways. Typical interactions at the operator/terminals interface are manual keystrokes and the printing or displaying of received characters. Typical interactions at the medium/terminal interface are the reading and punching of punched cards. Typical interactions at the application program/access method interface are the issuance of system calls, co-routine calls, and inter-process communications, and the setting and clearing of flags. Information can be transferred between computer programs either by physical movement of computer words or by buffer reallocation (transfer of buffer "ownership" without physical movement of the stored information).

All of the user/system interactions just described are examples of what the standard calls "interface events" (Figure 3-6a). As defined in the standard, an interface event is a discrete transfer of user or overhead information from the physical possession and control of one entity (user or system) to that of the opposite entity (system or user). User information includes all information intended to cross both user/system interfaces. All other information (e.g., ANSI (1971), ENQ, ACK, and SYN characters, off-hook and on-hook signals) is overhead information.

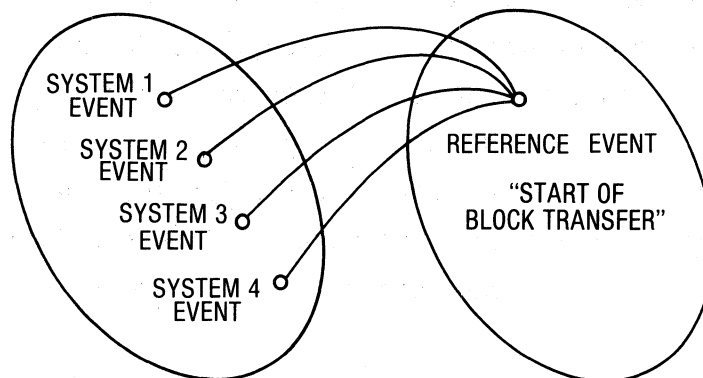
In any description of performance, certain key interface events are identified as events to be counted, timed, or compared in calculating performance parameter values. As noted earlier, most existing standards and specifications identify such key events by reference to particular system-specific signals (e.g., off-hook). The FED STD 1033 model departs from this approach by defining the performance parameters in terms of more general, system-independent reference events. Each FED STD 1033 reference event is a "generic event" which subsumes many system-specific interface events having a common performance significance; and each is defined in such a way that it can always be identified, if it occurs, in any



a. Nature of an Interface Event.



b. Overhead Transfer Effects.



<u>System</u>	<u>User Interface Device</u>	<u>System-Specific Event</u>
1	Character-asynchronous TTY	Typing a letter or figure
2	Buffered CRT terminal	Typing CR (end of line)
3	Host computer access method	Issuing WRITE system call
4	Punched card reader	Reading of card column

c. Example - Reference Event "Start of Block Transfer".

Figure 3-6. Reference event concept.

particular data communication transaction. The reference events collectively specify all information needed to describe performance in a comprehensive, user-oriented way.

The FED STD 1033 reference events are defined on the basis of the type of information transferred (overhead information or user information); the type of effect the transfer produces; and the particular user interface involved. Two basic types of overhead transfer effects are distinguished (Figure 3-6b):

1. Change in an entity's state of "commitment" to a particular information transfer transaction. Three possible commitment states are defined for each entity: Idle, Committed, and Closing.
2. Change in an entity's state of "responsibility" for creating the next transaction event at a particular user interface. Two possible responsibility states are defined for each entity: Active (responsible) and Waiting (not responsible).

All overhead information transfers of significance to user-oriented performance assessment can be represented as changes in these basic transaction states; and no further information (other than event times) is needed to define the associated performance parameters. Thus, these transaction state changes are the reference events associated with overhead information transfers.

An example will help to clarify the relationship between system-specific interface events and the associated reference events. A user's action in lifting a telephone handset off-hook transfers one bit of overhead information (the new hookswitch position) from the user to the system. This system-specific interface event transfers the calling user (and the system) from the Idle state to the Committed state; and makes the system Active, and the user Waiting, relative to the next transaction event (dial tone). These commitment and responsibility state changes are the reference event that corresponds to going off-hook. If we know the nature and time of occurrence of this reference event, we need no further information about the system-specific event which generated it in order to define (for example) the start of the access function. The same reference event might be generated by a completely different interface event in another system: An example is issuance of a Connect system call in the ARPA network.

User information transfer events normally do not affect "commitment" or "responsibility" as described above; they simply move information from the

⁹The model divides the system into two "half-system" entities to accommodate the possibility of independent effects at the two user interfaces.

physical possession and control of one entity to that of another. Here again, certain key information must be specified to support performance assessment (e.g., user interface, information content, and event time); and that information can be specified in terms of system-independent reference events. Parameters describing transfer performance can then be defined on the basis of these system-independent events.

An example of a system-independent user information transfer event defined in FED STD 1033 is the Start of Block Transfer. Such an event must obviously be identified to define performance parameters such as Block Transfer Time and Block Transfer Rate. In order to define the reference event Start of Block Transfer, we must clearly identify (1) what is meant by a "block", and (2) when the transfer of a block between end users should be regarded as "started". FED STD 1033 defines a user information block as "a contiguous group of user information bits delimited at a source user/system interface for transfer to a destination user as a unit." The transfer of a block is said to have started when two conditions have been met:

1. The user information contained in the block is physically present within the system facility.
2. The system has been authorized to transmit that information.

The latter criterion (authorization) is the most natural way to establish the block boundaries as well; i.e., authorizing transmission of a given unit of information identifies that unit as a FED STD 1033 block. Authorization may either be an explicit user action (e.g., typing Carriage Return at a buffered CRT terminal) or an implicit part of inputting the user information itself (e.g., typing a single character at an asynchronous terminal).

Given the above definitions, the nature of the information unit called a "block" and the physical events associated with block transfer will differ widely from one system to another (Figure 3-6c). Nevertheless, in every system configuration some specific information unit can be identified as a FED STD 1033 block, and the start of transfer of that unit can be determined, using the above criteria. Hence, the reference event Start of Block Transfer can always be identified. A similar approach is used in defining the other user information transfer reference events.

In describing the need for the FED STD 1033 model, we raised two questions which have not been explicitly answered as yet:

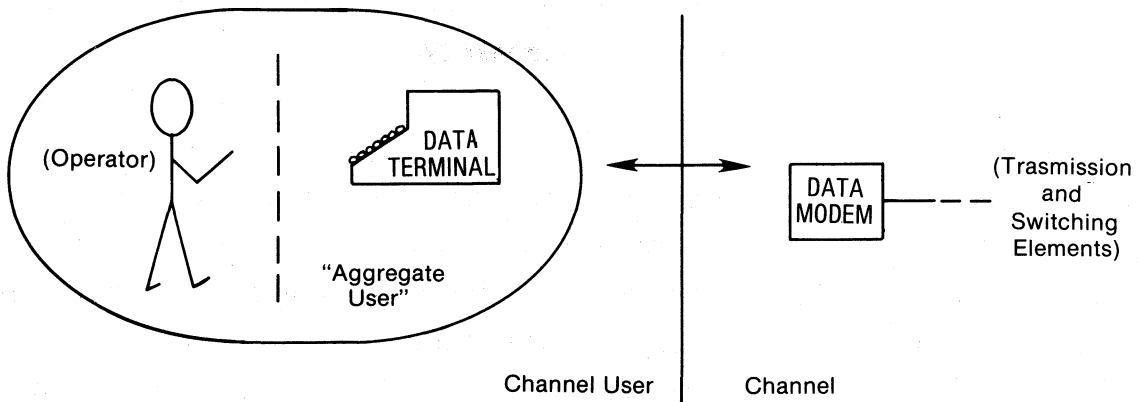
1. How can the process of telecommunication be described in a way that is meaningful and familiar to the end user, and yet not restricted to a particular type of interface or interaction sequence?
2. How should the performance parameter definitions be related to such a description?

We are now in a position to answer these questions. The process of telecommunication should be described as a chronological sequence of reference events, each specifying the time and impact or content of an associated overhead or user information transfer. The performance parameter definitions should then be based on the information specified in such a system-independent event history. The FED STD 1033 model implements exactly such an approach.

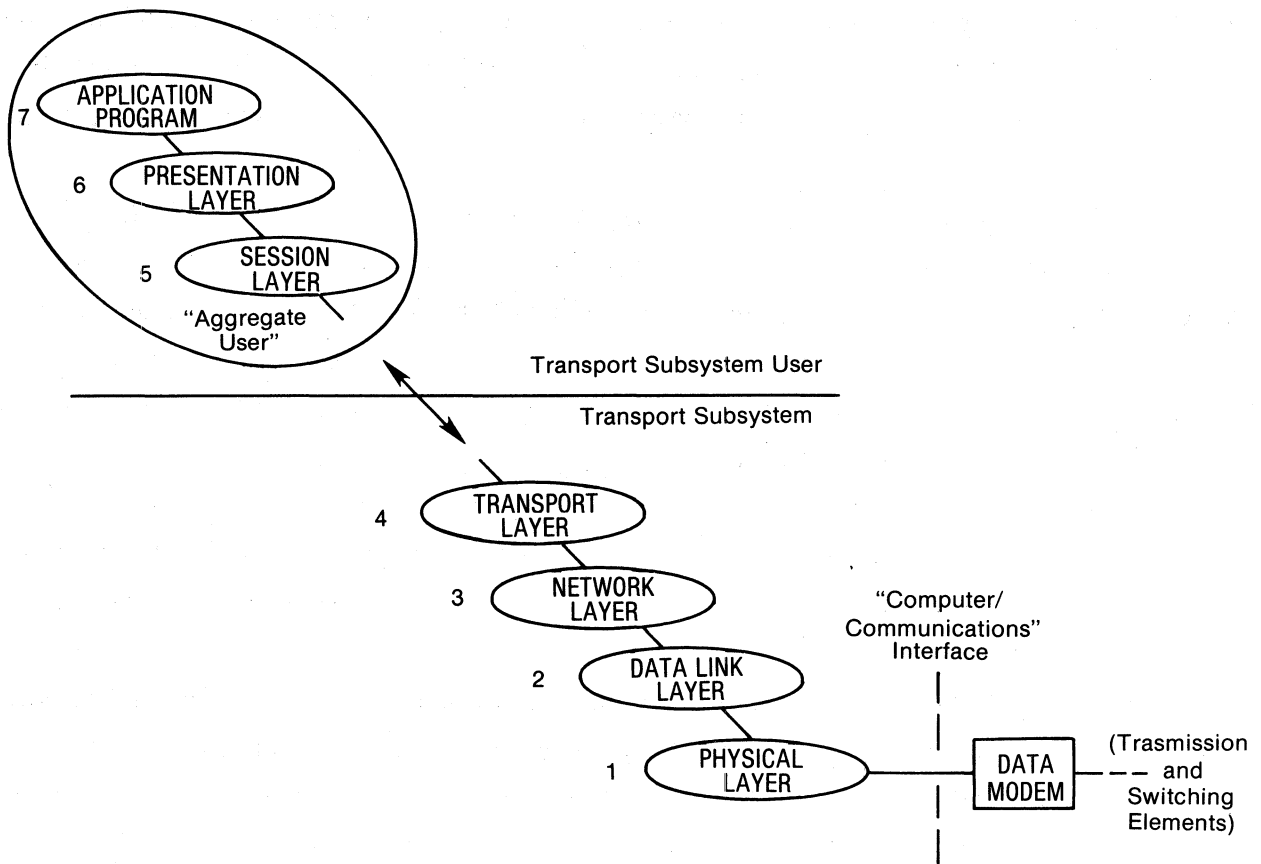
Two concluding remarks regarding the FED STD 1033 model are appropriate at this point. The first deals with the question of relative detail. Readers of the standard will note that the model is presented there in somewhat more detail than was used here. A more complete presentation was needed in the standard to support the development of performance measurement procedures. Nevertheless, with a few minor additions provided later, the model information presented here is sufficient to understand the FED STD 1033 parameters and to apply them in performance specification.

The second remark deals with alternative uses of the FED STD 1033 model. Although the model was developed primarily to represent end-to-end services, it is not restricted to such applications: any digital telecommunication process can be represented as a chronological sequence of reference events. In order to apply the model (and therefore the standard) to a digital subsystem, it is only necessary to (1) define the two interfaces of interest, (2) identify the specific events occurring at these interfaces, and (3) associate each specific interface event with a corresponding model reference event. The FED STD 1033 parameters can then be applied directly to the subsystem, since the parameter definitions are all based on the reference events.

Figure 3-7 illustrates two possible subsystem applications. In the first, Figure 3-7a, the subsystem interface is placed at the DTE/DCE physical interface; and the operator and terminal are regarded as an "aggregate user" of the information transfer channel. In the second, Figure 3-7b, the subsystem interface is placed between the Session and Transport layers of the OSI protocol hierarchy; and the Application, Presentation, and Session layers are regarded as an "aggregate user" of the Transport Subsystem (CCITT, 1980). Such applications can be useful



a. Operator and Terminal as an "Aggregate User" of the Information Transfer Channel.



b. Application, Presentation and Session Layers as an "Aggregate User" of the Transport Subsystem.

Figure 3-7. Aggregate user concept.

in allocating end-to-end performance objectives to purchasable components and services and, conversely, in determining the impact of subsystem choices on end-to-end performance.

3.3.2 Function Definition

Performance has little meaning as an isolated concept; to be useful, a description of performance must be clearly related to some particular function. The second step in developing the FED STD 1033 parameters was therefore to define a set of specific communication functions to be used as the focus of the performance description effort.

The five primary communication functions included in FED STD 1033 are defined in terms of particular model reference events as follows:

The access function begins on issuance of an Access Request signal at the originating user interface, and ends (normally) on the next subsequent input of a source user information bit or block to the system. It encompasses all activities traditionally associated with physical circuit establishment (e.g., dialing, switching, ringing, modem handshaking) as well as any activities performed at higher protocol levels (e.g., X.25 virtual circuit establishment). Making the end of access coincident with the start of input of user information to the system reflects the user view that no data communication service has actually been provided until user information begins to flow.

The bit, block, and message transfer functions describe the flow of user information between end users at three distinct levels of detail. Each function begins on the start of transfer of the associated information unit at the source user interface, and ends (normally) on completion of delivery of that unit to the intended destination. Each function encompasses all formatting, transmission, storage, error control, and media conversion activities performed between start of transfer and completion of delivery, including internal retransmissions if required. All three user information transfer functions must be considered in a comprehensive performance specification, as discussed below.

The disengagement function begins on issuance of a Disengagement Request signal at either user interface, and ends (normally) on return of a corresponding Disengagement Confirmation signal. A separate disengagement function is defined for each end user; the two functions may either be linked (as in the case of circuit-switched systems) or independent (as in the case of message-switched systems).

The terms Access Request, Disengagement Request, and Disengagement Confirmation are general descriptors of purpose (i.e., names of reference events) rather than particular interface signals. An Access Request is any interface signal issued for the purpose of initiating an information transfer transaction. The

corresponding commitment state change in the issuer is from Idle to Committed. Two specific Access Request signals have been cited earlier: The "off-hook" signal in the public switched network, and the Connect request in the ARPA network. Another typical Access Request signal is the Open Destination (OPNDST) VTAM Macro in IBM's Systems Network Architecture (SNA).

A Disengagement Request is any interface signal issued for the purpose of terminating an entity's participation in an information transfer transaction. The corresponding commitment state change in the issuer is from Committed to Closing. The Disengagement Request signals corresponding to the three Access Request signals just cited are the "on-hook" signal in the public switched network; the Close system call in the ARPA network; and the Close Destination (CLSDST) macro in SNA.

A Disengagement Confirmation is any interface signal issued for the purpose of confirming termination of an information transfer transaction. The corresponding commitment state change in the issuer from Closing to Idle. In the latter two systems cited above, Disengagement Confirmation is indicated by an explicit interface signal (completion flag). In the public switched network, Disengagement confirmation is an implied event which must be verified by subsequent user action (going "off-hook" and checking for dial tone).

The bit transfer, block transfer, and message transfer functions each serve a distinct purpose in the description of user information transfer performance. The bit transfer function fulfills the need for a "common denominator" to facilitate performance comparison - performance parameter values can always be compared at the bit level. The block transfer function describes performance relative to the information unit that is most relevant to the user in his internal operations - the user information block. The message transfer function provides a formal basis for defining the so-called "secondary" parameters, which describe the long-term availability of a data communication service. The "message" information unit is essentially a sample size, determined on the basis of measurement precision objectives as described in Crow and Miles (1976). ANSI Task Group X3S35 has adopted the term "sample" in preference to "message" to emphasize this fact.

The ANSI Task Group has also suggested a useful clarifying example in defining the end of the access function. In the case where the user interface device is a buffered CRT terminal, the end of access is defined to occur when the first user information character is typed, even though actual transmission of that character may not begin until the next subsequent carriage return is typed

(authorizing transmission of the associated line block). The time difference between these two events can be quite substantial if the user's typing speed is slow, or if text editing is involved.

Figure 3-8 summarizes the intended interpretation of the standard with regard to the above points. The start of input of a block (Event 1) and authorization to transmit that block (Event 2) may or may not be coincident. If they are not, the end of access should be associated with the former event, and the start of block transfer with the latter. Note that the End of Block Transfer is associated with user notification (Event 4) in all cases.

An important characteristic of the primary communication functions defined in FED STD 1033 is that they are user dependent: i.e., their successful completion depends, in general, on events which must be produced by a user. As noted earlier, there is a problem in using parameters based on such functions to describe required system performance: the supplier has no control over user performance, and hence cannot ensure that user dependent parameter values will be met. FED STD 1033 overcomes this problem by explicitly describing the influence of user delay on the primary parameter values by means of separate "ancillary" parameters. The definition and use of these parameters is described in Section 3.3.5.

3.3.3 Outcome Definition

In defining performance parameters for a function, there is a clear need to identify the possible outcomes, or end results, that might occur on any given performance of that function. The third step in developing the FED STD 1033 parameters was to define such a set (or "sample space") of possible outcomes for each of the five primary communication functions. The possible outcomes for any given function can be grouped in three general categories:

1. Successful Performance. The function is completed within a specified maximum performance time, and the result or output is exactly what was intended. A familiar example is successful connection to the correct called party in a voice telephone call.
2. Incorrect Performance. The function is completed within the specified maximum performance time, but the result or output is somehow different from what was intended. A familiar example is connection to a "wrong number" (as a result of a system switching error) in a voice telephone call.
3. Nonperformance. The function is not completed within a specified maximum performance time. A familiar example is the blocking of a voice telephone call attempt by a circuit busy signal.

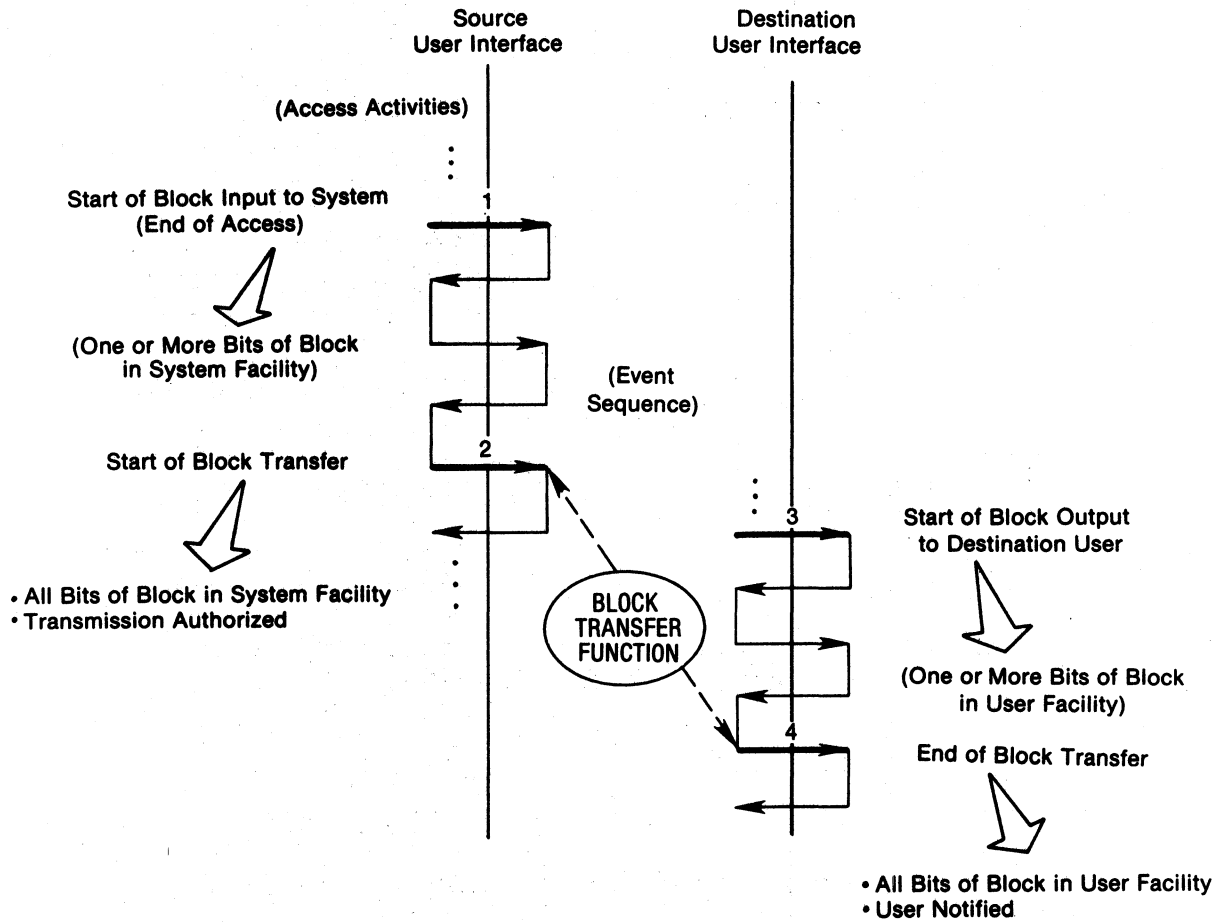


Figure 3-8. Block transfer function definition.

These three outcome categories are significant because they correspond very closely with the three basic performance concerns (or "criteria") most frequently expressed by end users. Successful performance is associated with a user concern with efficiency or "speed"; i.e., in the case of successful performance, the user's concerns center on performance time or rate. Similarly, incorrect performance is associated with a user concern with accuracy; and nonperformance is associated with a user concern with reliability. FED STD 1033 uses these three general performance criteria as an overall framework for organizing the primary performance parameters.

Efficiency, accuracy, and reliability have extremely broad application in the assessment of performance. The questions "how fast, how accurately, and how reliably" apply to the performance of any function, irrespective of what the function "does" or how it is internally accomplished. Examples can be readily found in fields as diverse as energy conversion, manufacturing, transportation, and data processing, among others. The three criteria apply to user functions supported by communications as well as to communication functions, a fact that is helpful in developing "user requirements" for communication service. This subject will be discussed more fully in the Application Manual.

FED STD 1033 divides the Incorrect Performance and Nonperformance categories into more detailed outcomes to enable the definition of specific performance parameters. In general, the system outputs produced during the performance of a function can be "incorrect" in three ways: they may be incorrect in content, they may occur at an incorrect location, or they may include duplicate or other unwanted "extra" information. Failure to produce the expected output of a function can be a consequence, in general, of either system or user nonperformance. Thus, the standard distinguishes six possible outcomes of an individual "trial performance" of a typical primary function:

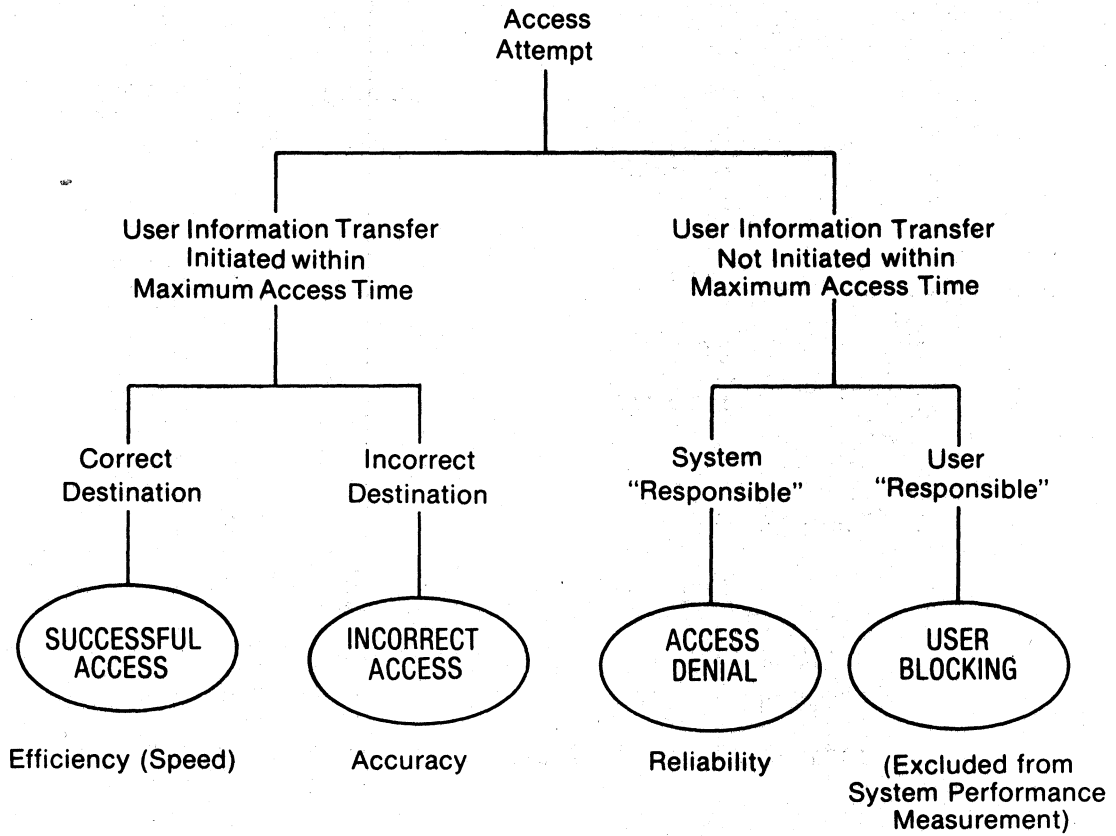
1. Successful Performance. The expected output occurs and is correct in both location (user interface) and content (delivered information).
2. Content Error. The expected output occurs at the correct location, but is incorrect in content.
3. Location Error. The expected output occurs at an incorrect location.
4. System Nonperformance. The expected output does not occur within the maximum performance time, as a result of either issuance of a blocking (busy) signal or excessive delay on the part of the telecommunication system.

5. User Nonperformance. The expected output does not occur within the maximum performance time, as a result of either issuance of a blocking (busy) signal or excessive delay on the part of a user.
6. Extra Event. An unwanted (extra) output occurs in addition to that expected.

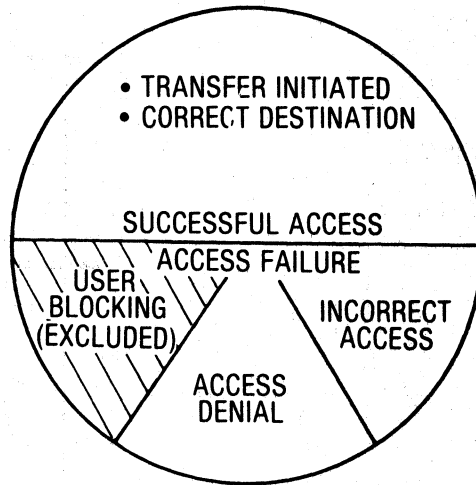
Outcome "sample spaces" for the five primary functions were defined by selecting pertinent outcomes from the above list, and specializing their meaning to the particular function in question. Figure 3-9 shows how this was done in the case of the access function. The standard defines four possible access outcomes: Successful Access, Incorrect Access, Access Denial, and User Blocking. Successful Access is the case where user information transfer is initiated as intended within a specified maximum access time. Incorrect Access is the case where transfer is initiated within the maximum time, but the transaction involves a user other than the one intended by the originator (i.e., a "wrong number"). Access Denial is the case where an access attempt fails as a result of either issuance of a blocking signal or excessive delay by the system. User Blocking is the case where an access attempt fails as a result of either issuance of a blocking signal or excessive delay by a user.

Familiar examples of system and user blocking signals are the "circuit busy" and "user busy" signals in the public switched network. User Blocking outcomes are excluded in defining the access performance parameters. Two of the six "typical function" outcomes defined earlier are not pertinent in the case of the access function: Content Error and Extra Event. In virtually all cases, such errors will result in either Incorrect Access or Access Denial.

Figure 3-10 shows the possible outcomes the standard defines for the block transfer function. Successful Block Transfer is the case where a transmitted block is delivered to the intended destination (within a specified maximum block transfer time), and the delivered block is completely correct in content. Incorrect Block is the case where a transmitted block is delivered to the intended destination, but the delivered block content includes one or more bit errors, additions, or deletions. Misdelivered Block is the case where a transmitted block is delivered to a destination other than that intended by the source. The block may be either correct or incorrect in content. Lost Block is the case where a transmitted block is not delivered to the intended destination within the maximum block transfer time, and the failure is attributable to the communication system. Refused Block is the case where a transmitted block is not delivered to the intended destination within the maximum block transfer time, and the failure is attributable to a



a. Possible Outcomes of an Access Attempt.



b. Sample Space Representation.

Figure 3-9. Access outcome definition.

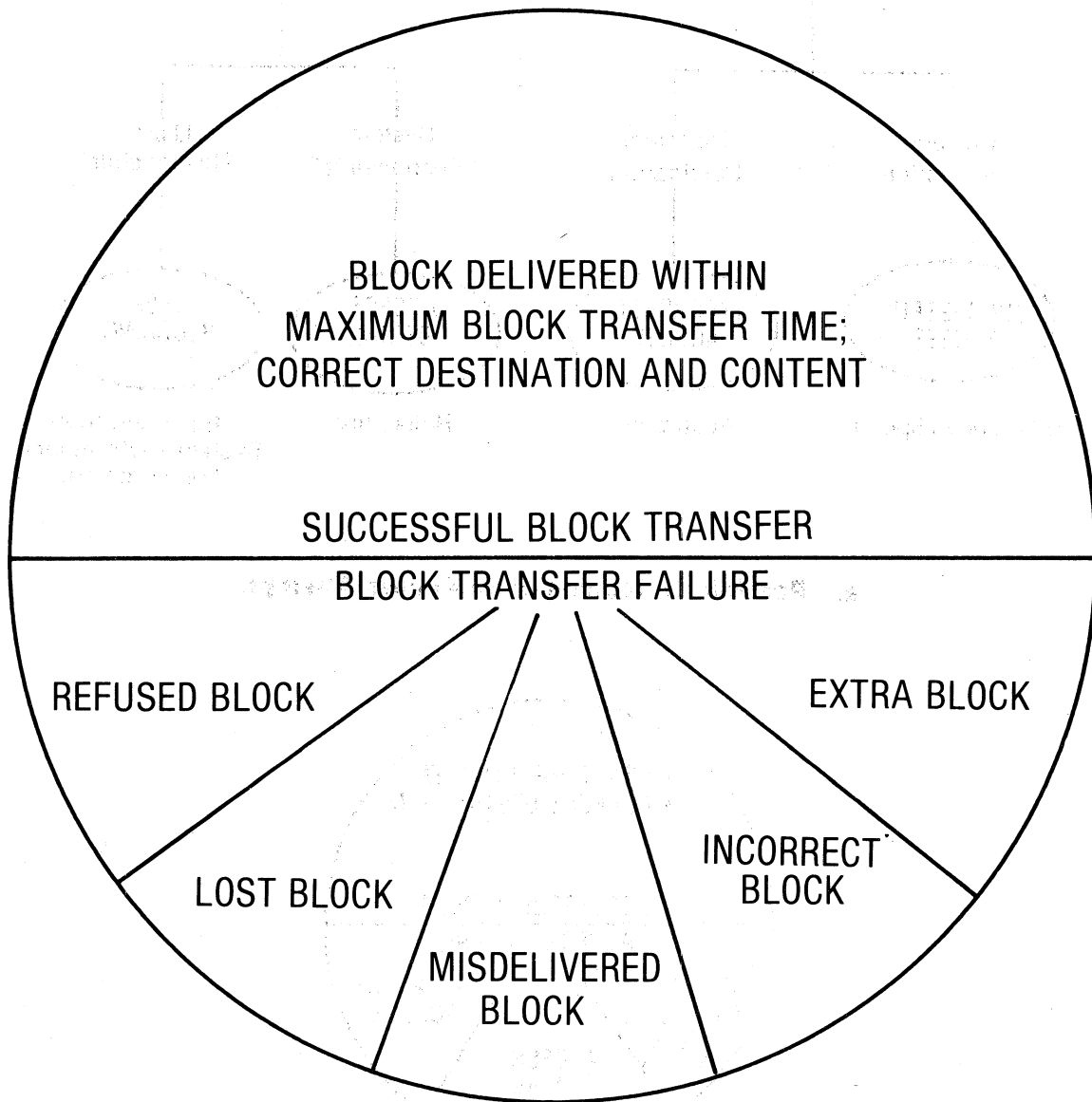


Figure 3-10. Block transfer outcomes.

user.¹⁰ Extra Block is the case where the system delivers to a destination a block that was not output by the source (e.g., a duplicate block).

FED STD 1033 defines the bit transfer and message transfer outcomes in a similar manner, and the corresponding "sample spaces" have the same form as that shown in Figure 3-10. The message transfer outcomes are basically collections of bit and block transfer outcomes; they are used in defining the "secondary" performance parameters as discussed in Section 3.3.5.

Figure 3-11 shows the possible outcomes the standard defines for the disengagement function. Successful Disengagement is defined to occur when the disengaging user and his local "half-system" are returned to the idle state (i.e., are freed to initiate a new transaction) within the specified maximum disengagement time. As noted earlier, this outcome is often indicated by an explicit Disengagement Confirmation signal issued by the system; but in some cases, it must be confirmed by a subsequent Access Request. Disengagement Denial is the case where the disengaging user (and his local half-system) are not returned to the idle state within the maximum disengagement time, and the failure is attributable to the communication system. User Disengagement Blocking is the case where the disengaging user (and his local half-system) are not returned to the idle state within the maximum disengagement time, and the failure is attributable to the user.

Figure 3-12 summarizes the possible outcomes (end results) FED STD 1033 defines for each of the five primary communication functions. Specific examples of each outcome are presented in Section 4.

3.3.4 Parameter Selection

The final step in developing the FED STD 1033 parameters was to select and define a minimum set of parameters to describe system performance relative to each function and outcome. Figure 3-13 illustrates how this was accomplished in the case of the access function. Access performance was described in terms of three specific parameters, one associated with each of the three general performance criteria noted earlier. The standard defines the selected access parameters essentially as follows.¹¹

¹⁰ A destination user might "refuse" a block, for example, by failing to allocate necessary buffer space.

¹¹ Terminology and notation differ slightly from that used in the standard.

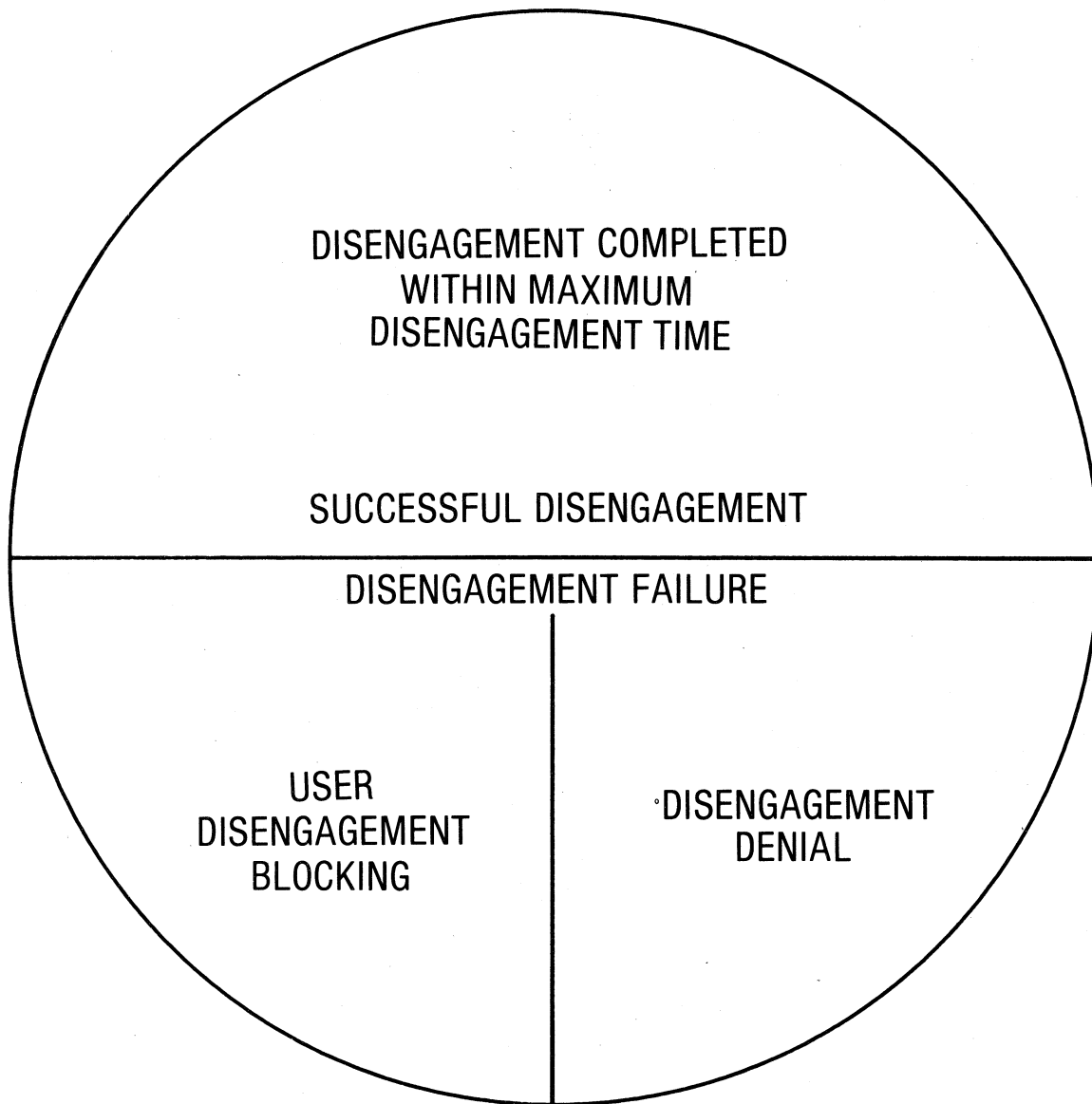


Figure 3-11. Disengagement outcomes.

PRIMARY FUNCTIONS	OUTCOMES INCLUDED IN SAMPLE SPACE					
	SUCCESSFUL PERFORMANCE	CONTENT ERROR	LOCATION ERROR	SYSTEM NON-PERFORMANCE	USER NON-PERFORMANCE	EXTRA EVENT
ACCESS	✓		✓	✓	✓	
BIT TRANSFER	✓	✓	✓	✓	✓	✓
BLOCK TRANSFER	✓	✓	✓	✓	✓	✓
MESSAGE TRANSFER	✓	✓	✓	✓	✓	✓
DISENGAGEMENT	✓			✓	✓	

Figure 3-12. Outcome summary.

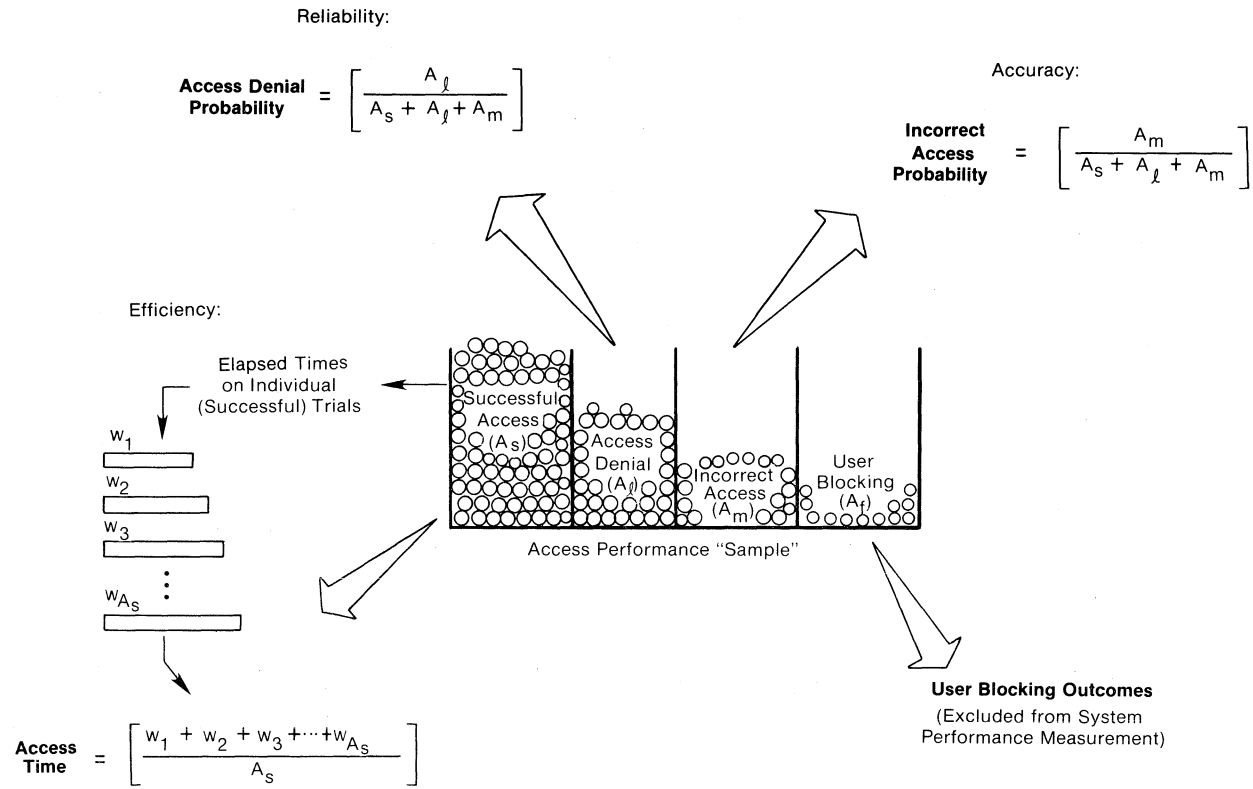


Figure 3-13. Access parameters.

1. Access Time - Average value of elapsed time between the start of an access attempt and Successful Access. Elapsed time values are calculated only on access attempts that result in Successful Access.
2. Incorrect Access Probability - Ratio of total access attempts that result in Incorrect Access (i.e., connection to an unintended destination) to total access attempts included in an access performance sample (excluding User Blocking outcomes).
3. Access Denial Probability - Ratio of total access attempts that result in Access Denial (e.g., system blocking) to total access attempts included in an access performance sample (excluding User Blocking outcomes).

A key aspect of the FED STD 1033 parameter definitions is their expression in mathematical form. As noted earlier, this approach eliminates the ambiguity associated with purely narrative definitions, and also provides a standard procedure for calculating performance parameter values. The mathematical parameter definitions are based, in each case, on the concept of an access performance "sample" - i.e., a large number of successive access trials distributed, like apples, in appropriate outcome "bins". Each Successful Access outcome has an associated elapsed time value (the total time required to complete that particular attempt).

Values for the access parameters may be calculated directly from the data in an access performance sample. The value of the efficiency parameter Access Time is calculated by adding the individual elapsed times (w_i) for all A_s Successful Access outcomes, and then dividing by A_s . The value of the accuracy parameter Incorrect Access Probability is calculated by dividing the total number of Incorrect Access outcomes (A_m) by the total number of outcomes in the access sample, excluding the User Blocking outcomes - i.e., dividing A_m by $(A_s + A_\ell + A_m)$. Similarly, the value of the reliability parameter Access Denial Probability is calculated by dividing the number of Access Denial outcomes A_ℓ by $(A_s + A_\ell + A_m)$. User Blocking outcomes are excluded in calculating the access failure probabilities to ensure the comparability of values measured under different usage conditions.

The preceding section referred to a "maximum access time" beyond which an access attempt is declared a failure for performance assessment purposes. To ensure comparability, the standard fixes this "timeout" point at three times the Access Time specified for the service: i.e., three times the delay the user "expects to see" on any given access attempt. Note that this timeout constant has significance only in the assessment of performance; access attempts that extend beyond the timeout point need not be abandoned. Additional characteristics of the Access Time distribution (e.g., variance or 95-percent points) may also be of interest in some cases.

The same general approach used in the access case was followed in selecting and defining performance parameters for the user information transfer and disengagement functions. A separate probability parameter was defined to express the likelihood of each possible failure outcome; and an "average elapsed time" parameter was defined, in each case, to quantify successful performance. Bit and Block Transfer Rate and Rate Efficiency parameters were also defined, to express performance from the standpoint of "throughput" and resource utilization. A complete list of the primary performance parameters specified in FED STD 1033 is provided in Section 3.3.6.

3.3.5 Secondary and Ancillary Parameters

Although the primary parameters described above provide a relatively detailed description of data communication performance, they fall short of completeness in two respects:

1. They do not provide the kind of macroscopic, long-term performance view users traditionally associate with the concept of availability.
2. They are user dependent, and thus cannot be applied directly in situations where it is necessary to describe unilateral system performance.

A small set of additional "secondary" and "ancillary" performance parameters were included in the standard to meet these needs.

Figure 3-14 illustrates the approach used in defining the secondary (availability) parameters. Very briefly, the sequence of transmissions between a specified pair of users is divided into a series of consecutive performance measurement periods or samples, each corresponding to the "message" information unit defined earlier. Values for each of five "supported" primary performance parameters are calculated on the basis of the outcome of each successive message transfer function. The calculated values are compared with corresponding outage thresholds to define the "secondary outcome" of that trial performance of the message transfer function as either Operational Service state or Outage state. Finally, appropriate time and probability parameters are defined to describe the resulting sequence of availability state transitions.

In assessing availability performance, the service connecting a user pair is observed only during the User Information Transfer (UIT) phase: i.e., the time, during each transaction, between Successful Access and disengagement of the last committed user (Seitz, 1980). There is no correspondence, in general, between the duration of an individual user information transfer phase and the length of a "message". As noted earlier, the "message" information unit is essentially a

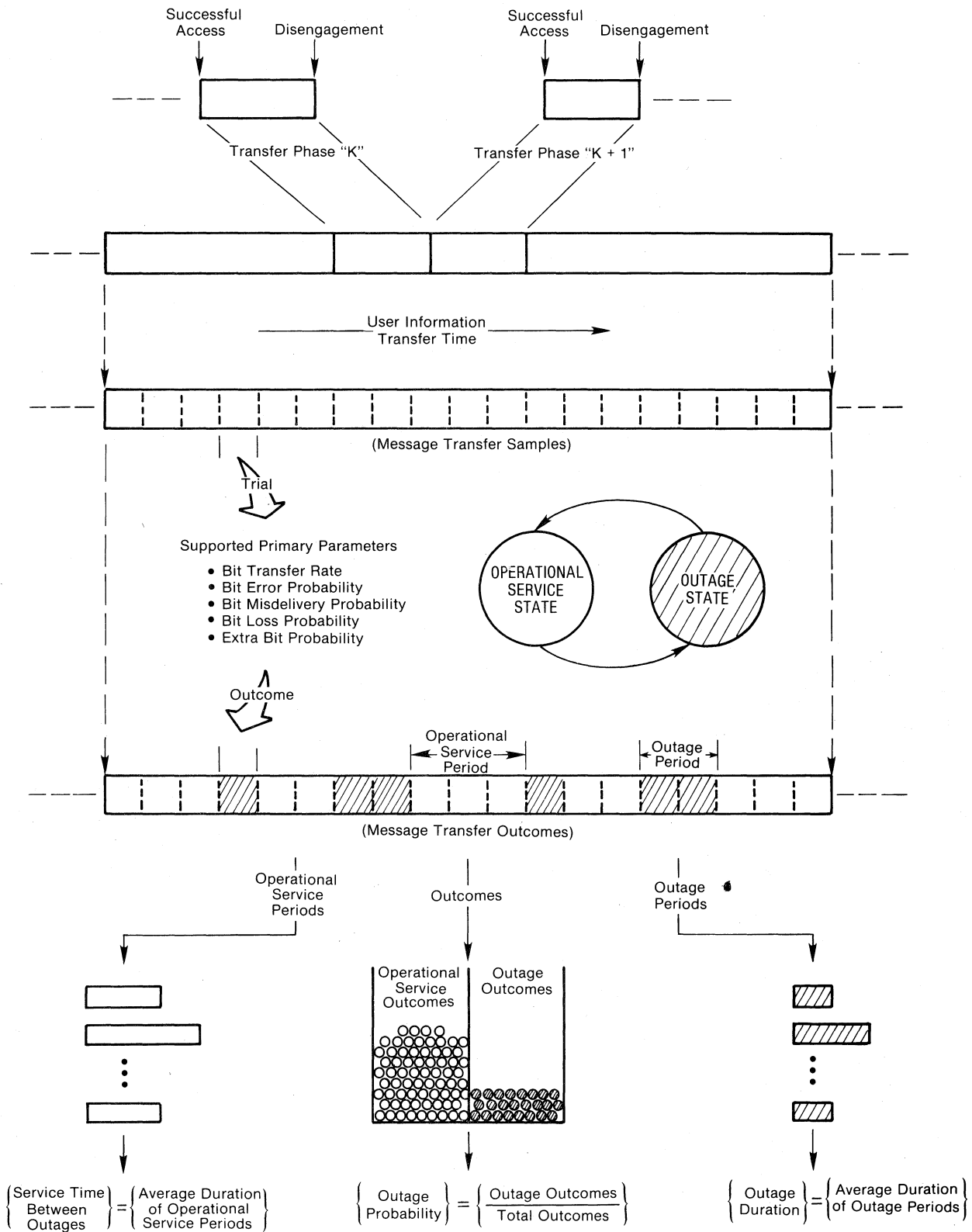


Figure 3-14. Secondary parameter development.

sample size (a fixed number of transferred bits); and a given "message" may thus span more than one transaction.

Five primary user information transfer parameters are defined as supported performance parameters: the four bit transfer failure probabilities (Bit Error Probability, Bit Misdelivery Probability, Bit Loss Probability, and Extra Bit Probability) and Bit Transfer Rate. Outage thresholds for the supported performance parameters are defined as a function of the corresponding "nominal" values specified for the service as follows:

1. The outage threshold for Bit Transfer Rate is defined as one-third (1/3) of the nominal Bit Transfer Rate.
2. The outage thresholds for the four bit transfer failure probabilities are defined as a function of the corresponding nominal probability values by expressing the nominal value as a power of ten (for example, 10^{-6}) and then dividing the exponent by two (producing, for example, a threshold value of 10^{-3}). This procedure corresponds to taking the square root of the nominal (specified) probability value.

A service is defined to have been in the Operational Service state (during the preceding performance measurement period) whenever the measured values for all supported parameters are better than their associated outage thresholds. A service is defined to have been in the Outage state whenever the measured values for one or more supported parameters are worse than their associated thresholds. This definition process produces (in the measurement record) a sequence of alternating Operational Service and Outage periods, each having a known duration in the User Information Transfer (UIT) time. Each period comprises an integer number of messages or samples. The secondary performance parameters provide a statistical description of this two-state random process. They are defined in the standard essentially as follows.

Service Time Between Outages - Average value of elapsed User Information Transfer time between entering and next leaving the Operational Service state.

Outage Duration - Average value of elapsed User Information Transfer time between entering and next leaving the Outage state.

Outage Probability - Ratio of total message transfer attempts resulting in the Outage state to total message transfer attempts included in a secondary parameter measurement.

These parameters are termed "secondary" to emphasize the fact that they are defined on the basis of measured primary parameter values, rather than on the basis of direct observations of interface events.

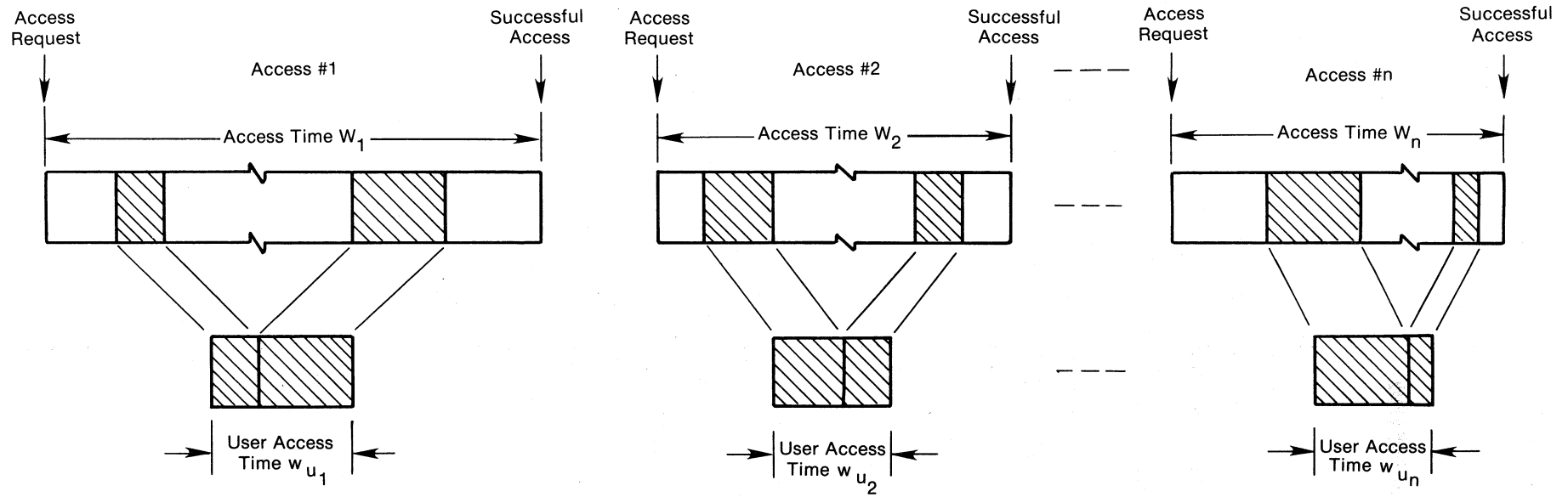
The final category of FED STD 1033 parameters to be described are the ancillary parameters. As discussed earlier, the primary communication functions defined in FED STD 1033 are, in general, user dependent; and there is a need, then, for a quantitative means of expressing the influence of user delay on the primary parameter values. The ancillary parameters fulfill that need.

Very briefly, the ancillary parameters are developed by dividing the total performance time for an associated primary function into alternating periods of system and user "responsibility"; and then calculating the average proportion of total performance time for which the users are "responsible". As a simple illustration, consider the voice telephone access example discussed earlier (Fig. 3-3). The total performance time for the access function is the time between the calling user's off-hook action and the called party's answer. This total performance time can be divided into alternating periods of system and user responsibility by noting, at any time, which entity must produce the next interface event. During the period between off-hook and dial tone, the system is responsible; during the period between dial tone and positioning of the first dialed digit, the user is responsible; and so on. The ancillary parameter User Access Time Fraction expresses the average proportion of total Access Time attributable to the user activities.

Figure 3-15 illustrates the approach used in defining the ancillary parameters in more detail, again using the primary function of access as an example. The figure depicts a series of successful access attempts, each having a total access time w and a total user access time w_u . The latter quantity represents the total access time attributable to user responsibility on each particular trial. The ancillary parameter User Access Time Fraction is calculated by adding the user access time values over a suitable number of successful access attempts, and then dividing by the corresponding sum of the total access times. Only Successful Access outcomes are considered in estimating User Access Time Fraction in order to avoid biasing the average with unrepresentative values.

A similar approach is used in defining ancillary performance parameters for the block transfer, message transfer, and disengagement functions. No ancillary parameter is defined for the bit transfer function, since its values can be inferred from the corresponding block transfer parameter. The standard thus defines a total of four ancillary performance parameters: User Access Time Fraction, User Block Transfer Time Fraction, User Message Transfer Time Fraction, and User Disengagement Time Fraction.

In describing the ancillary parameters, it should be noted that there are cases where it is not possible to place unilateral responsibility for completing a



$$\left\{ \begin{array}{l} \text{User Access} \\ \text{Time Fraction} \end{array} \right\} = \left\{ \begin{array}{l} \text{Average User Access Time} \\ \text{Average Total Access Time} \end{array} \right\}$$

$$= \left\{ \frac{w_{u_1} + w_{u_2} + \dots + w_{u_n}}{W_1 + W_2 + \dots + W_n} \right\}$$

Figure 3-15. Ancillary parameter development - access example.

function on the user or the system. Activities may proceed concurrently at the two interfaces, with the user responsible at one and the system responsible at the other. The message transfer function provides a straightforward example. A "message" typically consists of many separate blocks. At a given moment during transfer of a message, the source user may be inputting one block to the system while the system is outputting a previous block to the destination. Neither entity is solely responsible, at that moment, for the overall function of message transfer; responsibility is "split" equally between the two.

FED STD 1033 accommodates such concurrency by defining responsibility separately for each user interface, and counting intervals of split responsibility at half their actual value in calculating total user performance time. Thus, if a message transfer interval included 2 minutes of unilateral user responsibility and 2 minutes of split responsibility, the total message transfer time attributed to the users would be 3 minutes. (Split responsibility is nonexistent in many data communication transactions, and it can often be disregarded with negligible effect on a performance specification.)

The ancillary parameters have two specific uses:

1. They enable calculation of "user-independent" values for the associated efficiency parameters - i.e., the values that would be observed if all user delays were zero.
2. They provide a basis for identifying the entity "responsible" for timeout failures - the user or the system.

Each of these uses is described more fully in Section 4.6.

3.3.6 Problem Solutions - Summary

We noted earlier that the development of FED STD 1033 required the solution of three key technical problems. The technical approach adopted in the standard provides a solution to each of these problems, as summarized below.

1. System Dependence. The standard solves this problem through the expedient of the user-oriented performance model. The model reduces all user/system interactions to a small set of general reference events which can be identified in any system; and the performance parameter definitions are then based on these system-independent events.
2. Detailed Parameter Definition. The standard solves this problem by using sample spaces and mathematical equations as the major parameter definition tools. Sample spaces encourage the analyst to consider, and carefully define, all relevant outcomes of a performance trial. Equation definitions eliminate the ambiguity often associated with purely narrative definitions.

3. User Dependence. The standard solves this problem through the use of the ancillary performance parameters. These parameters provide a basis for "factoring out" user influence on the waiting time, time rate, and rate efficiency parameters; and a means of determining whether the user or the system is "responsible" for timeout failures.

Figure 3-16 summarizes the performance parameters ultimately selected for inclusion in Interim Federal Standard 1033. A total of 26 parameters were selected, including 19 primary parameters, 3 secondary parameters, and 4 ancillary parameters. Each selected parameter is described in detail in the following section.

4. UNDERSTANDING THE PARAMETERS

4.1 Introduction

Suppose you were handed the parameter table of Figure 3-16, with no prior explanation, and asked to use it in specifying a communication service requirement. What questions would you ask about each parameter before beginning the specification? For most potential users, the key questions about each parameter would include the following:

- What is the meaning of this parameter, in simple, straightforward, user-oriented terms? How is it related to other widely-used performance parameters?
- Why is the value of this parameter significant to data communications users? What are its best and worst possible values, and what are their implications?
- What typical values might be specified for this parameter, in characterizing (a) performance requirements for familiar user applications, and (b) performance capabilities of existing data communication systems and services?
- How do the values for this parameter influence, and how are they influenced by, the key decisions in communication system design?

This section answers these questions by means of tutorial "essay descriptions" of the FED STD 1033 parameters. The individual parameter essays are organized by function and category in the manner suggested above: i.e., access parameters, user information transfer parameters, disengagement parameters, secondary parameters, and ancillary parameters. A separate essay is provided for each primary parameter, with the exception that corresponding bit- and block-oriented transfer parameters (e.g., Bit Error Probability - Block Error Probability) are described together. The secondary parameters and the ancillary parameters are each described

FUNCTION	PERFORMANCE CRITERION			
	EFFICIENCY	ACCURACY	RELIABILITY	
ACCESS	• ACCESS TIME	• INCORRECT ACCESS PROBABILITY	• ACCESS DENIAL PROBABILITY	• USER ACCESS TIME FRACTION
BIT TRANSFER	• BIT TRANSFER TIME	• BIT ERROR PROBABILITY • BIT MISDELIVERY PROBABILITY • EXTRA BIT PROBABILITY	• BIT LOSS PROBABILITY	• USER BLOCK TRANSFER TIME FRACTION
BLOCK TRANSFER	• BLOCK TRANSFER TIME	• BLOCK ERROR PROBABILITY • BLOCK MISDELIVERY PROBABILITY • EXTRA BLOCK PROBABILITY	• BLOCK LOSS PROBABILITY	• USER MESSAGE TRANSFER TIME FRACTION
MESSAGE TRANSFER	• BIT TRANSFER RATE • BLOCK TRANSFER RATE • BIT RATE EFFICIENCY • BLOCK RATE EFFICIENCY	• OUTAGE PROBABILITY		• USER DISENGAGEMENT TIME FRACTION
DISENGAGEMENT	• DISENGAGEMENT TIME	• DISENGAGEMENT DENIAL PROBABILITY		
SERVICE CONTINUATION	• SERVICE TIME BETWEEN OUTAGES			
SERVICE RESTORAL	• OUTAGE DURATION			

Legend




	Primary Parameters
	Secondary Parameters
	Ancillary Parameters

Figure 3-16. FED STD 1033 parameters.

in a single essay to emphasize interdependencies and definitional similarities. Readers are referred to the standard and its supporting reports for more rigorous parameter definitions, and for application details.

4.2 Access Parameters

Requesting access to a communication service is a little like going to the post office to mail a letter. Your objective, in each case, is not to spend time in the service facility (waiting on hold or standing in line behind other customers); that process simply wastes time that could be used productively. Your objective, in each case, is to get your message started on its way to the intended destination as soon as possible.

An ideal communication service (or an ideal postal service) would accept your message and start it on its way to the intended destination immediately, with no delay, every time you requested service. As a rate payer or tax payer, you realize that such an ideal service might be prohibitively expensive. But you still judge the service, as an end user, on the basis of how closely it approaches that ideal.

The fundamental user concerns about service performance are also similar in the two cases:

- Efficiency - How long will I have to wait to get my message started on its way, assuming I am successful in doing this?
- Accuracy - What is the likelihood that the service facility will process my service request incorrectly (e.g., misconnection or a wrong zip code), thus establishing an information path that directs my message to the wrong destination?
- Reliability - What is the likelihood that I will be denied service on any given request (e.g., the telecommunication system issues a "circuit busy" signal, or the postal clerk closes his window and leaves for the day)?

FED STD 1033 defines three primary performance parameters which directly express these user concerns: Access Time, Incorrect Access Probability, and Access Denial Probability.

4.2.1 Access Time

Access Time is the average time the user must wait, after requesting communication service, for the system to begin accepting his/her information for transmission. Computation of Access Time begins on issuance of an Access Request, or its implied equivalent, at the originating user interface; and ends on the first

subsequent transfer of user information from a source user to the system. Access Time values are calculated only on access attempts that result in Successful Access.

The Access Request event takes many different forms. Three examples of explicit Access Request signals have been cited earlier (Section 3.3.2). An Access Request can also be implicit, e.g., in the case where a user asks the system to "poll" him for possible messages at some specific future time. In the latter case, an access attempt is defined to begin at the prearranged time, even if the system does not issue a polling signal at that time.

Successful Access is defined to occur when at least one bit of user information is transferred from a source user to the system within the specified maximum access time. In the case of circuit-oriented transactions, there is an additional requirement: the intended nonoriginating user must have been contacted and committed to the transaction prior to the start of user information transfer. This requirement distinguishes Successful Access outcomes from Incorrect Access outcomes.

The relationship between Access Time and the traditional telephone switching parameters Dial Tone Delay and Time to Receipt of Audible Ringing has been discussed in Section 3.2. To recap briefly, Access Time describes the total time between off-hook and answer; the latter two parameters describe specific intervals of system performance within that time.

Access Time is closely related to another commonly-used switching parameter, Connection Establishment Time. This parameter has been defined as follows (ANSI, 1980):

"Connection Establishment Time represents the time interval required to establish an information transfer channel to the desired destination... Connection Establishment Time begins when network service is requested by going off-hook or activating the call request (CRQ) function at the DTE-DCE interface. It ends when clear to send (CB) or equivalent function is activated at the DCE-DTE interface at either the calling or called station, whichever transmits first."

Connection Establishment Time differs from Access Time in two major respects:

1. The starting and ending events are defined occur at the DTE/DCE interface rather than at the end user/communication system interface.
2. The ending event is a system-generated clear to send signal rather than the actual start of user information transfer.

The events used in defining Access Time are more appropriate in a user-oriented standard because they are observable at the end user interfaces and are system

independent. The difference in timing between the two event pairs can be substantial, particularly in "layered architecture" systems like that depicted in Figure 3-5c.

Access Time also has a close kinship with the "average waiting time" parameter defined in queueing theory (e.g., see Kleinrock, 1976). The latter parameter describes the average time a customer must spend waiting in queue, on any given arrival, before receiving some desired service. In the case of telecommunications, transfer of user information is the desired service; issuing an Access Request denotes queue entry; and the start of user information transfer denotes the end of waiting and the beginning of service.

Access Time differs from "average waiting time" (and the other time parameters mentioned) in one important respect: unlike them, it is the average of a truncated distribution. Figure 4-1 illustrates the meaning of this difference. If we measure a large number of individual delay values and plot the relative frequency of occurrence of each possible value, the result is a histogram or distribution of delay values. In general, such a distribution will be unbounded on the right, since extremely long delays will occasionally occur. It is desirable to exclude such "outlying" values in calculating an average for two practical reasons:

1. Their observation requires, in the limit, infinite patience on the part of the observer.
2. They can unduly influence an average because of their large magnitude.

FED STD 1033 excludes abnormally long delay values in calculating Access Time by truncating, or cutting off, the Access Time distribution at a value three times the "nominal" value specified for the service.¹² The standard counts access attempts which last longer than this "maximum access time" as failures for performance assessment purposes, and describes their relative frequency of occurrence with the parameter Access Denial Probability. The same approach is used in defining all other time averages in the standard. The timeout constant 3 was chosen, somewhat arbitrarily, as a value which would (1) be generally consistent with user expectations about service performance, and (2) include most of the area under a typical delay time distribution. For further discussion, see Crow (1979).

Why is Access Time, an average wait for service, significant to data communications users? These are two distinct reasons:

¹²The "nominal" value is the mean before truncation.

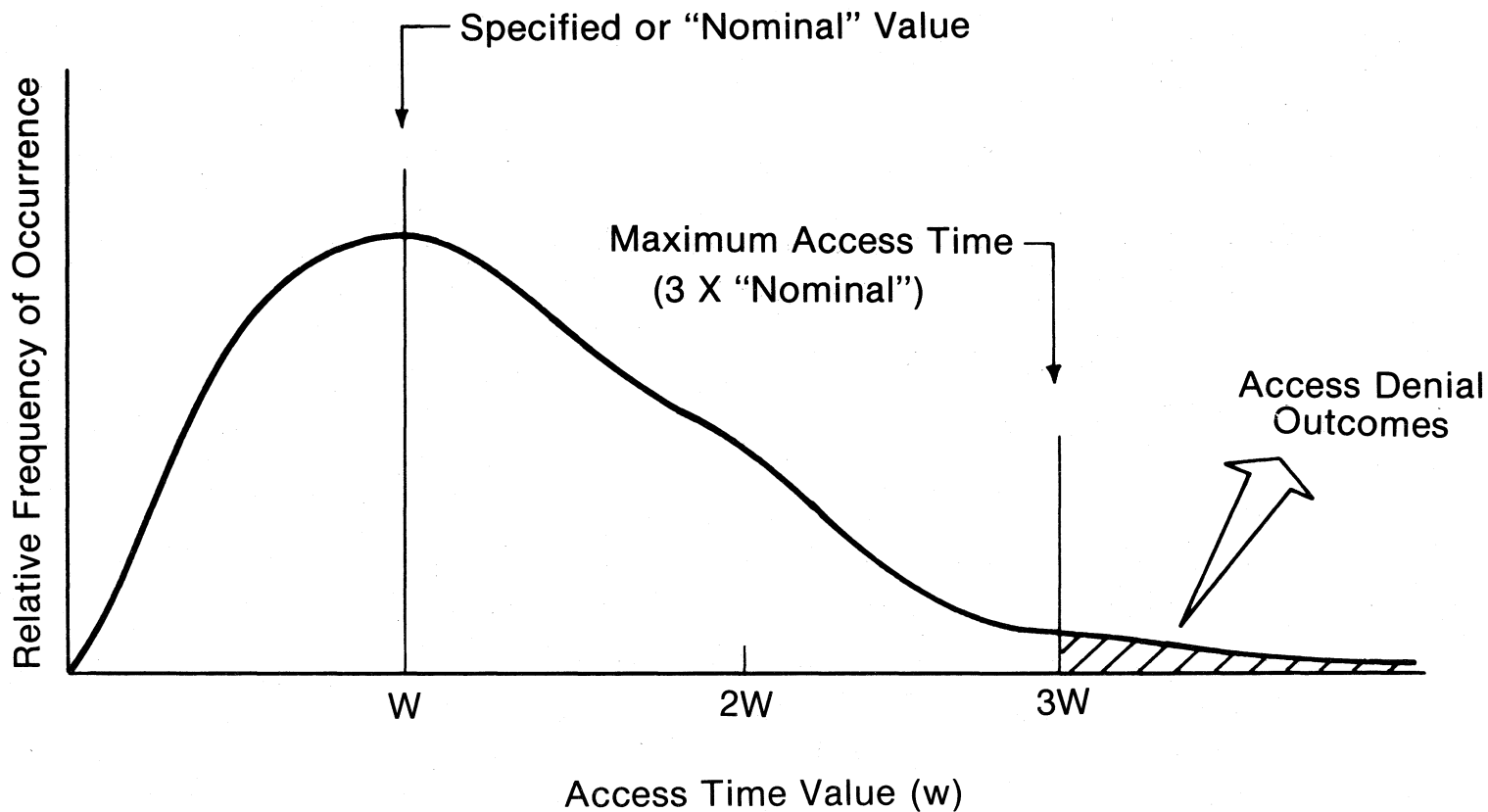


Figure 4-1. Truncation of the access time distribution.

1. Time is Money. Users value their time, and quite properly view time spent establishing communication as unproductive time. The cost to the user of communication delay can often be expressed in direct monetary terms (e.g., value of goods not produced).
2. Information "Ages". Virtually all information becomes less relevant, and therefore less valuable, with the passage of time. As an extreme example, a message warning two aircraft of an impending collision has no value after the collision has occurred. Last week's gold prices are of little value to an investor who must make a decision today.

In essence, Access Time is the price we pay for the economic benefit of sharing a communication resource with other users.

As noted earlier, an ideal communication service would begin transmitting a user's information as soon as access was requested; and the best possible value for Access Time is therefore zero. There is no theoretical upper limit on an Access time specification, but measured values for the parameter can never exceed the "three times nominal" timeout described earlier. Obviously, extremely long Access Time values imply a service that is essentially useless.

Appropriate user requirements for Access Time vary widely as a function of the user application. At the lower extreme are critical, real-time applications such as military command and control. Access Time requirements here may be in the millisecond range. At the upper extreme are routine record communications and electronic mail: if the only requirement is for "next-day delivery," a delay of several hours between preparation and transmission of a message may be quite acceptable. A recent ITS study recommends 0.9-percentile¹³ Access Times in the range of 0.15 to 4.0 seconds for interactive Autodin II users (Nesenbergs, et al., 1980). An average Access Time of 15 seconds has been specified in a Request for Quotations (RFQ) recently issued by the Environmental Protection Agency for the communications portion of a nation-wide time-sharing network (EPA, 1980).

Access Time values are strongly influenced by system design. Access delay is primarily caused by resource sharing (i.e., switching), and it is therefore not surprising that dedicated, nonswitched services provide the shortest Access Times. One might expect Access Time to be zero in dedicated services, but this is often not the case. Even though communication facilities are not shared, they may not be used continuously. If they are not, a short access delay will often be encountered at the beginning of each usage period while the system synchronizes equipment

¹³0.9-percentile values are exceeded 10% of the time. The corresponding averages can be considerably lower.

at the two user locations (Gray, 1972; Kimmett and Seitz, 1978). Such delays will normally be in the millisecond range if no user dependence is involved.

FED STD 1033 distinguishes two general categories of resource-sharing communication services: circuit-oriented and message-oriented. Circuit-oriented services require that the intended "nonoriginating user" (called party) be contacted and committed to a transaction prior to the start of user information transfer. Traditional circuit-switched and modern virtual-circuit services fall in this category. Message-oriented services allow user information transfer to begin without such a commitment. Traditional message-switched and modern datagram services fall in this category.

Access Times are normally longer in circuit-oriented services than in message-oriented services because the process of obtaining destination user commitment takes time. Typical Access Time values in message-oriented services are in the range of 1 to 5 seconds. Typical Access Time values in circuit-oriented services are in the range of 5 to 10 seconds. An average Access Time of 7 seconds has been measured for ARPA network terminal operators using Telnet, a virtual-circuit protocol (Payne, 1978).

The decision to defer destination user commitment to the transfer phase increases the transfer delays in message-oriented services, as discussed in Section 4.3.1. The effect of user dependence on Access Time is addressed separately in Section 4.6.

4.2.2 Incorrect Access Probability

Incorrect Access Probability expresses the likelihood that user information will be transmitted on an improper circuit path as a result of a system error during the access process. As noted earlier, it is defined as the ratio of total Incorrect Access outcomes to total access attempts included in a performance sample, excluding access attempts that fail as a result of User Blocking.

Incorrect Access is essentially the case of a "wrong number." It occurs when the system establishes an improper physical or virtual circuit connection during access, and then does not correct the error before the start of user information transfer. Incorrect Access can occur only in circuit-oriented services, since no physical or virtual circuit is established between end users in message-oriented services. Incorrect Access is distinguished from Successful Access by the fact that the intended nonoriginating user is not contacted and committed to the session prior to the start of user information transfer.

What kinds of system errors cause Incorrect Access? Perhaps the simplest cause is a transmission error in communicating the signaling information from the

originating user to the first switch, or between switches. A second possible cause is a switch error in translating the signaling information into (for example) a physical crosspoint connection. The latter errors will normally be infrequent and random (e.g., as a result of a marginal switch component); but they can also be systematic in some cases (e.g., as a result of an improper numbering change). Obviously, systematic Incorrect Access outcomes can also be caused by various deliberate "spoofing" and tampering actions.

Incorrect Access is closely associated with what common carriers and switch manufacturers refer to as "mishandled" or "misprocessed" calls (Kobylar and Malec, 1973; Malec, 1975). However, Incorrect Access Probability differs from the "misprocessed call" probability in two respects:

1. Misprocessed calls typically include calls that are not completed (i.e., the switch does not respond) as well as calls that are mis-connected.
2. Misprocessed call probability describes the performance of a switch rather than that of an end-to-end system.

The latter difference can be very significant in systems which provide "automatic answerback" or other connection verification features, as discussed below.

Incorrect Access also has an obvious association with the concept of misdelivery. The following definition of misdelivery is typical of those encountered in the literature (DCA, 1975):

"Misdelivery is defined as the delivery of a segment [i.e., message] in violation of the originally specified addressing information."

Incorrect Access and misdelivery are often related as cause and effect in circuit-oriented systems. If a system establishes a circuit connection to an incorrect (but compatible) destination during access, and does not detect the error prior to the start of user information transfer, it is highly likely that at least some user information will be misdelivered to that destination. It should be recognized, however, that Incorrect Access does not invariably result in misdelivery. The reason this is so is that the test for Incorrect Access is a negative test: i.e., Incorrect Access is declared (in circuit-oriented systems) if the intended nonoriginating user is not contacted and committed to the transaction prior to the start of user information transfer. This test does not distinguish between the case where some other user was contacted and the case where the commitment step in access was somehow "short-circuited" by the system, with no other user contacted.

Misdelivery will normally occur in the former case; but in the latter case, the end effect will be data loss.

These two possible consequences of Incorrect Access, misdelivery and loss, make Incorrect Access Probability very significant to data communications users. In the case of misdelivery, the risk to the source user is twofold:

1. The data may be delivered to a destination user who has the desire and capability to exploit it to the source's disadvantage. This risk is particularly great when sensitive information is transmitted over public telecommunication facilities without benefit of encryption. An example of sensitive, but unclassified data would be details of prospective financial transactions.¹⁴
2. The source user may be led to believe that his information has been delivered to the intended destination, when in fact it has not. His subsequent actions will then be based on a false assumption. As an example, a weather information source might incorrectly assume that a storm warning has been delivered to a threatened ship, and thus stop transmitting.

Only the latter risk is applicable in the case of loss, since the user information is not delivered to any destination.

Like all probabilities, Incorrect Access Probability has possible values between zero and one. A value of zero would indicate that Incorrect Access is impossible, a situation that can only be realized in systems that perform no switching during the access phase. Such systems include message-oriented systems, which perform their switching during the user information transfer phase, and non-switched or "dedicated" systems, which perform no switching at all. A measured Incorrect Access Probability value of one would indicate that Incorrect Access is certain (i.e., always occurs); such values are possible in the case of systematic switching errors, as noted above.

Quantitative data on user requirements for Incorrect Access Probability is scarce. Nesenbergs et al., (1980) suggest a range of values in the neighborhood of 10^{-10} for interactive Autodin II users; a somewhat more stringent value (10^{-11}) is specified in the Autodin II System Performance Specification (DCA, 1975). Neither estimate appears to be based on a quantitative user impact assessment. Incorrect Access may be no more than a nuisance in a benign communication environment; in such situations, users may well specify a relatively arbitrary, easily attainable value (e.g., 10^{-3}). A value of 10^{-5} is specified in the EPA RFQ cited earlier (EPA, 1980).

¹⁴NTIA is currently conducting a program to assess communications protection options (Lemp, 1980).

System performance data on Incorrect Access Probability is also relatively scarce. As noted earlier, Incorrect Access can be caused by errors either in transmitting or in interpreting the signaling information. Errors in transmitting signaling information are much more frequent in older systems employing direct current, multi-frequency, or single frequency switching than in modern systems employing common channel signaling (e.g., 10^{-4} vs. 10^{-8}). A value of 10^{-4} appears to be a typical objective for "mishandled call" probability in a single switch (Kobylar and Malec, 1973). In estimating Incorrect Access Probability, this number would be reduced by the fact that not all "mishandled" calls are misconnected, and increased by the fact that a typical circuit normally involves several tandem switches. A value of 10^{-5} is probably a reasonable estimate of the likelihood of misconnection as a result of switching error in a typical circuit-switched system.

Of course, errors in signaling and switching may or may not cause Incorrect Access. The likelihood of Incorrect Access given such an error depends on two factors:

1. Whether the error results in contact with a terminal (or terminal function) compatible with that of the called party intended.
2. Whether the system provides circuit verification techniques such as automatic answerback (e.g., see AT&T, 1968).

The likelihood of connection with a compatible terminal depends on the mix of terminals in the network in question. Answerback schemes can reduce Incorrect Access Probability by several orders of magnitude.

In general, "virtual circuit" switching systems have a lower Incorrect Access Probabilities than conventional "physical circuit" switching systems as a result of their more effective use of end-to-end error control. Autodin II, for example, employs a 32-bit Cyclic Redundancy Check (CRC) on transmitted data; by a familiar "folk theorem" cited in Nesenbergs et al. (1980), this CRC check should provide an undetected error rate for circuit establishment messages better than 2^{-32} (10^{-10}).

One practical limitation of the parameter Incorrect Access Probability should be noted in conclusion: it does not consider situations where an unintended destination is contacted as a result of a user error in inputting the addressing information (e.g., misdialing). Such errors should be considered in establishing user requirements for system performance. This subject will be discussed more fully in the Application Manual.

4.2.3 Access Denial Probability

Access Denial Probability expressed the likelihood that a system will fail to provide the user with access to a communication service on any given request. It is defined as the ratio of total Access Denial outcomes to total access attempts included in a performance sample, excluding access attempts that fail as a result of User Blocking.

Access Denial (also termed system blocking) can occur in two basic ways: (1) the system issues a blocking signal to the originating user during the access period, thereby terminating the access attempt; or (2) the system delays excessively in responding to user actions during the access period, with the result that user information transfer is not initiated within the maximum access time.

What is a "system blocking" signal? In essence, it is a system's way of telling a user that it cannot provide him with communication service on a particular request because some required system facility is currently unavailable. The required facility (e.g., trunk circuit) may be unavailable because it is serving another user, or because it is in an outage condition; the two possibilities often cannot be distinguished at the end user interface.

As defined in FED STD 1033, a system blocking signal constitutes a definite denial, rather than a delay or deferral, of an access attempt. A familiar example of a system blocking signal is the two cycle-per-second "circuit busy" signal in the public switched network. Such a signal tells the user that the current access attempt will not succeed, no matter how long he hangs on; his best alternative is to hang up and try again. System blocking (denial) signals should be distinguished from signals which merely delay Successful Access (e.g., the familiar "all reservation clerks are busy - please do not hang up" recording).

Systems experiencing congestion or outage may not respond, or may delay excessively in responding, to a user's Access Request. Virtually everyone has experienced such a situation at one time or another in making a long distance telephone call. The system gives a few promising "clicks" after dialing, and then seems to "go dead"; more often than not, the optimistic user who waits is eventually disconnected. FED STD 1033 defines such nonresponses as Access Denials if they persist longer than three times the "nominal" Access Time for the service, as defined earlier. Access Denials are distinguished from User Blocking outcomes by comparison of ancillary parameter values as discussed in Section 4.6.

Access Denial Probability is closely associated with what is known as the Grade of Service or Blocking Probability in circuit-switched systems. In general, such systems cannot economically provide access to all users during the worst case

loading period or "busy hour"; instead, they are designed to serve all but a certain (small) fraction of calls attempted during that period. The fraction, P , of call attempts not served by a circuit-switched system during the busy hour is its Grade of Service or Blocking Probability. The symbol $P.01$ indicates that one call in a hundred will be blocked; $P.04$ indicates that four calls in a hundred will be blocked; and so on. Customers accept a small Blocking Probability in exchange for the economic benefits of resource sharing; but if the Blocking Probability is too high, they may abandon the service in favor of more reliable alternatives.

Access Denial Probability is also closely associated with the concept of availability. A typical definition of availability is that of ANSI (1974):

"The portion of a selected time interval during which the information path is capable of performing its assigned data communications function. Availability is expressed as a percentage."

If a user attempts access at regular intervals during a time period of interest, and the system is "down" say $Q\%$ of the time during that period, it follows that the user will be denied access $Q\%$ of the time, even if no blocking occurs during nonoutage periods. If blocking occurs with probability P during the latter periods, the observed Access Denial Probability will be $(P + Q)$. Thus, blocking probability and availability both contribute to Access Denial Probability in circuit-switched systems.

The significance of Access Denial Probability to the user depends on whether alternative means of communicating his/her information are available. If no alternative system is available, the user has no choice but to continue attempting to access the denying system; and the negative consequences are similar to those cited earlier for longer Access Times: i.e., loss of productive time and data aging. In general, a series of Access Denials is more detrimental to the user than a single access delay of equivalent duration, because each Access Denial nullifies previously completed access steps (e.g., dialing). There is a definite buildup of dissatisfaction with repeated Access Denials in the case of human users.

If an alternative means of communication is available to the user, Access Denial Probability expresses the likelihood that it will be needed. Access Denial Probability is therefore useful in assessing the necessity for, and the potential utilization of, backup services.

Access Denial Probability values range between zero and one. A value of zero implies that the user is never denied access, i.e., the system is completely non-blocking and has perfect reliability. At the other extreme, a value of one implies a service that always denies the user access, i.e., never actually "serves."

In considering user requirements for Access Denial Probability, it is important to distinguish between what the user actually needs and what he will accept if nothing else is available. There are switched communication services with call blocking probabilities of 0.4 or even higher (e.g., AUTOVON); but there also is abundant evidence of user dissatisfaction with such services (GAO, 1977). Access Denial Probabilities in the range of 1% to 5% are normally satisfactory in applications where data "aging" is slow (e.g., computer program development). Values of 10^{-3} or lower may be needed in critical real-time applications (e.g., military command and control). An end-to-end circuit availability of 99% has been suggested for the evolving digital DCS (Kirk and Osterholz, 1976). An Access Denial Probability of 10^{-2} is specified in EPA (1980).

The system design features that most strongly influence Access Denial Probability are (1) the resource-sharing (or switching) technique used, and (2) the inherent reliability of the system facilities. Many smaller communication systems attempt no resource sharing, and they are therefore nonblocking; a familiar example is a simple "dedicated line" interconnecting two users. Availability values for typical dedicated services are in the neighborhood of 98%. Assuming the uniform access attempt rate, this corresponds to an Access Denial Probability of 2×10^{-2} . Better values can be achieved through the use of backup-circuit provisions (Frank and Hopewell, 1974).

Large multi-user networks almost invariably employ some type of resource sharing. The overall strategy of resource sharing is to take advantage of intermittent user demand by deliberately designing certain costly system elements (e.g., switches, trunks) with less capacity than would be needed to serve all users simultaneously. Under normal usage the design capacity is adequate, and the benefits of economical service are realized. Under unusually heavy usage the design capacity is not adequate, and the transmission of some offered traffic must be deferred.

The choice of resource-sharing technique has an obvious impact on Access Denial Probability. Circuit-oriented systems defer excess traffic by denying access, and they therefore exhibit relatively high Access Denial Probabilities. Typical values for blocking probability in well-designed circuit-switched systems are in the range of one to four percent (AT&T, 1961; Duffy and Mercer, 1978). System outages contribute relatively little further degradation in many cases, because a user pair can be interconnected by many different physical paths (Martin, 1976).

Message-oriented systems defer traffic by prolonging system storage rather than by denying access, and they therefore exhibit relatively low Access Denial Probabilities. In most message-oriented systems, Successful Access can occur even when there is no physical connectivity between the source and the intended destination: in such a situation, a switching node connected to the source simply holds the message until connectivity is restored.

In general, then, Access Denial occurs in message-oriented systems only when the switching center serving the source user is down. Values for the down rate (unavailability) of individual message switching nodes range between 10^{-4} and 10^{-2} , with the latter value perhaps being more realistic. A measured down rate of 1.64% has been reported for the ARPA network Interface Message Processors (Kleinrock and Naylor, 1974).

4.3 User Information Transfer Parameters

Once a user has successfully gained access to a telecommunication service, his performance concerns naturally shift to the user information transfer functions. Once again, these concerns can be grouped in three general categories:

Efficiency - What delay will my information experience in traversing the network? What throughput, or rate of information flow, will the system allow? How is this user-to-user throughput related to the total system capacity allocated to my traffic?

Accuracy - What is the likelihood that the system will alter or misdeliver my information? What is the likelihood that it will deliver duplicate messages, or other "extra" information not output by me, during a transaction?

Reliability - What is the likelihood that the system will lose my information in transmission?

In general, the end user is indifferent to how his information is transported as long as his end-to-end performance needs are met.

The following paragraphs describe the fourteen primary user information transfer performance parameters specified in the Interim Federal Standard. These parameters are Bit/Block Transfer Time, Bit/Block Error Probability, Bit/Block Misdelivery Probability, Bit/Block Loss Probability, Extra Bit/Block Probability, Bit/Block Transfer Rate, and Bit/Block Rate Efficiency. Corresponding bit- and block-oriented parameters are described together, but differences in definition and impact are highlighted where appropriate.

4.3.1 Bit/Block Transfer Time

Bit Transfer Time and Block Transfer Time describe the total delay an information unit experiences in transit between end users. For each information unit (bit or block), computation of transfer time begins when the information unit has been input to the system and its transmission has been authorized. Computation of transfer time ends when the information unit has been output to the destination user, with appropriate notification to that user where required. In each case, parameter values are calculated only on successful transfer attempts.

The difference between Bit Transfer Time and Block Transfer Time lies primarily in the output time, i.e., the time it takes to move bits across the destination user interface. A bit is an elementary information unit whose transfer across the destination user interface normally occurs at a single point in time. A block is an aggregate information unit, possibly containing many thousands of bits, and its transfer across the destination user interface often occurs in a series of increments occupying a substantial period of time.

Bit Transfer Time measures the total transfer delay for a single (typical) bit - in the interim standard, the average of the first and last bits in each block. Block Transfer Time measures the total time required to transfer all bits in a block. Thus, Bit Transfer Time is relatively independent of output rate, whereas Block Transfer Time is strongly influenced by output rate. The difference between the two provides a measure of output rate. This difference could be perceived by an operator, for example, as the difference between a terminal that outputs data serially, character by character, and one that outputs entire lines in parallel. The relative merits of the two output schemes are at least partly a matter of user preference.

Bit Transfer Time is closely akin to a class of parameters variously referred to as "transmission time" or "propagation delay" (McManamon et al., 1975). It differs from them, however, in that it includes time spent in system storage as well as time spent in actual transmission. Kleinrock (1976) defines the "initial response time" of a network as "the average time from when the first bit is presented to the network until the first bit is delivered." This parameter differs from Bit Transfer Time only in that the latter is an average over the first and last bits of each block.¹⁵

¹⁵For an excellent discussion of "response time" and its many definitions, see Miller (1968). For a recent application of block transfer "response time" in comparing the performance of two value-added networks, see Rose and O'Keefe (1980).

Block Transfer Time is very closely related to the ANSI standard parameter Message Transfer Time (ANSI, 1980):

"MTT is the time in seconds that is required for a message to be transferred from a source frame buffer and accepted at the designated sink frame buffer. Where more than one link is involved in the transfer, it includes all of the time required for enroute storage and forwarding."

ANSI defines the term "message" as follows:

"A message is an arbitrary amount of information whose beginning and end are defined. The information may be contained in one or more frames which must all be accepted (for the message to be accepted) in order to stop the MTT measurement."

A "message" thus defined includes the FED STD 1033 block, as well as many other possible information units.

Message Transfer Time differs from Block Transfer Time in two respects. The first is in the definition of measurement starting events. ANSI defines the start of an MTT measurement as follows:

"MTT measurements start when both of the following have occurred: (a) transmission service has been requested, and (b) the information field for the first frame has been entered in the source frame buffer. Transmission service requests may be evidenced by: the issuance of a call request; the transition to off-hook; an operator initiated action; or other equivalent indication."

Thus defined, MTT includes Access Time as well as the time spent in retries if access is denied. Block Transfer Time excludes these access phase delays.

The second major difference between MTT and Block Transfer Time is in the choice of measurement interfaces. ANSI defines the end of an MTT measurement as follows:

"MTT measurement is stopped upon acceptance of the final frame of the message at the destination frame buffer."

Frame acceptance occurs at Level 2 in the OSI protocol hierarchy depicted in Figure 3-5. The end of block transfer, as defined in FED STD 1033, occurs when the block crosses the interface between levels 6 and 7. The time difference between acceptance at Level 2 and delivery to the end user will often be small, but it can be substantial if the Level 3-6 programs are inefficient or have a low priority under the host computer operating system.

Block Transfer Time also bears some resemblance to the so-called "writer-to-reader-delay" used in analyzing military message communications (Feldman et al., 1979; Armed Service Investigating Subcommittee, 1971). The latter parameter

differs from Block Transfer Time in that its measurement interfaces are more inclusive. The total writer-to-reader-delay in communicating a message includes administrative review time and physical transport time as well as time spent in telecommunication. Such user delays should be carefully analyzed in defining telecommunication performance requirements. This subject will be discussed more fully in the Application Manual.

Bit Transfer Time and Block Transfer Time also closely resemble a general transfer time parameter defined in the CCITT Green Book (CCITT, 1973):

"Transfer Time - The time that elapses between the initial offering of a unit of the user's data to a network by a transmitting data terminal equipment and the complete delivery of that unit to a receiving data terminal equipment.... A unit of data may be a bit, byte, packet, message, etc."

The only significant difference here is the interface at which the starting and ending events take place.

In describing the access performance parameters, we identified two distinct disadvantages of communication delay: loss of productive time and data "aging." Both disadvantages also apply in the case of bit and block transfer, but with a slightly different emphasis. When a user is attempting to access a system, he typically must devote a substantial portion of his time to that effort for as long as it takes to succeed. In contrast, when a user has transferred a unit of information into the system for delivery to a destination, he is unoccupied, at least as far as that information unit is concerned; and he may well use the transfer time in other productive ways. Thus, the loss of productive time is often less significant in the case of transfer than in the case of access. This is particularly true of "electronic mail" systems, where the preparation, input, and output time for a message may be negligible compared to its transfer time (e.g., minutes vs. hours).

If lost time is less significant in transfer than in access, data aging is often more significant. When a user is denied access to a communication service, he at least knows that his message is not on its way, and he can try again or take some alternative action. In contrast, he may have no way of knowing when transfer is being delayed excessively; and the consequences of the delay may therefore be more severe. Many modern networks address this user concern by providing an explicit "delivery confirmation" response to the sender. The following quote from a recent Bell System advertisement for its Dataphone[®] Digital Service puts the consequences of communication delay in direct monetary terms:

"Delays can have very expensive consequences. On the order of \$1,000 per hour per circuit for time-sharing firms. Twenty-five times that for a company with a ship held in port by faulty documentation."

The possible values for Bit and Block Transfer Time range between zero and a practical upper bound defined by the "three times nominal" timeout defined earlier. A value of zero implies an infinite speed of transfer between source and destination (i.e., no system); extremely long values suggest that the system may function more like a sink than a pipe.

As in the case of Access Time, user requirements for Bit and Block Transfer Time vary over a wide range. At the lower extreme are real-time process control and teleprocessing applications, where average one-way transfer times much less than a second are specified (Martin, 1976; DCA, 1975; Kelley, 1977; EPA, 1980). The upper extreme probably occurs in the case of electronic message services, where "next-day delivery" is the key performance objective. In general, longer transfer times make the difference between Bit Transfer Time and Block Transfer Time less significant, and conversely.

The transfer time for a bit or block can be loosely divided into three components: modulation time, propagation time, and storage time. Modulation time is the minimum time a signal element must be maintained at the input to a circuit in order to ensure its detection at the output. It corresponds to the so-called "baud time" of a modem, and is inversely proportional to the signaling bandwidth. Modulation time may actually determine the minimum Bit Transfer Time on short, low-speed channels. As an example, a 20-mile cable circuit operating at 150 bits per second has a modulation time per bit over six times as long as the propagation time (Kimmitt and Seitz, 1978).

Propagation time is the total time a signal takes to traverse the physical distance between the two ends of a transmission circuit. The shortest propagation times are provided by terrestrial radiating systems such as microwave, which combine high propagation velocity (about 186,000 miles per second) with relatively direct signal paths. Cable systems also provide relatively direct signal paths, but their transmission velocities are much lower (e.g., 20,000 miles per second). Synchronous satellites exhibit much longer propagation times because of the longer path distances involved (e.g., 250 milliseconds for a 45,000-mile, single hop path).

Storage time includes all time during which a unit of user information is not physically moving through a communication system towards its destination. In all but the simplest systems, storage time is the dominant factor in determining

both Bit and Block Transfer Time. There are two principal motives for temporarily storing user information within a system during its transfer between end users:

1. Data Aggregation. Systems normally collect many serially transmitted bits in a single block or frame at each end of a transmission link to facilitate error detection and other control functions. Data may also be aggregated in a destination terminal in order to deliver it to the user in meaningful "chunks" (e.g., ASCII characters or computer words).
2. Resource Sharing. Message-oriented systems store user information at various internal switching nodes to increase utilization of the associated transmission links. Systems may also store user information to facilitate the sharing of user resources: "mail box" and "call hold" features provide examples of such storage (e.g., see AT&T, 1978).

Simple circuit-switched services with unbuffered terminals have bit and block transfer times among the lowest available - e.g., 30 to 100 milliseconds for typical transmission path lengths.¹⁶ Circuit-switched services with buffered terminals have somewhat longer transfer times because the blocks are longer (e.g., 80 characters); typical values for such services are in the range of 100 to 300 milliseconds. The ARPANET, a prototype packet-switching network with a virtual-circuit protocol, was designed to provide end-to-end delays less than one-half second for typical messages of a few thousand bits (Roberts and Wessler, 1970). Measured results indicate that actual transfer times in the ARPANET are in fact lower (Kleinrock, 1976).

Transfer times are substantially higher when traditional message switching is employed, because the messages are stored in their entirety at each switching node. The end-to-end message transfer times for DCA's AUTODIN I are probably typical (Armed Services Investigating Subcommittee, 1971):

<u>Message Precedence</u>	<u>Transfer Time</u>
Flash	<10 minutes
Immediate	<30 minutes
Priority	<3 hours
Routine	<6 hours

As noted earlier, the upper extreme on transfer time occurs in electronic mail systems. Particularly in the case where the destination user must take some action to "read his mail," delays on the order of a day or more are not unusual. The

¹⁶The bit and block transfer times are identical in many such services because all bits in a block (character) cross the user/system interfaces in parallel.

ancillary performance parameters provide a method of "factoring out" the user component of such delays, as discussed in Section 4.6.

4.3.2 Bit/Block Error Probability

Bit Error Probability and Block Error Probability express the likelihood that a unit of information transferred from a source to the intended destination will be delivered with incorrect binary content. The numerator of each probability ratio is the number of information units (bits or blocks) delivered to the intended destination with content errors; and the denominator is the total number of information units transferred (by intent) between the source and destination in question. In each case, the denominators exclude lost, misdelivered, and extra information.

In the case of Bit Error Probability, "incorrect content" normally means simple binary inversion between source and destination, i.e., a transmitted one becomes a received zero or vice versa. ANSI Task Group X3S35 has considered a number of more complex cases, including code conversion and the representation of user information in nonbinary symbols, and has suggested the following additional guidelines:

1. In the case of code conversion, error comparisons should be based on the intended and actual bit patterns at the destination user interface.
2. In the case where information crosses the user/system interfaces in the form of nonbinary symbols (e.g., ASCII characters), the input or output symbols should be translated into bits on the basis of the binary representation physically closest to the user.

"Incorrect content" in a delivered block is defined to exist whenever (a) one or more bits in the block are incorrect, or (b) some, but not all bits in the block are lost or extra. Thus, any unintended change in the information content of a delivered block identifies the block as incorrect. In general, the Block Error Probability for an n-bit block will be between one and n times the sum of the Bit Error, Bit Loss, and Extra Bit Probabilities, depending on how many failure outcomes occur in each incorrect block.

Bit and Block Error Probability are perhaps the most widely used communication performance parameters, and there is little need to relate them to others for the purpose of familiarization. Nevertheless, it is important to emphasize that both parameters apply, in the first instance, to end-to-end services as defined earlier; and their values should reflect the error-producing or error-removing effects of data terminals and higher level protocols.

Bit Error Probability is similar to the ANSI parameter "residual error rate" (ANSI, 1980) in that both measure errors which remain after error control. However,

the latter parameter includes undelivered and "misaccepted" (misdelivered or extra) bits in the numerator, and uses the total number of transmitted (source) bits as the denominator.

The significance of Bit and Block Error Probability to end users is also relatively apparent, but a brief discussion may be helpful. Two general categories of error effects can be distinguished, depending on whether the end user does or does not detect the error prior to using the delivered information. User detection of delivered errors is most probable in the case where the user is a human terminal operator. If the error is isolated and occurs in redundant text (e.g., misprinting of a single character in a text message), the operator can normally infer the intended meaning; and the error may be no more than a minor nuisance. If the error is more extensive or occurs on nonredundant text (e.g., total garbling of a line or an error in numerical data), the impact on the user will be much more significant. Typically, the destination user must re-contact the source, request a retransmission, and then defer any action based on that information until the retransmission is received. In essence, the users are performing the function of error control in a costly and inefficient manner.

The effects of delivered errors are more serious, in general, when they are not detected at the destination prior to actual use of the delivered information. This will almost always be the case when no human operator is involved. The many possible effects of undetected errors can be summarized by saying that they cause the destination user to make decisions and take actions based on erroneous information. In critical applications such as strategic weapons control and electronic funds transfer, such mistakes can be very costly.

Bit and Block Error Probability values vary between 0 and 1, with a practical upper limit of 0.5 on the former. In each case, a value of 0 implies that incorrect information is never delivered to the end users; i.e., total reliance on system outputs is warranted. A Bit Error Probability of 0.5 means that any delivered bit is just as likely to be wrong as right; and therefore no useful information can be communicated.¹⁷ A Block Error Probability value of one indicates that every delivered block contains at least one incorrect, lost, or extra bit.

User requirements for Bit and Block Error Probability depend, as one would expect, on the consequences of errors. Narrative message applications are among the least stringent (e.g., 10^{-2}) because their high inherent redundancy makes

¹⁷It is interesting (though academic) to note that useful information can be gleaned from a bit stream when the Bit Error Probability is greater than 0.5, by inverting each bit.

user correction possible. It has been estimated, for example, that normal English text is 50% redundant compared to a random character sequence (Shannon, 1948). Very high Bit Error Probability values may be tolerated at the output of digital subsystems used in transmitting voice; it has been shown, for example, that Continuous Variable Slope Delta systems can produce "acceptable" speech with channel Bit Error Probabilities approaching 10^{-1} (McRae et al., 1976).

As suggested earlier, user requirements for Bit and Block Error Probability are most stringent in applications where the cost of errors is high. A Bit Error Probability requirement of 10^{-12} has been specified for AUTODIN II users having error controlled access circuits (DCA, 1975); a more recent study have suggested a less stringent (and more realistic) value of 10^{-10} (Nesenbergs et al., 1980). Bit Error Probability requirements for normal teleprocessing applications are in the range of 10^{-5} to 10^{-8} ; a value of 8×10^{-6} is specified in the EPA (1980).

Some feeling for the significance of these numbers can be obtained by relating them to output rate. A 10^{-5} Bit Error Probability corresponds to approximately one bit error every 17 minutes at 100 bits per second; every two minutes at 1 kilobit per second; or every 10 seconds at 10 kilobits per second. A 10^{-10} Bit Error Probability corresponds to one bit error every 32 years, every 3 years, or every 4 months, respectively, at the same output rates.

In describing the Bit and Block Error Probabilities of existing systems, it is important to distinguish between values observed at the transmission channel interfaces and at the end user interfaces. So-called "raw channel" Bit Error Probabilities vary from 10^{-3} (for HF digital systems) to 10^{-6} (for all-digital, nonradiating local area networks). A value of 10^{-5} is probably typical for the public switched network (AT&T, 1971). For any given transmission speed, the raw channel error probability is primarily determined by two factors: (1) the signal-to-noise ratio at the receiver input, and (2) the effective transmission bandwidth. These can be effectively traded off in many cases (e.g., see Utlaut, 1978).

The raw channel error performance of a data communication system can be vastly improved through the use of error control techniques (Hamming, 1950; Kuhn, 1963). The most commonly used technique today is simple error detection and retransmission, also called ARQ. Well-designed ARQ systems can produce output channel Bit Error Probabilities in the range of 10^{-8} to 10^{-10} with negligible coding redundancy, almost irrespective of the raw channel error probability. Unfortunately, such systems also severely restrict throughput as the channel error probability approaches the reciprocal of the block size. This disadvantage can

be mitigated, in many cases, by hybrid ARQ/Forward Error Correction systems (Nesenbergs, 1975).

4.3.3 Bit/Block Misdelivery Probability

Bit and Block Misdelivery Probability express the likelihood that a unit of information delivered to a given destination user will, in fact, have been intended for some other user. The numerator of each probability ratio is the total number of misdelivered information units (bits or blocks); and the denominator is the total number of information units transferred between the source and destination in question. In essence, these parameters answer the following question: "of all the bits (or blocks) actually transferred between source A and destination B, how many were intended for some destination other than B?" Expressing misdelivery probability on a bit basis is not intended to imply that individual bits will be misdelivered; such outcomes will normally occur in groups of one or more blocks.

How can misdelivery occur? One obvious cause in circuit-oriented systems is Incorrect Access: i.e., a source user is connected to the wrong destination during the access phase. Misdelivery can also occur in message-oriented systems, as a result of routing errors. Misrouting of a message can either be a random event caused (for example) by an undetected error in a message address field, or a systematic occurrence caused (for example) by an incorrect address table in a message switching center. Errors of the latter type may be a result of software "bugs", hardware failures, operator errors, or even deliberate tampering.

The significance of misdelivery to the source user has been discussed earlier in connection with Incorrect Access Probability. Briefly, the two chief risks are (1) exploitation of the misdelivered information by an unfriendly recipient, and (2) inappropriate actions based on the false assumption that the information has been successfully delivered.

Most readers will recall the incident in which a U.S. Navy ship, the U.S.S. Liberty, was attacked and severely damaged by Israeli forces during the 1967 Arab-Israeli war. Many may not be aware, however, of the crucial role the misrouting of messages played in causing that incident. Here is a brief summary of that role, abstracted from a report of the Armed Services Investigating Subcommittee (1971) to the U.S. House of Representatives:

"Hostilities commenced between Israel and the United Arab Republic on June 5, 1967 ... During the afternoon of June 7th, the Joint Chiefs of Staff decided to reposition U.S.S. Liberty to move her farther from the coasts of the belligerent nations. In implementing that decision,

a series of five messages from JCS and U.S. commanders in the European Command were directed to U.S.S. Liberty and other addresses. None of those messages had reached Liberty by 1200Z hours on June 8th, 13-1/2 hours after the first message was released for transmission."

No less than three of these messages were misrouted, one on two separate occasions. The latter misroutings are summarized in the Subcommittee report as follows:

"The information copies of the message, addressed to U.S.S. Liberty and Commander, Task Force 64, were finally transmitted at 0350Z, but once again, those messages for addresses in the Mediterranean area were misrouted to Naval Communications Station, Philippines. A subcommittee witness testified that the misrouting was due to an erroneous routing indicator which had been assigned to the message by a civilian clerk in the Army Communications Center, Pentagon. Upon its arrival at the Naval Communications Station, Philippines, the error was recognized, the routing indicator was corrected to Naval Communications Station, Morocco, and the message was retransmitted within an hour. That correction should have taken those copies of the message to the Mediterranean area and ultimately to the addressees, except that the message was routed to pass through the Army Communications Station, Pentagon. That station, instead of transmitting the messages to the Navy Communications Station, Morocco, to which they were addressed, sent them to National Security Agency, Fort Meade, MD., where they were filed without further action. The only explanation given for this inexcusable conduct was that clerical personnel had misread the routing indicator. Needless to say, those messages had not reached either U.S.S. Liberty or Commander, Task Force 64, by 1200Z hours, June 8, 1967."

The Subcommittee summarizes its report on the Liberty communications snafu as follows:

"The circumstances surrounding the transmission of those messages could be considered a comedy of errors were it not for the tragic results of the failure to move U.S.S. Liberty. At 1210Z hours, June 8, 1967, U.S.S. Liberty was attacked by Israeli aircraft and, at 1235Z hours, she was torpedoed by Israeli patrol boats. As a result of those attacks, 34 officers and men were killed, while 75 were wounded, and the ship sustained such severe damages that it was never restored to duty. At the time of those attacks, U.S.S. Liberty, through no fault of hers, had not received any of the above-described messages. If the communications system had been responsive, she should have had several hours during which she could have placed some distance between herself and the coast, thereby probably avoiding the attack."

In each of the above cases, the misrouting was a result of human error. Unfortunately, machines make errors too. Kleinrock (1976) discusses two such errors which occurred in the early days of the ARPANET:

"Although transmission line errors are handled by cyclic error detecting codes that are hardware generated and checked by the line modems, no protection against hardware errors in the IMP [Interface

Message Processor] was provided originally. This led to some amusing network crashes. For example, there was the case of an IMP that improperly generated routing update messages claiming it had zero-delay paths to all destinations in the net! This IMP became an absorbing node for an unlimited amount of traffic, finally bringing the network to its knees. In another case a certain amount of chaos was caused when an IMP claimed it was the UCLA IMP (which it was not!). These hardware errors are now detected by the inclusion of a (16-bit) software checksum that accompanies all packets in their journey through the net."

In sum, a healthy skepticism is justified in assessing claims that data misdelivery "cannot occur."

Possible Bit and Block Misdelivery Probability values range between zero and one. A value of zero implies that misdelivery is impossible; or, in a measurement context, that all traffic transferred between a specified source and destination user was correctly delivered. A measured value of one suggests an addressing error in which all transmitted traffic is systematically misrouted.

Moving on to the subject of user requirements, the AUTODIN II specification (DCA, 1975) calls for a "segment" misdelivery probability of 10^{-11} . This number applies directly to both Bit and Block Misdelivery Probability, since misdelivery outcomes normally occur in block or "segment" groups. More recently, Nesenbergs et al. (1980) suggest the same target value. For reference, such a value is sufficient to enable a user pair to exchange ten million packets per day for 27 years before the first misdelivery occurs.

Bit and Block Misdelivery Probabilities like those specified above are impossible to measure, and they are thus, in one sense, academic. Nevertheless, such values can have a substantial influence on system design, and this often justifies their inclusion in specifications. Misdelivery probability can be reduced to negligible proportions by at least one "brute force" approach: that of providing a protected, dedicated, hard-wired line between the source and destination in question. Although expensive, such an approach may be justified in situations where the consequences of misdelivery are especially grave.

The single design feature that most strongly influences Bit and Block Misdelivery Probability in switched systems is the error control technique. Depending on the number of compatible terminals in a network, systems without error control on the addressing information may experience misdelivery probabilities in excess of 10^{-5} (Kimmett and Seitz, 1978). Error control provisions such as those employed in common channel signaling systems and in the ARPA network will reduce these probabilities substantially, perhaps to the neighborhood of 10^{-9} in a benign

environment. CCITT Study Group VII suggests an "illustrative figure" for datagram misdelivery probability of 10^{-6} (CCITT, 1978). As one would expect, these values are greatly increased by the presence of deliberate tampering threats. In secure communication systems, encryption techniques are routinely employed to foil such efforts (Feistel, 1973; Popek, 1974).

The cause-effect relationship between Incorrect Access and Bit or Block Misdelivery and has been discussed earlier. As a general rule of thumb, it can be said that the values for the two parameter types will seldom differ by more than an order of magnitude in circuit-oriented systems; the latter values may be somewhat lower as a result of user detection of Incorrect Access events prior to the completion of message transfer.

4.3.4 Bit/Block Loss Probability

Bit Loss Probability and Block Loss Probability express the likelihood that a system will fail to deliver a unit of information output by a source to the intended destination within a specified maximum transfer time. The numerator of each probability ratio is the total number of information units (bits or blocks) lost as a result of system performance failures; and the denominator is the total number of information units output by the source, excluding any not delivered as a result of user nonperformance. The timeout period on both bit and block transfer attempts is three times the nominal (specified) value of the parameter Block Transfer Time. Block Loss is distinguished from Block Refusal (i.e., nondelivery for which the user is responsible) by comparison of ancillary parameter values, as discussed in Section 4.6.

How can a system "lose" a user's information? There are at least seven distinct ways. The first is through signal interruption. In systems that do not provide block storage and retransmission, a fade or other attenuation may interrupt the signal representing a sequence of bits, with the result that they are simply "wiped out" of the delivered data stream. This is a very real phenomenon in asynchronous systems; in fact, the character loss rate exceeds the character error rate by more than an order of magnitude on short, low-speed data links in the public switched network (AT&T, 1971).

The second possible cause of data loss is timing errors. Whenever two communicating elements in a digital system are driven by different clocks, there is the possibility that one or more bits will not be sampled by the receiver while they are being presented by the sender. Unless the "slip" is later corrected by error detection and retransmission, the unsampled bits will be lost by the

system. The Satellite Business Systems digital interface specification includes an explicit description of this phenomenon (SBS, 1978):

"For those cases in which the port, configured as DTE, must accept transmit and receive timing from the external equipment, a performance degradation known as slipping will occur. Slipping is the loss or duplication of data during a transmission due to timing differences between two independent synchronous transmission systems operated in tandem.... If the external system is fast relative to the SBS system, 64 bits will be lost each time a slip occurs."

A third cause of data loss is ARQ protocol failures. Simple ARQ protocols control retransmission by means of positive acknowledgment (ACK) and negative acknowledgment (NAK) signals which are returned from receiver to sender after each block transmission. Block loss will occur whenever a NAK is changed (by transmission errors) to an ACK, since the sender will assume the block was delivered when in fact it was discarded. The probability of such transformations can be made extremely small by appropriate error coding, but not all systems employ such methods.

A fourth cause of data loss is hardware or software "crashes" in system switching computers. A simple illustration of a hardware crash is the case where the power supplying a semiconductor memory in a message switching computer is suddenly interrupted. The system will then lose all messages stored in that memory unless special backup provisions have been made. Software crashes are often even more serious, since the affected switch may continue to operate, in an essentially "crazy" fashion, for some time before the failure is detected. The IMP crashes described earlier by Kleinrock indicate some of the possibilities here. Sunshine (1975) proves that it is impossible to prevent data loss or duplication in all cases when one side of a protocol fails with memory loss.

A fifth cause of data loss is a flow control protocol failure known as a "lockup". Kleinrock (1976) describes one of many lockup conditions actually observed in the ARPANET as follows:

"Reassembly lockup, the most famous of the ARPANET deadlock conditions, was due to a logical flaw in the original flow control procedure. It occurred when partially reassembled messages could not be completely reassembled since the congested network prevented their remaining packets from reaching the destination; that is, each of the destination's neighbors had given all of their store-and-forward buffers to additional messages heading for that same destination (at which no unassigned reassembly buffers were available). Thus the urgently needed remaining packets could not pass through the barrier of blocked IMPs surrounding the destination."

While noting that this and several other lockup conditions have been eliminated by subsequent changes in flow control procedures, Kleinrock points out that "indeed, whenever one introduces conditions on the message flow, there exists the danger that these conditions cannot be met and then the message flow will cease. Reassembly and sequencing are examples of such conditions."

A sixth possible cause of data loss is the deliberate discarding of packets to eliminate network congestion. This strategy is the rule rather than the exception in modern packet switching networks, particularly those that employ "datagram" protocols. CCITT (1978) suggests an "illustrative figure" for the probability of such a discard (with notification to the sender) of 10^{-3} ; the corresponding value for discard without notification is 10^{-4} . Sloan (1979) proposes an explicit procedure for limiting the maximum lifetime of packets in a packet-switching network. Kleinrock et al. (1976) present measured results indicating that "on the average, every hundredth message that enters the ARPANET will not reach its destination. The reason for this undesirable behavior is that many destination hosts are tardy in accepting messages." Remember that the host front end is a part of the system as far as the end user is concerned.

A final cause of data loss in communication systems is internal misrouting. Data which is misrouted in a communication system may or may not be delivered to an incorrect destination; but in either case, it is lost as far as the source and intended destination are concerned. Various causes of such errors have been discussed in the preceding section.

The impact of data loss on users has also been discussed earlier, inasmuch as "loss" and "excessive delay" have similar impacts. The loss of a few early bits in a block may cause the user to misinterpret the meaning of succeeding bits in some applications; the effect on the user may then be the same as if the system's delivered Bit Error Probability had suddenly jumped to 0.5.

Bit and Block Loss Probability values range between 0 and 1, with a measured value of 0 meaning no loss and a measured value of 1 suggesting a system with an open circuit or infinite sink. Perceived requirements for user applications range from values as high as 10^{-3} (in normal, redundant message text) to values as low as 10^{-11} (in highly critical military applications). In character-asynchronous systems, the former value corresponds to about two errors on a printed page of text. Teleprocessing user requirements are typically intermediate between these extremes, in the range of 10^{-5} to 10^{-8} . A Block (character) Loss Probability of 8×10^{-6} is specified in EPA (1980).

As in the case of the error probabilities, the key design impact on the loss probabilities is the choice of error control technique. Systems that do not provide data storage and retransmission are always vulnerable to signal interruptions and slips, and there are specialized applications (e.g., space-to-earth communications, military missions requiring radio silence) in which retransmission protocols are not possible. Character loss probabilities for unbuffered, asynchronous services in the public switched network are in the range of 10^{-3} to 10^{-4} (AT&T, 1971). SBS (1978) relates the frequency of slips with bit timing accuracy in the following table:

<u>Accuracy of External Supplied Timing</u>	<u>Number of Bit Times Between Slips (Loss or Duplication of Data)</u>
± 1 in 10^9	1.0×10^{10}
± 1 in 10^8	4.3×10^9
± 1 in 10^7	6.1×10^8
± 1 in 10^6	6.4×10^7
± 1 in 10^5	6.4×10^6
± 1 in 10^4	6.4×10^5

If we assume that 64 bits are lost on each slip, the above data translate into Bit Loss Probability values between about 6×10^{-8} and 4×10^{-3} .

In most modern systems with well-designed retransmission protocols, the predominant cause of data loss will be switch "crashes" and network congestion. Given Sunshine's demonstration that retransmission protocols cannot be made loss-proof in the face of node crashes, and Kleinrock's measurements of a 1% to 2% down rate for the ARPA network IMPs, it seems highly questionable that user Bit Loss Probability requirements like 10^{-11} are attainable.

4.3.5 Extra Bit/Block Probability

Extra Bit Probability and Extra Block Probability express the likelihood that the information delivered to a destination user will contain duplicate bits or blocks, or other "extra" information not output by the source. The numerator of each probability ratio is the total number of extra information units (bits or blocks) received by a particular destination user; and the denominator is the total number of information units received by that user. Unless Misdelayed Bits are explicitly identified in a measurement process, they will be counted as Extra Bits (Seitz and McManamon, 1978).

How can a system include "extra" information in a sequence of bits delivered to a destination user? The most frequent cause is the inadvertent duplication of previously delivered data. Three of the seven phenomena just discussed as causes of data loss can also cause data duplication: timing errors, ARQ protocol failures, and hardware or software "crashes." Timing errors between system elements cause duplication, rather than loss, whenever the clock in the sending element is slower than that in the receiving element. In such a situation, input data will be sampled twice at periodic intervals; and if the error is not corrected later by error detection and retransmission, the duplicate data will be delivered to the destination user.

ARQ protocol failures cause data duplication in essentially the complement of the way they cause loss. Any time an ACK is changed to a NAK as a result of transmission errors, the data sender will unnecessarily retransmit the block in question; and two copies of the block will then exist at the receiver. Both copies may be delivered to the destination user if the protocol in use does not assign unique ID's to each packet. The same thing may happen when an ACK is lost in so-called positive acknowledgment, retransmission on timeout (PAR) protocols.

Probably the dominant cause of data duplication in modern communication systems is hardware or software "crashes". When a switch crashes, its memory about current status of information in transit may well be lost. Most switches are programmed to retransmit dubious blocks in such a circumstance, and some of these may already have been delivered.

Some communication networks deliberately transmit duplicate copies of user information to improve transfer reliability or speed, and then eliminate the duplicates by "filtering" at the destination. One modern network which does so is the National Weather Service's Automation of Field Operations and Services (AFOS) network. This approach is also used, more or less informally, in many military message switching systems. Referring to one of the five errant messages to the U.S.S Liberty, the Armed Services Investigating Subcommittee (1971) makes a rather caustic reference to one such case:

"In order to ensure getting this message to its addressees, it was transmitted concurrently over two alternative relay paths. The necessity for the alternate transmission was quickly demonstrated by the loss of the message at the Pirmasens, Germany, Army DCS relay, the first station on one of the transmission paths. As a result of that loss, there was no further transmission of that copy of the message. The explanation offered for the loss of that message was that 'the station was being operated under a combination of adverse conditions caused by the consolidation of commands and relocation of units from France.

Heavy traffic volumes resulted from the extensive relocation of units and retermination of teletype circuits. The number of qualified personnel was inadequate to ensure error-free processing of traffic."

A few military message switching networks actually provide users with an explicit SUSDUP code to identify suspected duplicates. Duplicate messages may also be created by redundancy (overlapping) in addressee tables in some networks.

The impact of data duplication on the user is substantially different from that of data loss. "Extra" information has no impact whatsoever on the source user, since his entire output is, in general, delivered as intended. The impact of extra information on the destination user depends on the type of user and on how clearly the duplicated data is delimited from other, nonduplicate information. A clearly delimited, complete duplicate of a previously delivered message is normally no more than a minor nuisance to a human end user: he will simply throw the duplicate out. At the other extreme, duplication of even a few bits of numerical data may cause a computer application program to completely misinterpret an input file, thereby producing a meaningless or misleading output.

Extra Bit and Extra Block Probability values theoretically range between zero and one, but 0.5 is probably a more realistic upper bound. A value of 0.5 suggests that every block output by the source is delivered to the destination twice.

Data on user requirements for Extra Bit and Extra Block Probability is exceedingly scarce. Nesenbergs et al. (1980) suggest a value of 10^{-10} to 10^{-11} for interactive data communication services in the future DCS, based on the premise that Extra Bits have essentially the same effect as Incorrect Bits. EPA (1980) specifies an Extra Character Probability of 8×10^{-6} for the teleprocessing application cited earlier, apparently based on the same premise.

The key design impact on Extra Bit and Extra Block Probability is, once again, the choice of error control technique. Character-asynchronous circuit-switched systems with no retransmission or buffering will provide the lowest possible values (essentially zero), since such systems contain no storage in which duplicate information could be created. Traditional message-switching systems probably exhibit the highest values (e.g., as high as 10^{-3}) because of their long-term storage of entire user messages in each switching node. Kimmett and Seitz (1978) estimate a value of about 10^{-6} for a star-connected message-switching network with modern outage recovery features.

4.3.6 Bit/Block Transfer Rate

Bit Transfer Rate and Block Transfer Rate describe the average rate at which user information is successfully transferred between a given source and

destination user. The numerator in each case is the total number of user information units (bits or blocks) successfully transferred during a defined sampling interval (corresponding to a "message"); and the denominator is the duration of that interval in User Information Transfer (UIT) time.

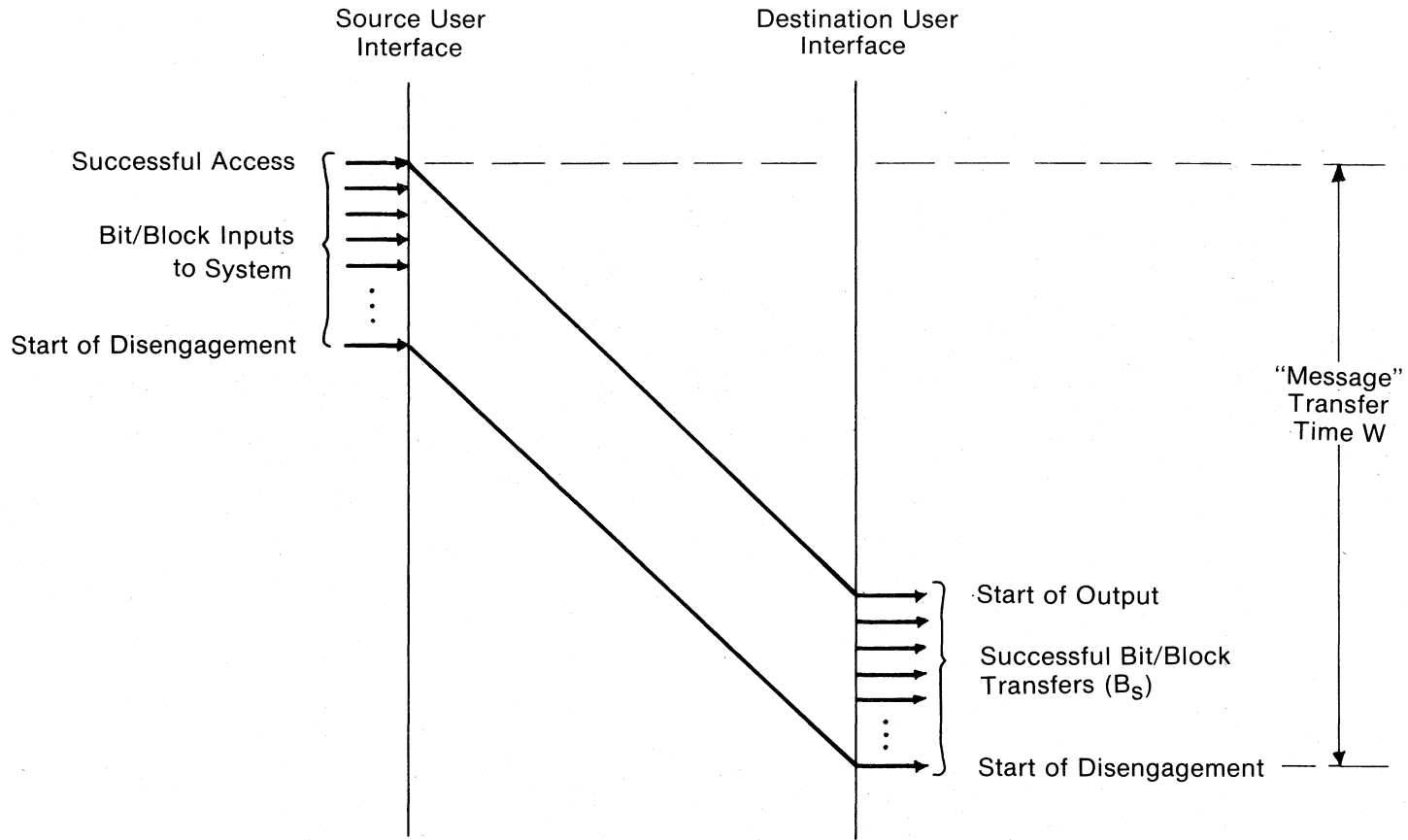
Only successful transfer outcomes are counted in calculating values for Bit and Block Transfer Rate. Each sample (or "message") encompasses a fixed number of user information bits, defined on the basis of required measurement precision as specified in Crow and Miles (1976). The UIT time for a user pair includes all time, within transactions involving that pair, between Successful Access and the start of disengagement of the last committed user. It thus includes all time during which any user information is in transit between that pair.

The sampling interval used in calculating Bit and Block Transfer Rate will often be contained within a single information transfer transaction, but it may be formed, where necessary, by concatenating the user information transfer phases of successive transactions. All idle, access, and disengagement time between successive transfer phases is excluded from the UIT time when this is done (Seitz and McManamon, 1978).

Figure 4-2 defines the Bit and Block Transfer Rate parameters in pictorial form, assuming (for simplicity) that the rate sampling interval corresponds to the transfer phase of a single transaction. In that case the "message" consists of all user information bits output by the source during the transaction; the rate numerators consist of all successfully transferred bits or blocks (B_s); and the rate denominators consist of the total time (W) between Successful Access and the start of disengagement at the destination. Note that start of output at the destination may or may not precede the start of disengagement at the source; i.e., the input and output of user information may or may not overlap in time.

Bit and Block Transfer Rate are both measures of useful information flow or "throughput." A better understanding of how these parameters describe flow can be obtained by considering some of the variabilities inherent in their measurement. In the context of Figure 4-2, consider the following:

1. The time delay between input of an information unit at the source and output of the corresponding unit at the destination may be milliseconds or hours, depending on the type of system.
2. Information input and/or output may be bursty in nature, with instantaneous rates far higher than would be sustainable on a continuous basis.



$$\left\{ \begin{array}{l} \text{Bit (or Block)} \\ \text{Transfer Rate} \end{array} \right\} = \frac{\left\{ \begin{array}{l} \text{Total Number of Successfully} \\ \text{Transferred Bits (or Blocks)} \end{array} \right\}}{\left\{ \begin{array}{l} \text{Total "Message" Transfer Time} \end{array} \right\}} = \left[\frac{B_s}{W} \right]$$

Figure 4-2. Bit and block transfer rate definitions.

3. One system may have a capability to store thousands or even millions of user information bits in transit between a source and destination, while another may have extremely limited storage capacity.

In essence, the FED STD 1033 rate parameters distribute the Successful Bit Transfer and Successful Block Transfer outcomes observed during a performance period uniformly over the entire period, beginning with input of the first bit at the source and ending with output of the last bit at the destination. These parameters do not attempt to "factor out" periods of source or destination inactivity during user information transfer; they do not attempt to describe variability in flow; and they do not attempt to "scale up" a system's throughput to reflect "pipelining" or other system storage capabilities. In sum, they measure average actual, rather than instantaneous or potential, user information transfer.

A brief survey of previously defined flow measures will further clarify the meaning of the FED STD 1033 rate parameters. ANSI (1974, 1980) has defined four standard transfer rate parameters: Transfer Rate of Information Bits (TRIB); Link Transfer Rate of Information Bits (L-TRIB); Network Transfer Rate of Information Bits (N-TRIB); and User Message Data Rate (UMDR). Key excerpts from these parameter definitions are provided below.

- Transfer Rate of Information Bits. The TRIB criterion expresses the ratio of the number of Information Bits accepted by the receiving Terminal Configuration during a single Information Transfer Phase (Phase 3) to the duration of that Information Transfer Phase. TRIB is expressed in bits per second. Information Bits are all bits excluding start-stop elements (if used) and parity bits contained in Information Characters. Information Bits are defined to be accepted by the receiving Terminal Configuration if a positive acknowledgment to a transmission block is received by the sending Terminal Configuration.
- Link Transfer Rate of Information Bits. L-TRIB is the number of information bits transferred and accepted in a data communication link during a specified time interval divided by that time interval; it is expressed in bits per second. In a Primary to Secondary, point-to-point or multipoint configuration, the number of information bits used in determining L-TRIB is the sum of the information bits transmitted and received by the Primary. In a balanced configuration, it is the sum of the information bits transmitted and received by either station. Information bits in frames that are not accepted are excluded.
- Network Transfer Rate of Information Bits. N-TRIB is a measure of the total information flow rate in a network. It is determined by dividing the sum of all accepted information bits leaving all exit ports of the network by a continuous time interval measurement.
- User Message Data Rate. User Message Data Rate, a message-based measure of performance, is determined by dividing the number of user-defined data bits in a message by the Message Transfer Time.

Of these four ANSI standard parameters, TRIB is the most similar to the FED STD 1033 transfer rates. TRIB differs from Bit Transfer Rate primarily in that it defines "Information Bits" and the "Information Transfer Phase" on the basis of a particular, character-oriented communication control protocol, ANSI's X3.28 (ANSI, 1971). Note that the denominator of TRIB is always the duration of a single Information Transfer Phase, irrespective of how many bits are transferred during that period. TRIB is, by definition, a unidirectional measure of flow.

The parameter L-TRIB differs from Bit Transfer Rate in two fundamental ways:

1. It is measured at the frame buffer outputs (i.e., at Level 2 in the OSI protocol hierarchy of Figure 3-5) rather than at the end user interfaces.
2. It is a bidirectional flow measure; i.e., both transmitted and received bits are counted in its numerator.

Bidirectional flow values may be determined using the FED STD 1033 parameters by averaging the separate Bit and Block Transfer Rate values characterizing each direction of flow.

The parameter N-TRIB differs from Bit Transfer Rate (as well as the other ANSI parameters) in that it measures the total network throughput for all users rather than the throughput for a particular user pair or group. Like L-TRIB, N-TRIB is measured at the frame buffer outputs (Level 2) rather than at the end user interfaces.

The parameter UMDR differs from Bit Transfer Rate in two respects: (1) it is measured, once again, at the frame buffer interfaces; and (2) it includes Access Time (and any time required to respond to Access Denials) in the denominator. A description of the ANSI standard parameter Message Transfer Time, which forms the denominator of UMDR, has been provided in Section 4.3.1.

A somewhat more technical definition of throughput is offered by Kleinrock (1976):

"The network throughput... we define as γ_{jk} msg/sec between source j and destination k ... more than one message (say up to m) is allowed to be in transit through the network at the same time.... If Z_{jk} is the network delay average over all messages passing from j to k ... then at most we have

$$\gamma_{jk} = \frac{m}{Z_{jk}}$$

If \bar{b} is the average message length (in bits), then the average throughput in bits per second is simply $\bar{b}\gamma_{jk}$ (at most)."

The latter quantity, $\bar{b}_{y_{jk}}$, is essentially an upper bound on Bit Transfer Rate. It represents the Bit Transfer Rate that would be measured if the "pipe" were always "full", i.e., if the Bit Transfer Time were negligible compared to the total rate measurement period.

Bit and Block Transfer Rate are significant to the user primarily as measures of the overall "responsiveness" and "capacity" of an information transfer service. To clarify this a bit, consider your expectations when you turn on the garden hose at your home. Once the faucet is on, you expect water to begin flowing out the far end of the hose in abundance within a very short time. If the flow of water is delayed excessively after you turn on the faucet, or if water dribbles rather than gushes out once flow does start, you are likely to be dissatisfied with your hose or your water service. Bit and Block Transfer Rate provide a quantitative measure of these concerns as they apply to the flow of user information in data communication systems.

Values for Bit and Block Transfer Rate range between zero and a practical upper limit determined by the Signaling Rate of the service (see Section 4.3.7). Low measured values imply little useful flow, i.e., either (1) the input or output is negligible, (2) the transfer delay is excessive, or (3) the information that is delivered is incorrect. Conversely, high values imply high input and output rates, low transfer delays, and good delivered accuracy.

In specifying user requirements for Bit and Block Transfer Rate, it is well to keep in mind that the user himself is often the major cause of flow and delay restrictions in data communication systems. It makes little sense, for example, to specify a data channel with a Bit Transfer Rate of 2000 bits per second if the input rate is limited by the source user's typing ability to less than one-fiftieth that rate (e.g., 50 words per minute). Similar constraints exist, in many cases, on the output side. The key point here is that advances in resource-sharing technology are making it increasingly feasible and economical for users to specify transfer rate requirements on the basis of their actual capability to generate and absorb traffic. These capabilities are often far lower than the data transfer capabilities of traditional dedicated and circuit-switched services.

The FED STD 1033 transfer rate parameters encourage usage-sensitive rate specifications by including user delays (e.g., think time, typing delays) in their denominators. As noted earlier, the equivalent user-independent values can always be determined using the ancillary parameters (Section 4.6). In general, user requirements for Bit and Block Transfer Rate should be derived in the context of a "data stream model" like that described in Jackson and Stubbs (1969). Such an approach will be outlined in the Application Manual.

Within the above framework, what are some typical values for Bit and Block Transfer Rate in actual user applications? Grubb and Cotton (1975) present the following table of "transfer rate requirements" for interactive teleprocessing services, based on three separate usage measurements. (References 5, 6, and 7 in the table are referenced in this report as Schwartz et al., (1972), Jackson and Stubbs (1969), and Fuchs and Jackson (1970), respectively.) Note that the term "system" in this table refers to the teleprocessing computer, including the application program (an end user in FED STD 1033 terms).

	Speed in Bits per Second			Average of Both
	Signaling Speed	User	System	
TYMNET (ref. 5)	110.0	3.5	35.0	-
GE Information Services (ref. 5)	110.0	-	49.0	-
	300.0	-	147.0	-
Jackson, Stubbs and Fuchs of Bell Telephone Labs (ref. 6 & 7):				
moderately loaded scientific	110.0	3.4	61.6	25.2
heavily loaded scientific	110.0	1.9	14.7	10.5
moderately loaded business	150.0	5.6	58.1	28.0

The effects of user delay on throughput are evident in the above table. As Grubb and Cotton point out, the strong asymmetry of the operator-to-computer and computer-to-operator paths suggests that separate rate specifications for the two paths may be appropriate.

Transfer rate requirements for operator-to-operator applications are normally somewhat higher than the operator-to-computer values cited above, because a more relaxed communication format and less "think time" are involved. Values in the range of 10 to 20 bps are probably typical of operator-to-operator transactions when "listening time" is included; corresponding "continuous transmission" values would be about twice as high.

Transfer rate requirements are typically much higher (and less variable) in applications where human operators are not involved. Current Bit Transfer Rate values for existing computer-to-computer transactions are in the range of 10^2 to 10^4 bits per second. Requirements on the high end are increasing as distributed processing and process control applications expand.

Carriers traditionally specify offered services in terms of the (continuous) Signaling Rate of the transmission channel rather than in terms of user information

transfer rate. Martin (1976) lists some 20 widely available data communication services with Signaling Rates ranging from 45 bps (for switched sub-voice grade channels) to 500,000 bps (for dedicated wideband channels). Among switched services, the highest Signaling Rates commonly available are about 56 kbps. Typically, the dedicated wideband services interconnect switches, concentrators, or multiplexers (i.e., system components) rather than individual end users. For a sampling of planned service offerings, see AT&T (1978) and Xerox (1978).

Because the FED STD 1033 transfer rate parameters describe the flow of user information between end users, their values will normally differ substantially from the corresponding "raw channel" Signaling Rates. The Bit and Block Transfer Rate Efficiencies directly relate these two quantities as described in the following section.

4.3.7 Bit/Block Rate Efficiency

Bit Rate Efficiency and Block Rate Efficiency describe the average proportion of the information transfer capacity connecting a user pair that is actually used (successfully) in transferring user information between that pair. The numerator of Bit Rate Efficiency is Bit Transfer Rate (as defined in the preceding section); and the denominator is the Signaling Rate of the communication service interconnecting the source and destination users. Block Rate Efficiency is defined in a similar manner, as the product of the Block Transfer Rate and the average block length divided by the Signaling Rate.

FED STD 1033 defines the Signaling Rate of a communication service as follows:

"Signaling Rate is defined as the maximum rate, in bits per second, at which binary information could be transferred (in a given direction) between users over the telecommunication system facilities dedicated to a particular information transfer transaction, under conditions of (1) continuous transmission, and (2) no overhead information."

For a single channel, Signaling Rate is expressed as

$$R_{\max} = \frac{1}{\tau} \log_2 \eta$$

where η is the number of significant conditions of modulation (levels) of the channel, and τ is the minimum time interval (in seconds) for which each level must be maintained. In the case where an individual end-to-end telecommunication service is provided by parallel channels, Signaling Rate is expressed as

$$R_{\max} = \sum_{i=1}^W \frac{1}{\tau_i} \log_2 \eta_i$$

where w is the number of parallel channels, τ_i is the minimum interval for the i th channel, and n_i is the number of levels for the i th channel. In the case where an end-to-end telecommunication service is provided by tandem channels, the end-to-end Signaling Rate corresponds to the lowest signaling rate among the component channels. The expression $1/\tau$ corresponds to the so-called "baud rate" of a circuit (Martin, 1976).

Signaling Rate has a clear interpretation in the case of traditional dedicated and circuit-switched services: it is the clock rate of the modems at each end of the circuit. Thus, for example, the conventional 1200, 2400, 4800, and 9600 bit per second modem "speeds" each correspond to a channel Signaling Rate as defined in FED STD 1033. In systems employing synchronous time-division multiplexing (e.g., T1 Carrier), the Signaling Rate corresponds to the subchannel transmission rate allocated to each user pair (e.g., 64 kbps in the T1 case).

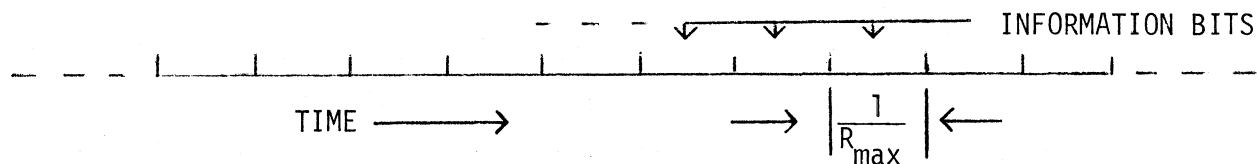
The meaning of Signaling Rate is less obvious in the case of asynchronous time-division multiplexing systems (e.g., packet and message switched systems). The "continuous transmission, no overhead" provision always allows a conceptual user-to-user Signaling Rate to be defined in such systems: that rate corresponds to the maximum flow the system could support between the given source and destination if all available system capacity were allocated to that purpose. The so-called "Max-Flow Min-Cut Theorem" of network theory states that this maximum flow is numerically equal to the capacity of the minimum "cut" between the two users, i.e., the lowest capacity collection of circuits whose removal from the network would stop all flow between the users (Frank and Frisch, 1971). Kleinrock (1976) defines a specific procedure for calculating this maximum flow.

Rate efficiency as defined in FED STD 1033 is actually among the most commonly used data communication performance parameters. For a sampling of recent uses in system optimization and error control, see Kleinrock et. al., 1976; Cacciamani and Kim, 1975; and Boustead and Metha, 1974. A survey of these and other references reveals that many different names (e.g., line efficiency, throughput efficiency, information transfer efficiency, transmission efficiency) are used in defining the same basic ratio. In each case, the goal is to express actual productive output (bits successfully transferred) as a proportion of maximum possible output (transfer capacity allocated). Rate efficiency is similar, in concept, to the traditional Power-In/Power-Out ratio used in assessing energy conversion equipment.

One instructive way of subdividing rate efficiency into component elements is the following:

$$\left\{ \begin{array}{c} \text{Rate} \\ \text{Efficiency} \end{array} \right\} = \left\{ \begin{array}{c} \text{Coding} \\ \text{Efficiency} \end{array} \right\} * \left\{ \begin{array}{c} \text{Transmission} \\ \text{Efficiency} \end{array} \right\} * \left\{ \begin{array}{c} \text{Input} \\ \text{Efficiency} \end{array} \right\}$$

To distinguish these factors, we can divide the user information transfer phase into a succession of equally spaced "slots" in time, each capable of containing one bit of information:



In a system with 100% rate efficiency, all of the slots would be filled with successfully transferred user information bits. Such a situation is not practically attainable for three reasons:

1. A certain proportion of the slots must be set aside for the transmission of overhead information, to support system functions such as block delimiting, error control, and flow control. Coding efficiency expresses the average proportion of the slots that are not allocated to overhead functions.
2. Transmission delays and imperfections will cause some of the remaining slots to be left empty, or filled with bits that are ultimately rejected or delivered in error. Each such failure excludes one slot from the possibility of containing a Successful Bit Transfer outcome. Transmission efficiency expresses the average proportion of user information slots that are not so excluded.
3. In general, the source user will not input user information bits to the system at the maximum possible rate. Input efficiency expresses the average proportion of the available user information slots that are, in fact, filled with user information bits.

The latter factor is discussed more fully in connection with the ancillary parameter User Message Transfer Time Fraction in Section 4.6.

Rate efficiency is significant as one measure of the cost effectiveness of a data communication service. Low rate efficiency implies that very little of the allocated transmission capacity is actually serving the users' needs. This in turn implies that the service may be overpriced or artificially subsidized, since the cost of providing a service is directly related to its capacity. This economic concern applies directly to the individual users in cases where a fixed transport capacity is assigned to each user pair (e.g., dedicated services). It applies to the overall user population in cases where a given transport capacity is shared among users on a demand-assignment basis.

On the surface, it would seem that higher rate efficiencies are always better than lower ones. This is often not true in the limit, however, for two practical reasons: (1) very high efficiencies may be costly to achieve (e.g., unrealistically high subsystem accuracy and reliability requirements); and (2) other desirable performance characteristics may be sacrificed in order to drive efficiency to very high values. One such characteristic, termed "reserve capacity," is discussed in Section 4.6.

The rate efficiencies are fundamentally design parameters, and the primary motive for specifying their values as user requirements is to place certain constraints on system design. Minimum rate efficiency values might be specified in a large system procurement, for example, to eliminate design solutions that are excessive in their use of rf spectrum or other resources that are difficult to quantify in monetary terms. Minimum rate efficiency requirements can also be used to ensure that resource sharing opportunities are considered in all candidate designs.

The three efficiency factors defined earlier provide a useful framework for discussing the impact of system design on rate efficiency. There is a fundamental trade-off between coding efficiency and transmission efficiency in data communication system design: a basic goal of error control, in fact, is to maximize the product of these two factors. Figure 4-3 shows the nature of this trade-off as it applies to retransmission error control systems. In such systems, a group of m user information bits is combined with a group of k error control (parity or CRC) bits to form an error control block of length n . The parity bits are used to detect transmission errors at the destination, and all received blocks with detected errors are retransmitted. The coding efficiency of such a system is m/n , and the transmission efficiency (assuming continuous transmission) is the average proportion of the user information bits that are successfully transferred on the first trial. This proportion decreases with increasing block size, since a bit error (requiring block retransmission) must ultimately occur. If too small a value is chosen for n , the result is high transmission efficiency but low coding efficiency; conversely, if too large a value is chosen for n , the result is high coding efficiency but low transmission efficiency. Kuhn (1963) and others have shown that there is an optimum block size, given any Signaling Rate and Bit Error Probability, for which rate efficiency is maximized.

A similar trade-off exists in the case of Forward Error Correction (FEC) error control systems, except that the major penalty for a wrong choice on the high end is degradation in delivered accuracy. So-called "hybrid" error control systems

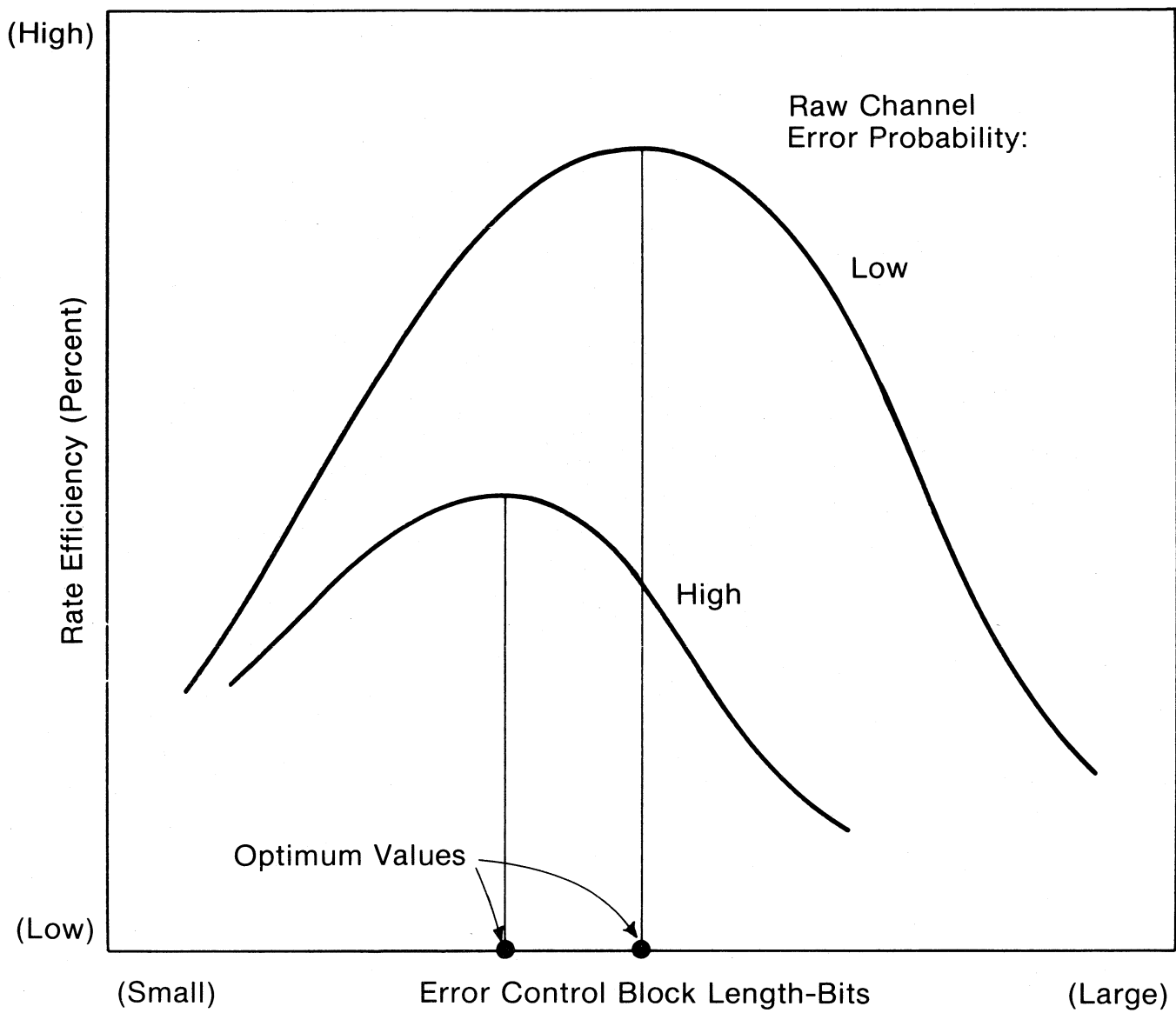


Figure 4-3. Relationship between rate efficiency and error control block length.

combine strong error detection with limited error correction to enable the use of longer, more efficient error control blocks (Nesenberg, 1975).

Two general design approaches can be used to increase the input efficiency of a communication system or subsystem:

1. The input of user information can be placed under system (terminal) control.
2. The transmission needs of many users can be aggregated in a resource-sharing device (e.g., concentrator).

The latter option involves a trade-off of two opposing cost factors: transmission cost and the cost of the resource-sharing equipment. There is a strong trend in modern design towards the use of sophisticated resource-sharing devices ("fancy switches") for two reasons: growing spectrum scarcity and the impressive cost/performance improvements of integrated circuit technology.

System values for Bit and Block Rate Efficiency are substantially influenced by both the design trade-offs just discussed. Perhaps the highest user-to-user rate efficiency values (better than 90%) are provided by synchronous services using low-error, low-delay terrestrial circuits and automated (terminal controlled) input. Synchronous operation and low error probability allow high coding efficiency; low delay and low error probability provide high transmission efficiency; and terminal-controlled input ensures high input efficiency.

Relaxing any of the above constraints decreases rate efficiency. Asynchronous operation decreases coding efficiency by adding start-stop (overhead) bits to the transmitted bit stream. High error probability decreases the product of coding and transmission efficiency in the manner described earlier. In-transit storage and lengthy propagation delays decrease transmission efficiency by "emptying the pipe". Manual (operator-controlled) input decreases input efficiency by adding "think time" and other user idle periods to the transfer interval.

Traditional message-switching systems have the lowest rate efficiency values on an user-to-user basis, but this requires a bit of explanation. In message switching systems the transmission links interconnecting an S-D pair are used successively, rather than concurrently, in transferring information between that pair. At any given point in transfer, at most one link is actively transmitting S-D information; the others are unused as far as that particular S-D pair is concerned. It is this "unused" capacity that makes the user-to-user rate efficiency of message switched services so low - the "pipe" is never full.

The important qualifying point, of course, is that the "unused" links in an end-to-end S-D path may be used to transport messages between other user pairs.

Thus, the rate efficiency of individual links in a message switching system (and the efficiency of the overall system) may be quite high when all user-to-user traffic is considered. Individual user-to-user rate efficiencies measured over a common time interval may be added to estimate the latter (composite) values.

The following excerpts from Kleinrock et al. (1976) provide a good brief summary efficiency issues encountered in the design and operation of the ARPANET, a prototype packet switching network:

- "About 64 percent of the traffic currently being carried by the ARPANET is background traffic [routing messages, line status messages, and status reports]. A large percentage of the background traffic is due to routing messages. The number of data bits per second is only about one half of one percent of the line capacity. The line utilization including all types of overhead is 6.73 percent."
- The best line efficiency (i.e., percentage of data bits) one can hope to achieve is about 20 percent (a conservative estimate of the 23.44 percent shown), because the delay increases indefinitely as the net saturates."
- "This, of course, is an average number. In particular cases one may get a far better line utilization. However, if the overall traffic characteristics remain constant, not more than roughly 10 of the 50 KBPS will, on the average, be available for process-to-process communication."
- "It appears that in some cases the freedom which the ARPANET protocols provide its implementers has been misused. In order to reduce the overhead, much more thought must be spent on the efficient implementation and use of network protocols, rather than only on their feasibility."
- "In view of these results we hope that in the future the design, implementation, and use of communication protocols take more account of the effect of overhead on the user-to-user throughput and thereby improve the network performance."

4.4 Disengagement Parameters

At one time or another, practically everyone has experienced the frustration of attempting to disengage from a system after receiving a requested service, only to find that the system delays disengagement excessively - i.e., "won't let you go." Whether the cause is a local telephone operator who forgets to "pull the plug" or a data communication system that loses your Close request, such a situation is a nuisance at best. User concerns with disengagement performance fall naturally into two categories:

Efficiency - How long will it take to complete disengagement from the system, assuming I am successful in doing this?

Accuracy, Reliability - What is the likelihood that my disengagement attempt will fail as a result of either error or nonperformance on the part of the system?

FED STD 1033 defines two disengagement performance parameters which address these specific user concerns: Disengagement Time and Disengagement Denial Probability.

4.4.1 Disengagement Time

Disengagement Time is the average time a user must wait, after requesting disengagement from an information transfer transaction, for the system to successfully accomplish the disengagement function. As noted in Section 3.3.2, a separate disengagement function is defined for each end user participating in a transaction. For each user, computation of Disengagement Time begins on issuance of a Disengagement Request by the user, and ends on subsequent issuance of a Disengagement Confirmation by the system. The parameter Disengagement Time is computed by averaging the individual disengagement times for all users participating in a transaction.

In some systems, (e.g., the public switched telephone network), no explicit Disengagement Confirmation signal is issued by the system. In such cases, the end of the disengagement function is defined to occur, for each end user, when that user is first able to initiate a new access attempt. Disengagement Time values are calculated only on disengagement attempts that result in Successful Disengagement.

Examples of specific Disengagement Request and Disengagement Confirmation signals have been cited earlier (Sec. 3.3.2). Identifying these signals in particular systems is normally straightforward, but there are two particular cases which should be mentioned. The first is the case of two-point circuit-oriented transactions. In such transactions, disengaging one user necessarily implies disengaging the other, since a "circuit" with only one end has no meaning. Both disengagement functions therefore start with a single Disengagement Request. The two functions nevertheless end with different events, as discussed earlier.

The second special case of interest is that of pre-emption. In some systems (e.g., AUTOVON), an ongoing information transfer transaction may be terminated by the system in order to free transmission resources needed by higher priority users. Although such events could be treated as system-initiated disengagements, it is more consistent with the service concept (and the attitude of the interrupted users) to treat them as outages. This is particularly true since a system with pre-emption may or may not notify low-priority users of impending disconnection. Thus, FED STD 1033 Disengagement Request signals are always user initiated. As in the case of Access Requests, they may be either explicit or implicit.

Disengagement Time appears to have no counterpart in previously defined data communication performance parameters. The ANSI standard parameter Total Overhead Time (ANSI, 1974) includes Disengagement Time as a component, but the contribution of the latter to the former will normally be small. There are some obvious similarities between Disengagement Time and Access Time, and in fact the disengagement and access functions are implemented by an identical "four-way handshake" in many virtual-circuit systems (e.g., the ARPANET). The two functions do differ, of course, in the definition of ending events.

Why, one might ask, should a user be concerned with the time the system takes to disengage him when his information has already been transmitted? The answer is that Successful Disengagement is often an essential prerequisite to other user activities. The most obvious such activity is communication with another distant user; but local communications may be affected as well. As an example of the latter situation, consider an operator who uses a data terminal to communicate with both distant and local computer programs. If his terminal remains logically connected to a distant program for a substantial period of time after he requests disengagement, he will be delayed in initiating communication with local programs; and his personal efficiency will suffer as a result. Such delays are not at all unusual in some distributed computing systems, including, in our experience, the ARPANET (Payne, 1978).

Disengagement Time values range between 0 and a measured upper bound equal to 3 times the "nominal" Disengagement Time specified for the service, as described earlier. A value of 0 implies that all users involved in a transaction are free to participate in new transactions immediately upon requesting disengagement. Very large values for Disengagement Time suggest a system that not only wastes the users' time, but its own as well.

Appropriate user requirements for Disengagement Time depend on the individual usage pattern. Values less than a second may be appropriate in applications where a user initiates information transfer transactions continuously (e.g., "round robin" polling systems). Disengagement Time adds directly to the total round robin cycle time in such systems, and thus contributes directly to the "age" of the transmitted data. Such aging is of particular concern in military command and control applications (e.g., the Naval Tactical Data System).

At the opposite performance extreme are applications where service usage periods are always, or almost always, preceded by a long idle period. An example would be a retail inventory control system where accumulated receipts are transmitted to a central computer for processing once per day. User requirements for

time sharing applications are typically intermediate between these extremes; as an example, a Disengagement Time of 10 seconds is specified in EPA (1980). It is normally appropriate (and technically feasible) to specify a Disengagement Time short enough to ensure that disengagement will not delay the next access attempt.

Data on Disengagement Times for existing systems is sparse. One can infer minimum values in the neighborhood of 1 to 2 seconds for modern circuit-switched systems, since shorter depressions of a telephone hookswitch are often used to signal an operator or activate special functions. Linfield and Nesenbergs (1978) cite a typical value for "disconnection time" in electronic switching systems of 2 seconds. Payne (1978) reports measured values for operator Disengagement Time in the ARPANET (a virtual-circuit packet switching system) in the range of 5.0 to 5.6 seconds. The latter values apply specifically to the Telnet protocol, and include 3.3 seconds operator typing time for the CLOSE request.

The design features that most strongly influence Disengagement Time are the type of resource sharing employed and its degree of automation. Services provided by dedicated lines typically offer the shortest Disengagement Times, because there are no shared system facilities which must be freed for use by other subscribers.¹⁸ Datagram and message-switched services typically have somewhat longer Disengagement Times because there are local buffers in the source switch which must be freed for other users. Circuit-switched and virtual-circuit systems typically have the longest Disengagement Times because there are shared resources (e.g., trunks) to be freed at both ends of the information path. Kimmatt and Seitz (1978) calculate Disengagement Time values of 0.5, 1.5, and 2.25 seconds for typical nonswitched, message-switched, and circuit-switched services, respectively.

4.3.3 Disengagement Denial Probability

Disengagement Denial Probability expresses the likelihood that a system will fail to detach a user from an information transfer transaction within a specified maximum time after he/she issues a Disengagement Request. It is defined as the ratio of total Disengagement Denial outcomes to total disengagement attempts included in a performance sample, excluding attempts that fail as a result of user nonperformance.

The Disengagement Denial outcome is indicated by either (1) failure of the system to issue a Disengagement Confirmation signal within the disengagement time-out period, in systems that provide such a signal; or (2) failure of the system

¹⁸The purpose of disengagement in such systems is simply to return the users to an established "idle" state after service usage.

to respond to a subsequent access attempt within the same period, in systems that do not. The duration of the disengagement timeout period is determined by the "three times nominal" rule described earlier. Disengagement Denial is distinguished from User Disengagement Blocking by comparison of ancillary parameter values as discussed in Section 4.6.

Disengagement Denial Probability is significant to data communications users for two reasons. The first is that it provides information about the shape of the Disengagement Time distribution. Like Access Time, Disengagement Time is the average of a truncated distribution (Fig. 4-1). The probability that an individual disengagement attempt will exceed three times the specified value will be relatively high if the spread (variance) of the Disengagement Time distribution is large, and vice versa. For systems with nearly constant Disengagement Times, Disengagement Denial Probability values will be very low - in essence, only "hard failures" such as switch crashes will cause disengagement timeout.

Disengagement Denial Probability is also significant to data communications users as a measure of system reliability. When a system gets "hung up" in disengagement, the effect on the users is often similar to that of an outage - the service is unavailable until the problem is corrected.¹⁹ Disengagement Denial differs from outage, however, in the fact that responsibility for correcting the problem often falls on the user rather than the communication manager or service supplier.

Possible Disengagement Denial Probability values range between 0 and 1. A value of 0 implies that the system never fails to disengage a user within the allotted timeout period. A value of 1 suggests a system that never lets the users go without some sort of unplanned reset action.

Appropriate user requirements for Disengagement Denial Probability depend, like these for Disengagement Time, on the service usage pattern. Here again, low values are appropriate in polling and similar applications, while quite high values can be tolerated in applications where service usage is normally preceded by a long idle period. Reliability requirements and the availability of backup facilities should also influence user requirement specifications. Nesenbergs et al. (1980) suggest a Disengagement Denial Probability requirement of 10^{-3} for interactive AUTODIN II users. A value of 10^{-5} is specified in EPA (1980).

¹⁹The ANSI availability definitions lump disengagement denial time with other "inoperative" (outage) time on the basis of this similarity.

Disengagement Denial Probability values are influenced by two general system design characteristics:

1. The relative complexity of the disengagement protocol employed.
2. The inherent accuracy and reliability of the facilities which implement that protocol.

In general, the lowest Disengagement Denial Probability values are found in message-switched and "datagram" services. In such services, the disengagement of each user is a simple, local function which involves only that user and his associated switch. Successful Disengagement of the source does not require communication with the destination, and it is therefore uninfluenced by transmission imperfections. Successful Disengagement of the destination requires only one passage of a Disengagement Request signal through the system, from the source to the destination.

Disengagement is more complex, and therefore more subject to failure, in systems that employ a virtual circuit protocol. In such systems, disengagement typically involves a "full four-way handshake" between transaction participants - e.g., transfer of a Close message from source to system, system to destination, destination to system, and system back to source. Successful Disengagement of the source thus requires two successful passages of a Close message through the system. If such a protocol is combined with a flow control mechanism which discards packets to control congestion or excessive delay, Disengagement Denial may be a rather frequent occurrence.

The ARPA network illustrates exactly such a situation. Kleinrock et al. (1976) report a 10^{-2} loss probability for messages entering that network. Logically, one would expect the loss probability for one or both of two Close requests to be about twice as high - a Disengagement Denial Probability of 2×10^{-2} . Payne (1978) reports a measured value for this parameter in the ARPANET of 3×10^{-2} .

4.5 Secondary Parameters

The primary parameters are intended to describe system performance during periods of normal, reasonably satisfactory operation. Users are understandably also concerned with the frequency and duration of total (or near-total) service failures - i.e., the long-term availability of service. FED STD 1033 defines three closely related "secondary" performance parameters which address these user concerns - Service Time Between Outages, Outage Duration, and Outage Probability. This section describes these three parameters in essay form, using the same general outline employed in the preceding sections: i.e., meaning, significance, values,

and design implications. The three parameters are described together to emphasize interdependencies and similarities.

The overall approach used in defining the secondary (availability) parameters has been described in Section 3.3.5 and illustrated in Figure 3-14. The essence of that approach is the user view that an outage is not an equipment failure, or a signal fade, or a software crash that happens somewhere inside the system; it is an unacceptable degradation in the performance of an end-to-end transfer service connecting two users. In the context of that definition, the meaning of the FED STD 1033 availability parameters can be explained as follows.

Service Time Between Outages describes how long, on the average, a system provides satisfactory transfer performance to a user pair between successive instances in which it fails to do so (outages). More precisely, Service Time Between Outages is the total (average) time from the start of the first sample to the end of the last sample in any consecutive group of transfer samples ("messages") in which satisfactory end-to-end performance is provided. Service Time Between Outages is measured against a discontinuous time scale which includes only actual User Information Transfer Time (Section 3.3.5); any access, idle, or disengagement time between successive user information transfer intervals is excluded. Satisfactory UIT performance (i.e., the Operational Service state) is declared on any transfer sample in which measured performance is better than threshold for each of the five supported performance parameters.

Outage Duration describes how long, on the average, a system remains unable to provide satisfactory transfer performance to a user pair in any given outage instance. It is the total (average) time from the start of the first sample to the end of the last sample in any consecutive sequence of transfer samples in which unsatisfactory end-to-end performance is provided. Outage Duration is also measured against a discontinuous UIT time scale. Unsatisfactory UIT performance is declared on any transfer sample in which measured performance is worse than threshold for any of the five supported performance parameters.

Outage Probability expresses the likelihood that the transfer service interconnecting a user pair will be declared to be in an Outage state on any given sample or trial. The numerator of the probability ratio is the total number of Outage samples observed during an availability measurement period; and the denominator is the total number of Outage and Operational Service samples observed during the same period. Samples having a Bit Transfer Rate below threshold as a result of user nonperformance (e.g., slow input) are excluded from both totals.

Service Time Between Outages and Outage Duration correspond, respectively, to the traditional availability parameters Mean Time Between Failures (MTBF) and Mean Time To Repair (MTTR). The former parameters are more specialized than the latter in two respects: (1) they specifically exclude nonoperating time, and (2) they embody a particular sampling procedure and associated "failure" and "repair" definitions. MTBF and MTTR normally describe the availability of a particular system component (e.g., terminal, transmission link, or switch); their FED STD 1033 counterparts describe the availability of an end-to-end transfer service. MTBF and MTTR are simply related to the performance parameter Availability as follows:

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

Availability is derivable from MTBF and MTTR, but the converse is not true - many different MTBF/MTTR combinations can produce a given availability value.

A frequently quoted definition for reliability is that of the Advisory Group on Reliability of Electronic Equipment (AGREE, 1957):

"Reliability is the probability of performing without failure a specified function under given conditions for a specified period of time."

Outage Probability is essentially a "specialized complement" of reliability as defined above: i.e., it is the probability that a specified function (message transfer) will not be performed successfully (with all supported parameter values better than threshold) under given conditions (particular source and destination) for a specified period of time (the message transfer interval). In essence, it is the likelihood that a system will be unable to maintain satisfactory performance during the transfer of a specified number of user information bits.

One might argue that Outage Probability is also derivable from MTBF and MTTR, but that is only true if bits are always transferred between end users at a uniform rate. This is rarely the case in practical applications, and Outage Probability often differs substantially from unavailability. This difference is illustrated in a later example.

The significance of the three secondary performance parameters specified in FED STD 1033 can be clarified by a simple analogy - the reliability of your personal automobile. Service Time Between Outages is the average trouble-free driving time you experience between "breakdowns" which make your car unusable. Outage Duration is the average time you must "do without" your car on any given breakdown. Outage

Probability is the likelihood that your car will break down on a trip between two specified locations. Many different design features will improve the values for these three parameters; but as a driver, your basic interest is in the desired end result - a long interval between breakdowns, rapid repair, and little likelihood of a spoiled trip.

Possible values for Service Time Between Outages and Outage Duration begin at zero and have no theoretical upper bound. Very low values for Service Time Between Outages or very high values for Outage Duration indicate a system that is rarely available, and conversely. Possible values for Outage Probability range between zero and one, with low values indicating high likelihood of successful transfer and conversely.

Data on user requirements for availability is plentiful, but the individual values for MTBF and MTTR are rarely distinguished. This is a deficiency in current specification practices - infrequent long outages and frequent short outages may produce the same overall availability, but have profoundly different effects on the end users. Typical availability values in current requirement specifications are in the range of 90% to 99.9%, with values above 98% much more common than those below.

The following excerpt from the "boiler plate" of a GSA Federal Supply Service ADP Schedule Price List is typical of a specification on the lower end:

"PERFORMANCE REQUIREMENTS

a. All equipment furnished under this contract shall perform the function for which it is intended in accordance with the manufacturer's specifications and other presentations at an average effectiveness level of 90%.

b. The average effectiveness level is a percentage figure determined by dividing the total productive time (time used) by the sum of the total productive time and the downtime (lost productive time) less travel time (not to exceed two (2) hours) multiplied by 100.

$$\frac{\text{Productive Time} \times 100}{\text{Productive Time} + \text{Downtime} - \text{Travel Time}} = \text{Effectiveness Level}$$

c. Downtime (lost productive time) for each incident shall be calculated from the time the Government has made a bona fide attempt during regular working hours to contact the contractor's designated representative at the prearranged contact point until the system or machine is returned to the Government in proper operating condition. If any downtime should be occasioned by fault or negligence of the Government, all such downtime shall be excluded for the purpose of calculating the average effectiveness level."

Note that "average effectiveness level" as defined here excludes maintenance travel time of up to 2 hours from the computation of Outage Duration.

EPA (1980) has specified availability requirements for the communications portion of a nation-wide time-sharing network as follows:

"The Contractor shall provide circuit availability levels of at least 0.999 within the following service parameters:

a. Outage Time

- (1) Not involving remote local loops
Mean = 0.3 hours; Standard Deviation = 0.2 hours
- (2) Involving remote local loops
Mean = 1.3 hours; Standard Deviation = 1.5 hours

b. Probability of Outage = no greater than 5×10^{-5} ."

These requirements are relatively stringent in comparison with those specified in other recent Federal procurements.

Availability specifications for existing data communication systems are mostly in the range of 98% to 99.9%. Here again, the relative contributions of MTBF and MTTR are rarely distinguished. As noted earlier, an availability value of 98% is probably typical of dedicated communication links (including the modems but not the terminals). This relatively low value is explained by the fact that such services normally have no backup provisions; a time-consuming maintenance action is normally required to restore service when an outage occurs. If one assumes an average of 4 hours for such action (i.e., average Outage Duration of 4 hours), the corresponding value for Service Time Between Outages is 196 hours - about one outage per month in a service used 8 hours per day, 5 days per week. The latter figure is approximately doubled by increasing the availability from 98% to 99%.

Specified availability values for circuit- and message-switched services are often slightly higher than those for dedicated services - typically, in the neighborhood of 99%. In the case of circuit-switched services, the primary source of improvement is the inherent redundancy circuit switching provides - there are many possible circuit paths between a given source and destination user. The effect of this redundancy is to reduce Outage Duration substantially. As an example, a user who encounters a bad connection can often restore satisfactory service himself, in a very short time, by simply hanging up and re-dialing.

An additional factor in the case of message switching is the fact that the transfer interval for a message (the "service time") is inflated by switch storage time (Kimmitt and Seitz, 1978). This effect makes direct comparison of availability

values between circuit- and message-switched systems somewhat misleading; Outage Probability is a better standard of comparison in such cases because it is based on units of output (bits transferred) rather than time.

Availability values for packet switching systems are normally dominated by the inherent reliability of the terminal switching nodes. The 1.64% "down rate" cited earlier for individual ARPA network IMPs would imply a subnetwork availability per user pair in the neighborhood of 97%.

In a filing submitted to the FCC, DATRAN (1975) made the following statements with regard to the quality and reliability of their Datadial[®] switched digital network (now operated by Southern Pacific Communications Company):

"DATRAN presently offers a guarantee of quality which states that refunds will be given if the error-free-second rate on a circuit is less than 99.95%. During an outage (defined as a period during which the error-second rate is greater than 50%), the quality standard (99.95% error-free-seconds) is definitely not being met so circuit outages become automatically included, and a separate guarantee would be redundant."

The availability of Datadial[®] network was thus specified to be better than 99.95%. A recent Bell System advertisement for its Dataphone[®] Digital Service offering contains the following statement:

"Availability is a key factor. The system is designed to be in service 99.96% of the time."

Neither of these specifications gives any indication of the underlying MTBF/MTTR components. In the hypothetical situation of a 4-hour MTTR, these service specifications would imply a Service Time Between Outages in the neighborhood of 10^4 hours - over a year's worth of 24-hour-per-day operation between outages! It should be noted, however, that both these specifications apply to subsystems from the end user perspective, since the "customer terminals" are not included. One would hope that Federal customers are not attaching terminals procured under the 0.9 "effectiveness level" cited earlier to the above services!

4.6 Ancillary Parameters

Telecommunication performance is user dependent. That fundamental fact underlies all that will be said here about the ancillary performance parameters, and indeed justifies their existence. It is a fact that is often disregarded in the design and operation of telecommunication systems, with the result that many existing data communication services are highly inefficient in meeting end user needs. This section shows how user dependence can be quantified and measured

via the ancillary parameters to improve end-to-end communication performance and cost effectiveness.

What exactly do we mean when we say that telecommunication performance is "user dependent"? The key points are these:

1. Most data communication systems require user inputs at various points in an information transfer transaction.
2. The user actions which generate those inputs inevitably take time. Often, the system has no alternative but to delay its own activities until the necessary user actions are accomplished.
3. The time required to complete the primary communication functions is therefore often dependent on user performance time.

In sum, the users and the system must normally be regarded as co-responsible entities who jointly determine overall communication performance. A primary purpose of the ancillary parameters is to describe the relative contributions of the users and the system to overall communication delay.

Two examples of user dependence have been presented in earlier sections of this report. In the voice telephone access example, Figure 3-3, we observed that the overall performance time for the access function in the public switched network depends on both the system's speed in signaling and switching and the user's speed in dialing and answering. We also saw, in Sections 4.3.6 and 4.3.7, the effect of user delay on message transfer performance. That effect is most apparent in cases where the user information is input by a terminal operator. In such cases, user performance time may increase the overall transfer time for a message by a factor of ten or more, reducing transfer rates and rate efficiencies proportionally.

User delays can also influence performance of the block transfer and disengagement functions. "Mailbox" services provide a familiar example of the former: in such services, the system cannot deliver stored information to a destination user until that user logs on and asks to read his mail. Certain types of flow control may also introduce user delays in block transfer. Disengagement is clearly user dependent in systems (such as the ARPANET) that require a "full four-way handshake" to close an established connection: the user not originating disengagement must respond to a Close request from the system before either user can be successfully disengaged (NCS, 1977).

The four FED STD 1033 ancillary parameters express these user influences on communication performance in quantitative terms. Each parameter relates directly to an associated primary function, and describes the average proportion of performance time for that function that is attributable to user delay. Ancillary

parameters are defined for the access, block transfer, message transfer, and disengagement functions. As noted earlier, no ancillary parameter is defined for the bit transfer function since its values can be inferred from the corresponding block transfer parameter.

There are relatively few precedents for the ancillary performance parameters in the formal literature of telecommunications - most published studies either ignore user dependence or "define it away" in some expedient, but often unrealistic, way. Three significant exceptions are worthy of note:

1. The study of Duffy and Mercer (1978) on network performance and customer behavior during Direct-Distance-Dialing call attempts in the United States. Among other findings, this study reports that "customer-determined components of the call setup time make up 71 percent of the total setup [access] time."
2. The study of Jackson and Stubbs (1969) on user/computer interactions in a typical remote-access timesharing system. A significant conclusion of this study is that "users themselves contribute substantially to the communications costs of their real-time computer access calls by introducing delays." Quantitative data from this study has been presented earlier.
3. The work of Kleinrock (1976) and others in applying queueing theory to computer networks. The concepts of customer "arrivals," inter-arrival times, and service times provide a natural framework for describing user dependence, although relatively few analysts have actually applied them to that problem. One such application has been described earlier in this report (Kleinrock et al., 1976).

The ancillary performance parameters are significant for two major reasons. First, each parameter can be used as a correction factor, to calculate "user-independent" values for the associated primary efficiency parameters. If W is the specified performance time for a primary function and p is the associated ancillary parameter value ($0 \leq p \leq 1$), the user-independent performance time for the function is

$$[1-p] \cdot W.$$

The factor $[1-p]$ is the average system performance time fraction - the complement of p . Similarly, given any specified rate or rate efficiency parameter value, R or Q , the corresponding user-independent value can be calculated as

$$\frac{R}{[1-p]} \quad \text{or} \quad \frac{Q}{[1-p]}$$

In each case, the user-independent values express the performance that would be provided by the system if all user delays were zero; i.e., if all user activities

were performed in zero time. As an example, assume the Access Time value for the telephone service of Figure 3-3 is specified as 25 seconds, with an associated User Access Time Fraction of 0.6. Then the user-independent Access Time value - the average total system delay during access - is $(0.4)(25) = 10$ seconds. User-independent values for the rate and rate efficiency parameters are higher than their user-dependent equivalents because the fraction $(1-p)$ appears in the denominator of the R and Q expressions.

The ancillary parameters also provide a basis for identifying the entity "responsible" for timeout failures: e.g., whether the user or the system should be charged with the failure when an access attempt is not completed within the maximum access time (Fig. 3-9). This decision is made by calculating a user performance time fraction for the particular (unsuccessful) trial in question, and then comparing the calculated value with the corresponding average ancillary parameter value. If the user fraction for the particular trial exceeds the corresponding average, the failure is attributed to the user; otherwise, the failure is attributed to the system.

As an example of this decision process, assume that in the telephone service specified above a particular access trial "times out" at 75 seconds (three times the specified Access Time). To determine whether the user or the system is "responsible", we would calculate the User Access Time Fraction for that trial, and compare the calculated value with 0.6 (the specified average value). If the calculated value was greater than 0.6, the failure would be attributed to User Blocking; otherwise, the failure would be attributed to Access Denial. This application of the ancillary parameters is described more fully in Seitz and McManamon (1978).

The ancillary performance parameters also have a direct significance which is independent of the two uses described above. That significance is slightly different for communication managers and suppliers, on the one hand, and communications users, on the other. To communication managers and suppliers, the ancillary parameters provide important information about the relative economy of a communication service. High values indicate that overall performance is dominated by user delays. In such situations, a potential for more economical service through resource sharing often exists. Data multiplexing is a familiar way of exploiting this potential. Low values for the ancillary parameters indicate that overall performance is dominated by system delays, and suggest that little resource sharing potential exists.

Communications users view the ancillary parameters from two perspectives, depending on the primary function in question. The key issue in the case of access and disengagement is ease of use. Low ancillary parameter values indicate a service that can be utilized with relatively little user investment in time and effort (e.g., an "off-hook" service); high values indicate a service that demands more of these resources (e.g., a service with lengthy, elaborate circuit establishment procedures).

The following rather humorous dialogue between the designer and a prospective user of a "computer-aided design" service illustrates user reaction to the latter type of service (Newman, 1979):

"Programmer: Now that you've drawn part of the circuit, you might want to change it in some way.

User: Yes, let's delete a component. How do we do that?

P: Point at the menu item labeled CD.

U: CD?

P: It stands for 'component delete.'

U: Ah. Well, here goes...hey, what happened?

P: You're in analysis mode: you must have selected AM instead of CD.

U: Funny, I was pointing at CD. How can I get out of analysis mode?

P: Just type control-Q.

U: [types C-O-N-T-R...]

P: No, hold down the control key and hit Q.

U: Sorry, silly of me...OK, I'll try for CD again.

P: Maybe aim a bit above the letters to avoid getting into analysis mode - no, not that much above - that's better.

U: Got it!

P: Now point to the component to delete it.

U: OK...nothing's happening; what am I doing wrong?

P: You're not doing anything wrong; you've deleted the component, but the program hasn't removed it from the screen yet.

U: When will it be removed?

P: When you type control-J to redraw the picture.

U: I'll try it...there we are; but only part of the component was removed!

P: Sorry, I forgot: you have to delete each half of the component separately. Just point to CD again.

U: Very well...now what's happened?

P: You're in analysis mode again: type control-Q.

U: Control...where's that Q? There it is...hey, why is the screen blank all of a sudden?

P: You typed Q, not control-Q, so the program quit to the operating system. I'm really sorry, but we've lost everything and we'll have to start all over again.

U: [groans] Could we postpone that until next week?"

We see here the impact a poorly designed user/system interface protocol can have on the usefulness (and utilization) of an offered service. Interface protocols have traditionally been quite simple in the case of data communication services,

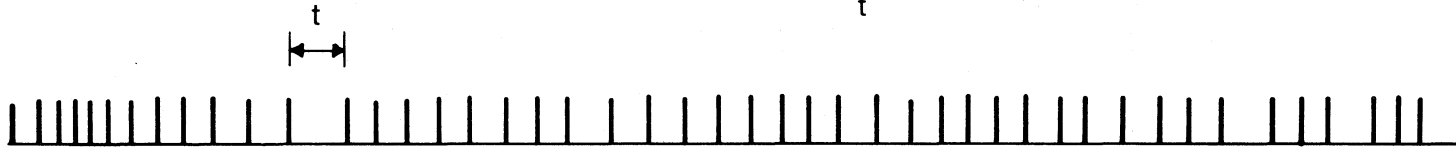
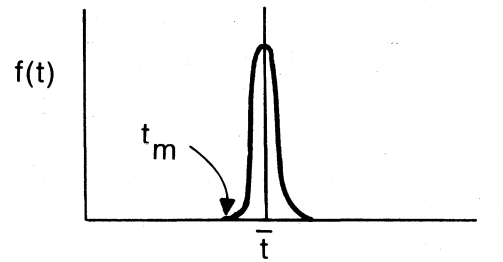
but there is a strong trend towards more powerful (and therefore more complex) protocols. The developing Open Systems Interconnection standards illustrate this trend: those standards will enable users to "customize" network services by specifying desired functions, features, and performance levels during the access phase (ISO, 1979).

Users view the ancillary performance parameters somewhat differently in the case of the user information transfer functions. Ease of use is still desirable, but its importance is often overshadowed by a concern with what might be called reserve capacity: the ability of the system to "keep up with the user" during momentary bursts of high-speed input. The more bursty the input, the more important such reserve capacity is. A familiar example of insufficient reserve capacity is a system that falls behind in "echoing" typed characters above a certain typing speed. High ancillary parameter values indicate that a substantial reserve capacity exists in a system, and conversely. A similar relationship holds the case of user-controlled output.

The above remarks provide a basis for some general conclusions about the user specification of ancillary parameter values. Since ease of use is the key issue in the case of access and disengagement, ancillary parameter values for these functions should normally be specified on the basis of the value of the user's time. Low specified values (e.g., < 0.1) are appropriate in applications where the user's time is extremely valuable; high values (e.g., > 0.9) may be tolerated in applications where the user has available time that cannot or will not be used in other productive ways. Examples of users in the former category are a tactical military commander, an air traffic controller, and a computer program controlling a critical real-time process. Users of a recreational "game network" like that proposed by Lucky (1979) might fall in the latter category.

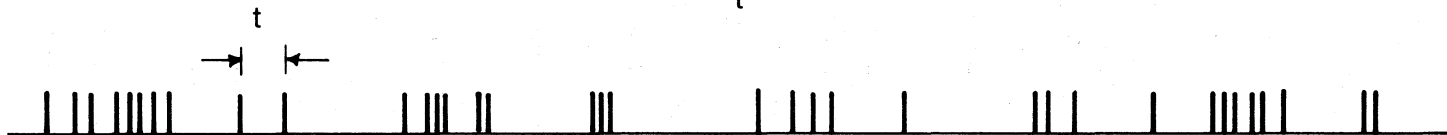
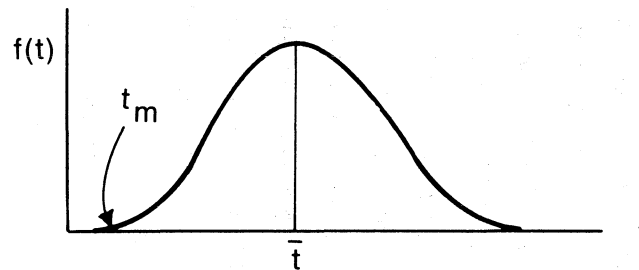
Figure 4-4 illustrates the influence of the user input pattern on the specification of User Message Transfer Time Fraction. If the user input pattern is uniform or nearly so, there is little need for reserve transfer capacity, and a relatively low ancillary parameter value is appropriate. If the user input pattern is very bursty, a substantial reserve capacity must be provided if the system is to keep up with the user during input bursts, and a higher ancillary parameter value is appropriate. If the minimum time between user inputs is t_m and the corresponding average time is \bar{t} , the user fraction of transfer time at the input interface is $[1 - (t_m/\bar{t})]$; and the reserve capacity needed to avoid delaying user input is $[(1/t_m) - (1/\bar{t})]$. The former expression approximates the ancillary parameter User Message Transfer Time Fraction when system propagation and storage times are small.

(Lower Ancillary
Parameter Values
Appropriate)



a. Nearly Uniform User Input Pattern.

(Higher Ancillary
Parameter Values
Appropriate)



b. Very Bursty User Input Pattern.

Figure 4-4. Impact of user input pattern on user message transfer time fraction.

A numerical example will help to clarify the above relationships. Assume two source users: a remote sensor generating user information characters at regular 100-millisecond intervals, and a typist generating user information characters intermittently, with average and minimum intercharacter times of 500 and 100 milliseconds, respectively.²⁰ Assume that system propagation and storage times are negligible in both cases. In the first case, the average character input rate $1/\bar{t}$ and the maximum character input rate $1/t_m$ are the same (10 characters per second); no reserve capacity is needed to accommodate bursts of user input $[(1/t_m)-(1/\bar{t}) = 0]$; and a low User Message Transfer Time Fraction is appropriate $[1-(t_m/\bar{t}) = 0]$. In the second case, the average character input rate and the maximum character input rate differ by a factor of five (2 vs. 10 characters per second); a reserve capacity of 8 characters per second is needed to accommodate "bursts" of user input $[(1/t_m)-(1/\bar{t}) = 8]$; and a relatively high User Message Transfer Time Fraction should be specified $[1-(t_m/\bar{t}) = 0.8]$.

It may seem paradoxical that a high user fraction is "bad" (from the user standpoint) in the case of access and "good" in the case of bursty message transfer. The explanation is this. In the case of access, the user is trying to obtain a service; and a larger user fraction means that he has to devote more time to that effort. In the case of transfer, the user has obtained the service; and a higher user fraction (for a given transfer rate) means that he has a faster, more responsive service. During transfer, the time between user inputs may be highly productive "think time"; this is often the case, for example, during operator interaction with a text editor or other data processing program. The ancillary parameters encourage more accurate specification of performance needs in either case.

The single design feature that most strongly influences ancillary parameter values is the user/system interface protocol. The lowest ancillary parameter values are observed in services where the system controls the transfer of information across both user/system interfaces: in such services, few or even no user actions may be required to complete a given primary function. Relatively high values are often observed in services where the users control (or participate in controlling) transfer across one or both interfaces: each required user action represents an opportunity for user delay.

In the case of access and disengagement, the ancillary parameter values are largely determined by the type of switching employed. Message-switched and datagram

²⁰These intercharacter times correspond to typing speeds of 20 and 100 words per minute, respectively.

services typically provide low ancillary parameter values: the functions of access and disengagement are inherently simple in such services, and few user/system interactions are involved. The opposite is true of circuit-switched and virtual-circuit services. Dedicated services usually provide relatively low User Access and Disengagement Time Fractions because the key user choices (e.g., desired destination) are "hard wired." User Access Time Fractions of 0, 0.4, and 0.19 are calculated for particular message-switched, circuit-switched, and non-switched services, respectively, in Kimmett and Seitz (1978).

The impact of system design on the ancillary user information transfer parameters is best clarified by a number of examples. Consider first the case where the source and destination terminals are a card reader and a card punch. The "end users" are then the punched card decks. In this situation, the system (terminal) controls both the input and output of user information; and both the User Block Transfer Time Fraction and the User Message Transfer Time Fraction will normally be zero. User actions would influence transfer performance only if an operator was required to replenish card supplies during a transaction. Note that this is true irrespective of the propagation and storage delays encountered in transmission.

As a second example, consider a conventional operator-to-operator communication session. Assume the input information is typed on-line and the output information is printed. The source user controls input, while the system controls output. In such a situation, the User Fraction of Block Transfer Time will normally be zero, but the User Fraction of Message Transfer Time may be quite high. Values in excess of 90% are not uncommon, as the Grubb and Cotton (1975) data indicates.

The important point here is that system propagation and storage delays may substantially influence these values. For example, suppose the source user types 15 words per minute into a terminal capable of accepting 150. Then the user fraction of input time is 90%. If system propagation and storage times are negligible (e.g., milliseconds), this value will correspond closely to the User Message Transfer Time Fraction. But if the transmission system is a message switching system that takes 6 hours to move the information from source to destination, the User Message Transfer Time Fraction will be quite small - less than 2% for a 100-word message. Again, this parameter describes the users' contribution to the total end-to-end transfer time for a message.

As a final example, consider the case of a simple "mailbox" service. Assume the user input is generated on-line as in the preceding example, but the system output is initiated only when the destination user specifically asks to read his

mail. The system notifies the destination user of queued mail via a "you have new mail" printout.²¹ In this situation, the destination user is often the major cause of end-to-end transfer delay - system propagation time may be negligible by comparison. For example, if the destination user reads his mail only once per day and messages are addressed to him at random intervals, the user fractions of block and message transfer time will exceed 90% for any input-to-notification delay less than 72 minutes!

These examples illustrate that the ancillary parameter values are strongly influenced by both user/system interface protocols and internal transmission times. To ensure cost effectiveness, both influences should be considered in developing communication performance specifications.

4.7 Summary

Figure 4-5 summarizes the 26 standard performance parameters in the context of a typical FED STD 1033 Performance Specification Form. As noted earlier in this report, these standard parameters can be used in three distinct ways: (1) to specify a particular data communication requirement, from the point of view of the end user; (2) to specify a particular service offering, from the point of view of the communication supplier; and (3) to compare various alternative means of meeting a stated user requirement, from the point of view of the communication manager. Guidelines for using the standard in each application will be provided in a planned sequel to this report, the Interim Federal Standard 1033 Application Manual.

ACKNOWLEDGEMENT

The author thanks Bob Linfield, Dana Grubb, and Evelyn Gray for their thoughtful and constructive review, and Kathy Mayeda for her assistance in typing and manuscript preparation.

²¹The ARPANET provides such a service, and several are being implemented in commercial value-added networks.

SERVICE PERFORMANCE SPECIFICATION

Part A - Primary Parameters

- | | | |
|--|--|---------------|
| 1. Access Time | | Seconds |
| 2. Incorrect Access Probability | | * |
| 3. Access Denial Probability | | * |
| 4. Bit Transfer Time | | Seconds |
| 5. Bit Error Probability | | * |
| 6. Bit Misdelivery Probability | | * |
| 7. Bit Loss Probability | | * |
| 8. Extra Bit Probability | | * |
| 9. Block Transfer Time | | Seconds |
| 10. Block Error Probability | | * |
| 11. Block Misdelivery Probability | | * |
| 12. Block Loss Probability | | * |
| 13. Extra Block Probability | | * |
| 14. Bit Transfer Rate | | Bits/Second |
| 15. Block Transfer Rate | | Blocks/Second |
| 16. Bit Rate Efficiency | | % |
| 17. Block Rate Efficiency | | % |
| 18. Disengagement Time | | Seconds |
| 19. Disengagement Denial Probability | | * |

Part B - Secondary Parameters

- | | | |
|--|--|-------|
| 20. Service Time Between Outages | | Hours |
| 21. Outage Duration | | Hours |
| 22. Outage Probability | | * |

Part C - Ancillary Parameters

- | | | |
|---|--|---|
| 23. User Access Time Fraction | | * |
| 24. User Block Transfer Time Fraction | | * |
| 25. User Message Transfer Time Fraction | | * |
| 26. User Disengagement Time Fraction | | * |

*Note: The probabilities and user performance time fractions are dimensionless numbers between zero and one.

Figure 4-5. Example service performance specification form.

5. ANNOTATED REFERENCES

- AGREE (1957), Reliability of military electronic equipment, Advisory Group on the Reliability of Electronic Equipment, U.S. Government Printing Office, Washington, DC. Foundation work on reliability theory.
- ANSI (1971), American National Standard, Procedures for the use of the communication control characters of American National Standard Code for Information Interchange in specified data communication links, X3.28-1971. Defines use of ASCII control characters in communications.
- ANSI (1974), American National Standard, Determination of performance of data communication systems, X3.44-1974. Original ANSI standard on data communication performance assessment, focusing on character-oriented protocols (e.g., ANSI X3.28).
- ANSI (1980), American National Standard, Determination of performance of data communication systems that use bit-oriented communication control procedures, X3.79-1980. Second ANSI standard on data communication performance assessment, focusing on bit-oriented protocols (e.g., ANSI X3.66). Expanded coverage of networks.
- Armed Services Investigating Committee (1971), Review of Department of Defense worldwide communications, Phase I, Committee on Armed Services, U. S. House of Representatives, May 10. Tragicomic history of U.S. military communication failures in three international crisis situations - the Israeli attack on the U.S.S. Liberty, the North Korean seizure of the U.S.S. Pueblo, and the North Korean shoot-down on an unarmed EC-121 reconnaissance aircraft.
- AT&T (1961), Switching Systems (American Telephone and Telegraph Company, 195 Broadway, New York, NY). Introduction to basic switching and traffic analysis.
- AT&T (1968b), Bell System Technical Reference, Model 37 teletypewriter stations for Dataphone^R service, September. Defines terminal features, options, and operating procedures.
- AT&T (1971), Bell System Technical Reference Pub. 41007, 1969-70 telecommunications network connection survey, April. Comprehensive report on transmission quality in the public switched network.
- AT&T (1978), Technical overview, Advanced Communications Service, American Telephone and Telegraph Company, Business Marketing Operations, Delivery System Strategy Group, November. Preliminary information on ACS implementation and service features.
- Boustead, C.N., and K. Metha (1974), Getting peak performance on a data channel, Data Communications magazine, July-August. Surveys factors that influence system throughput efficiency.
- Cacciamani, E.R., and K.S. Kim (1975), Circumventing the problem of propagation delay on satellite channels, Data Communications magazine, July-August. Compares stop-and-wait, continuous, and selective-repeat ARQ efficiency on satellite channels.

- CCITT (1973), Data Transmission, Green Book, Volume VIII, International Telecommunications Union, Geneva, Switzerland. Documents recommendations after the fifth Plenary. Presents standard definitions for several hundred key telecommunications terms.
- CCITT (1977), Member contributions on question 16/VII, quality of service, COM VII No's. 63-E, 101-E, 102-E, 104-E, 118-E, 133-E, 139-E, 140-E, 160-E, 169-E, 173-E, 174-E, 209-E, 210-E, 246-E, and 299-E. Contributions from various international sources on the quality of service in public data networks. Performance characteristics are proposed from both the administration (supplier) and customer (user) points of view.
- CCITT (1978), Datagram service and interface, Study Group VII contribution No. 133, COM VII No. 133-E, January. Defines datagram service quality objectives.
- CCITT (1980), Report of the Brighton meeting, Annex 3, proposed draft recommendation on elements of a network independent transport layer service, COM VII No. X3S37-80-39A, May. Exposition of network-independent transport service protocol elements and data units. User/system interaction "primitives" are defined.
- Crow, E.L. (1974), Confidence limits for digital error rates, Office of Telecommunications Report 74-51, November. Derivation of confidence limits applicable to small probabilities. Independence of successive trials assumed.
- Crow, E.L. (1978), Relations between bit and block error probabilities under Markov dependence, Office of Telecommunications, Report 78-143, March. Derivation of confidence limits for block error probability. Approximation for the probability of m errors in a block of n bits.
- Crow, E.L. (1979), Statistical methods for estimating time and rate parameters of digital communication systems, NTIA Report 79-21, June. Derivation of confidence limits and sampling procedures for the FED STD 1033 performance time, time rate, and rate efficiency parameters. Considers the effect of truncating a time distribution at three times its mean.
- Crow, E.L., and M.J. Miles (1976), A low-cost, accurate statistical method to measure bit error rates, Proceedings of the International Conference on Computer Communications, August. Practical procedure for bit error rate measurement.
- Crow, E.L., and M.J. Miles (1976), A minimum cost, accurate statistical method to measure bit error rates, Proceedings of the International Conference on Computer Communications, Toronto, Canada, August. Practical, understandable procedure for obtaining BER measurements of given precision.
- Crow, E.L., and M.J. Miles (1977), Confidence limits for digital error rates from dependent transmissions, Office of Telecommunications Report 77-118, March. Extension of Crow's 1974 report to the case of dependent errors. Two-state (Markov) error model assumed.
- DATRAN (1975), Comments of Data Transmission Company before the FCC in the matter of establishment of policies and procedures for consideration of applications to provide specialized common carrier services in the domestic public point-to-point microwave radio service and proposed amendments to parts 21, 43, and 61 of the Commission's Rules, Docket No. 18920, May 23. Proposes quality and reliability standards for "specialized" common carrier services.

- DCA (1975), System Performance Specification of AUTODIN II Phase I, Defense Communication Agency, November. States performance requirements for a "second-generation" packet switching network.
- Duffy, F.P., and R.A. Mercer (1978), A study of network performance and customer behavior during direct-distance-dialing call attempts in the U.S.A., Bell System Technical Journal, Vol. 57, January. Comprehensive measurements of DDD call attempts. Treats user delay and its influence on measured connection times.
- EPA (1980), Request for quotations, telecommunications network service, Request No. WA-80-D289/ldm, May 28. Specifies requirements for a nationwide time-sharing network to interconnect 300 and 120 bps terminals to host computers.
- FCC (1975), Further notice of inquiry and proposed rule making, specialized common carrier inquiry (Docket 18920), FCC 75-288, March 18. Defines policy issues technical questions on performance specification for "specialized" communication services.
- Feistel, H. (1973), Cryptography and computer privacy, Scientific American, Vol. 228, No. 5, May. Lucid introductory paper with practical examples.
- Feldman, N.E., W. Sollfrey, S. Katz, and S.J. Dudzinsky, Jr. (1979), Writer-to-reader delays in military communications systems, Rand Corporation Report No. R-2473-AF, October. Compares administrative and human processing delays with transmission delays for a typical military (Autodin I) message.
- Frank, H., and I.T. Frisch (1971), Communication, transmission and transportation networks (Addison-Wesley Publishing Co., Reading, MA). Rigorous mathematical presentation of network theory.
- Frank, H., and L. Hopewell (1974), Network reliability, Datamation magazine, August. Defines user-oriented reliability measures and presents case studies of network reliability enhancement. Includes quantitative cost/performance data.
- Fuchs, E., and P. Jackson (1970), Estimating distributions of random variables for certain computer communications traffic modes, Communications of the Association of Computing Machinery, Vol. 13, No. 12, December. Extends results of Jackson and Stubbs (1969).
- GAO (1977), Better management of defense communications would reduce costs, Report to the Congress by the Comptroller General of the United States, LCD-77-106, December 14. Incisive examination of current inefficiencies in military communications procurement. Excessive use of dedicated services is reported.
- Gray, J.P. (1972), Line control procedures, Proceedings of the IEEE, Vol. 60, November. Lucid presentation of protocol theory in terms of finite-state machines, with practical examples.
- Grubb, D.S., and I.W. Cotton (1975), Criteria for the performance evaluation of data communications services for computer networks, National Bureau of Standards Technical Note 882, September. Very readable survey of data communication performance parameters and issues. User perspective.

- GSA (1978), Federal Property Management Regulations, Title 41, Subchapter F, ADP and Telecommunications, Amendment F-35, November. Defines procedures for Federal agency procurement of data communications equipment and services. Identifies agency coordination and reporting responsibilities.
- GSA (1979), Interim Federal Standard 1033, Telecommunications: digital communication performance parameters, August 29. Official published version of the interim Federal Standard, available from the Office of the Manager, National Communications System Technology and Standards, Washington, DC 20305.
- Hamming, R.W. (1950), Error detecting and error correcting codes, Bell System Technical Journal, Vol. 29, April. Highly readable foundation of error control.
- ISO (1978), Reference model of open systems architecture, Version 3, ISO TC 97/SC 16 N117, November. Preliminary version of the proposed 7-layer protocol hierarchy.
- ISO (1979), Reference model of open systems interconnection, ISO/TC 97/SC 16 N 227, Version 4, August. First essentially complete exposition of the proposed 7-layer protocol hierarchy.
- Jackson, P., and C. Stubbs (1969), A Study of Multi-Access Computer Communications, Proceedings of the Spring Joint Computer Conference, May 14-16. Builds a "data stream model" from measurements of user/program interactions in a timesharing environment.
- Kelley, K.G. (1977), An evaluation of data transfer requirements for the future DCS, DCEC Technical Note 24-77, November. Projects future DCS traffic volumes and "response time" requirements.
- Kimmet, F.G., and N.B. Seitz (1978), Digital communication performance parameters for Federal Standard 1033, NTIA Report 78-4, Vol. II, application examples, May. Develops FED STD 1033 performance parameter values for three representative data communication services: nonswitched, circuit-switched, and message-switched.
- Kirk, K.W., and J.L. Osterholz (1976), DCS digital transmission performance, DCEC Technical Report 12-76, November. Defines end-to-end transmission performance objectives and relates them to transmission subsystems in a global reference network.
- Kleinrock, L. (1976), Queueing systems, Volume II, Computer applications (John Wiley & Sons, Inc., New York, NY). Outstanding, definitive text on design and analysis of computer communication networks. Much useful system performance information. Emphasis on packet switching.
- Kleinrock, L., and W.E. Naylor (1974), On measured behavior of the ARPA network, Proceedings of the National Computer Conference, May 6-10. Presents measurement results and analytic models derived from a week-long observation of ARPANET traffic.
- Kleinrock, L., W.F. Naylor, and H. Opderbeck (1976), A study of line overhead in the ARPA network, Communications of the ACM, Vol. 10, No. 1, January. Definitive analytical and experimental results on potential and actual ARPANET efficiency.

- Kobylar, A.W., and H. A. Malec (1973), System effectiveness trade-off in a space-time-space network for a digital exchange, GTE Automatic Electric Technical Journal, July. Defines reliability measures and presents simulation results for a typical PCM switch.
- Kuhn, T.G. (1963), Retransmission error control, IEEE Transactions on Communications Systems, Vol. CS-11, No. 2, June. Practical study with emphasis on block size optimization. Simple block parity error detection schemes are illustrated.
- Lemp, J. (1980), Telecommunications privacy and protection, Telecommunications Journal, Vol. 47, June. Describes Federal government concerns with passive eavesdropping on microwave radio transmissions and outlines protection alternatives.
- Linfield, R.F., and M. Nesenbergs (1978), Access area switching and signaling concepts, issues, and alternatives, NTIA Report 78-2, May. PABX and signaling alternatives for future Army access area communications systems are discussed.
- Lucky, R.W. (1979), Gamenet, IEEE Communications Magazine, Vol. 17, No. 6, November. Fanciful, futuristic view of tomorrow's electronic games.
- MacRae, D.D., F.A. Perkins, D.J. Risavy, and J.N. York (1976), 16 Kb/s Data modem techniques, RADC-TR-76-311, Harris Corp., October. Study of digital speech transmission quality.
- Malec, H.A. (1975), Telephone switching system reliability - past, present, and future. Proceedings of the National Telecommunications Conference, December 1-3. Surveys reliability and effectiveness measures for commercial telephone switching systems.
- Martin, J. (1976), Telecommunications and the computer, Second Edition, (Prentice-Hall, Inc., Englewood Cliffs, NJ). Comprehensive, readable book covering basic technology, administration, and applications of telecommunications.
- McFadyen, H.J. (1976), Systems network architecture: an overview, IBM systems Journal, Vol. 15, No. 1. Introduction to a special issue on SNA.
- McManamon, P.M., R.K. Rosich, J.A. Payne, and M.J. Miles (1975), Performance criteria for digital data networks, Office of Telecommunications Report No. 75-54, January. Early survey of user-oriented and engineering performance criteria and their relationships.
- Miller, R.B. (1968), Response time in man-computer conversational transactions, Proceedings of the Fall Joint Computer Conference, December 9-11. Comprehensive study of response time requirements for various human/system interactions.
- National Security Council (1979), National security telecommunications policy, Presidential Directive NCS-53, November 15. Sets out policy direction for development of the National Communications System.
- NCS (1977), Proposed Federal Standard 1033, Telecommunication: digital communication performance parameters, Federal Register, Vol. 43, No. 35, February 22. Initial publication of the standard for public comment.

- NCS (1978), Proposed Federal Standard 1033, Federal Register, Vol. 43, No. 180, September 15. Brief announcement of the revised standard's availability for final public comment.
- Nesenbergs, M. (1975), Study of error control coding for the U.S. Postal Service Electronic Message System, Office of Telecommunications, Institute for Telecommunication Sciences, U.S. Department of Commerce, Boulder, CO 80303, May. Surveys candidate error control techniques for a high data rate satellite network.
- Nesenbergs, M., W.J. Hartman, and R.F. Linfield (1980), Performance parameters for digital and analog service modes, NTIA Report 80- , November. Analysis of user and system requirements for digital and analog service of the future Defense Communications System.
- Newman, W.M. (1979), Principles of interactive computer graphics, 2nd edition (McGraw-Hill Publishing Company, New York, NY). Comprehensive text with extensive treatment of user interface design.
- NTIA (1980), Manual of Regulations and Procedures for Federal Radio Frequency Management, January. Voluminous rules governing allocation and use of Federal RF spectrum.
- NRC (1977), National Research Council Committee on Telecommunications, Summary of Office of Telecommunications Study Panel Meeting, Boulder, CO, April 11-13. Telecommunications experts from industry and academia assess ITS program priorities.
- Office of Management and Budget (1979), Circular No. A-76, Policies for Acquiring Commercial or Industrial Products and Services Needed by the Government, March 29. Provide guidelines for Federal agency procurement, emphasizing reliance on the private sector.
- Payne, J.A. (1978), ARPANET host-to-host access and disengagement measurements, NTIA Report 78-3, May. Describes access and disengagement performance measurements for ARPANET connections established via the Telnet protocol.
- Popek, G.J. (1974), Protection structures, Computer magazine, June. Examines computer security concerns and protection strategies.
- Roberts, L.F., and B.D. Wessler (1970), Computer network development to achieve resource sharing, Proceedings of the Spring Joint Computer Conference, 36, May 5-7. Early overview of the ARPA network.
- Rose, M.P., and J.P. O'Keefe (1980), User studies value-added network response times, Data Communications, 9, No. 4, April. Compares block transfer times and call blocking probabilities of Telnet and Tymnet.
- SBS (1978), Communications service interfaces available at the satellite communications controller, SBS 3201-004, January. SBS digital interface description.
- Schwartz, M., R. Boorstyn, and R. Pickholtz (1972), Terminal-oriented computer communications network, Proceedings of the IEEE, Vol. 60, No. 11, November. Tutorial presentation of four computer communication networks.

- Seitz, N.B. (1980), Measuring communication availability with Federal Standard 1033, Proceedings of the 1980 Reliability and Maintainability Symposium, January. The FED STD 1033 definition of "outage" is discussed.
- Seitz, N.B., and D. Bodson (1980), Data communication performance assessment, Telecommunications magazine, February. Brief, informal summary of the purpose, scope, and intended application of FED STD 1033.
- Seitz, N.B., and P.M. McManamon (1976), Review of responses to the FCC quality and reliability inquiry (Docket 18920), Office of Telecommunications Report 76-101, August. Summarizes industry responses the FCC (1975) inquiry and presents OT studies on user-oriented data communication performance assessment.
- Seitz, N.B., and P.M. McManamon (1978), Digital communication performance parameters for proposed Federal Standard 1033, NTIA Report 78-4, Volume I, standard parameters, May. Comprehensive presentation of the standard's technical basis.
- Shannon, C.E. (1948), A mathematical theory of communication, Bell System Technical Journal, Vol. 27, July. Profoundly significant, elegantly presented foundation of communication and information theory.
- Sloan, L.J. (1979), Limiting the lifetime of packets in computer networks, 4th conference of local computer networks, IEEE Computer Society, October 22-23. Proposes a procedure for detecting and discarding packets that encounter excessive transfer delays.
- Sunshine, C.A. (1975), Interprocess communication protocols for computer networks, Digital Systems Laboratory, Department of Electrical Engineering, Stanford University, Technical Report No. 105, December. Comprehensive, readable thesis on communication protocol analysis and design. End-to-end point of view.
- Utlaut, W.F. (1978), Spread spectrum: principles and possible application to spectrum utilization and allocation, IEEE Communications Society magazine, Vol. 16, No. 5, September. Tutorial summary of spread-spectrum principles and applications.
- Xerox (1978), Xerox Corporation petition for rule making, in re Amendment of Parts 2, 21, 87, 89, and 91 of the rules for the establishment of new common carrier Electronic Message Service (EMS) in the band 10.55-10.68 GHz, before the Federal Communications Commission, November 16. Summary of the proposed XTEN network.

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO. NTIA Report 80-55		2. Gov't Accession No.	3. Recipient's Accession No.
4. TITLE AND SUBTITLE INTERIM FEDERAL STANDARD 1033 REFERENCE MANUAL		5. Publication Date December 1980	
7. AUTHOR(S) Neal B. Seitz		6. Performing Organization Code NTIA/ITS	
8. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences, 325 Broadway, Boulder, CO 80303		9. Project/Task/Work Unit No. 910 4120	
11. Sponsoring Organization Name and Address U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences, 325 Broadway, Boulder, CO 80303		10. Contract/Grant No.	
14. SUPPLEMENTARY NOTES		12. Type of Report and Period Covered NTIA Report	
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>There is a growing need within the Federal government for a user-oriented, system-independent, <u>functional</u> means of specifying data communication performance. A recently published Federal Standard, Interim Federal Standard 1033, defines a set of standard performance parameters designed to meet that need. This report is basically an explanation and elaboration of that standard. The report first outlines the need for the standard, and the potential benefits of its use, from the viewpoint of the end user, the communication supplier, and the communication manager. The report then summarizes the objectives and content of the standard in informal, nontechnical terms. Finally, the report examines the meaning and importance of each standard parameter in a series of tutorial parameter "essays." Typical parameter values are presented, and design implications are discussed.</p>		13.	
16. Key Words (Alphabetical order, separated by semicolons)			
17. AVAILABILITY STATEMENT <input checked="" type="checkbox"/> UNLIMITED. <input type="checkbox"/> FOR OFFICIAL DISTRIBUTION.		18. Security Class. (This report) Unclassified	20. Number of pages 125
		19. Security Class. (This page) Unclassified	21. Price:

