An Overview of Signaling System No. 7

ABDI R. MODARRESSI, MEMBER, IEEE, AND RONALD A. SKOOG, MEMBER, IEEE

Invited Paper

In modern telecommunication networks, signaling constitutes the distinct control infrastructure that enables provision of ALL other services. The component of signaling systems that controls provision of services between the user and the network is the access signaling component, and the component that controls provision of services within the network, or between networks, is the network signaling component. There are international standards for both access signaling and network signaling protocols. From a network structure viewpoint, access signaling structures generally provide point-to-point connectivity between the user and a network node, while network signaling structures provide network-wide communication capability (directly or indirectly) between the nodes of the public network(s). Since the network signaling system acts as a traffic collector/distributor for many access signaling tributaries, its functions are more complex, its structure more involved, and its performance more stringent. This paper provides an overview of modern network signaling systems based on the Signaling System No. 7 international standard.

I. INTRODUCTION

In the context of modern telecommunications, signaling can be defined as the *system* that enables stored program control exchanges, network databases, and other 'intelligent' nodes of the network to exchange a) messages related to call setup, supervision, and tear-down (call/connection control); b) information needed for distributed application processing (inter-process query/response, or user-to-user data); and c) network management information. As such, signaling constitutes the control infrastructure of the modern telecommunication network.

Modern signaling systems are essentially data communication systems using layered protocols. What distinguishes them from other data communication systems are basically two things: their real time performance and their reliability requirements. No matter how complex the set of network interactions are for setting up a call, the call setup time should still not exceed a couple of seconds. This imposes quite a stringent end-to-end delay requirement on the signaling system. On the other hand, because of the absolute reliance of the telecommunication network on its signaling system, requirements for signaling network reliability (message integrity, end-to-end availability, network robustness, recovery from failure, etc.) are extremely demanding. For example, current objectives require the down-time between any arbitrary pair of communicating nodes in the signaling network not to exceed 10 min/year. This is at least two orders of magnitude smaller than the corresponding requirement in a general-purpose data network. Requirements on real-time performance and reliability of signaling systems are likely to become even more stringent with advances in technology and new application needs.

Over the last century or so, signaling has evolved with the technology of telephony, although the pace of this evolution has never been faster than in the last two decades, a period characterized by the marriage of computer and switching technologies. The advent of the Integrated Services Digital Network (ISDN) has further accelerated the pace of development and deployment of signaling systems to support an ever increasing set of "intelligent network" services on a worldwide basis. When viewed as an end-to-end capability, signaling in ISDN has two distinct components: signaling between the user and the network (access signaling), and signaling within the network (network signaling). The current set of protocol standards for *access signaling* is known as the Digital Subscriber Signaling System No. 1 (DSS1). The current set of protocol standards for *network signaling* is known as the Signaling System No. 7 (SS7).

This paper provides an overview of Signaling System No. 7. It is a somewhat abridged and updated version of a tutorial on SS7 that was published in 1990 [1]. Following this introduction, the salient features of SS7's Network Services Part (NSP) are described in Section II. Functionally, NSP corresponds to the first three layers of the Open System Interconnection (OSI) Reference Model. This section also provides a discussion of signaling network structures that, in conjunction with the NSP, provide ISDN nodes with a highly reliable and efficient means of exchanging signaling messages. Once this reliable signaling message transport capability is realized, each network node has to be equipped with capabilities for processing of the transported messages in support of a useful function like setting up of a call (connection). In an increasingly large number of cases, call setup has to be preceded by invocation of some distributed

Manuscript received October 23, 1991; revised December 18, 1991.

A. R. Modarressi is with AT&T Bell Laboratories, Columbus, OH 43213.

R. A. Skoog is with AT&T Bell Laboratories, Holmdel, NJ 07733.

IEEE Log Number 9108075.

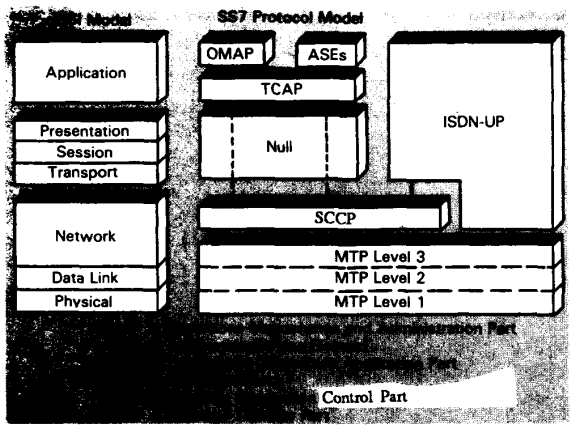


Fig. 1. SS7 protocol architecture.

application processes, the outcome of which determines the nature as well as the attributes of the subsequent call or connection control process. These nodal capabilities of call control and remote process invocation and management are part of the Signaling System No. 7 User Parts, which are described in Section III. In Section IV, we dwell on the very stringent performance requirements of signaling systems. These requirements reflect the critical nature of signaling functions and their real time exigencies. Finally, in Section V we sketch a broad outline of the likely evolution of network signaling in the remaining years of this century.

II. Signaling System No. 7 Network Services Part (NSP)

In this section, we describe the Signaling System No. 7 protocols that correspond to the first three layers (Physical, Data Link, and Network) of the OSI Reference Model. This component of the Signaling System No.7 protocol is called the Network Services Part (NSP), and it consists of the Message Transfer Part (MTP) and the Signaling Connection Control Part (SCCP). Figure 1 shows how these relate to each other and to the other components of the protocol. MTP consists of levels 1–3 of the Signaling System No. 7 protocol, which are called the Signaling Data Link, the Signaling Link, and the Signaling Network functions, respectively. SCCP is an MTP user, and therefore is in level 4 of Signaling System No. 7 protocol stack. MTP provides a connectionless message transfer system that enables signaling information to be transferred across the network to its desired destination. Functions are included in MTP that allow system failures to occur in the network without adversely affecting the transfer of signaling information. SCCP provides additional functions to MTP for both connectionless and connection-oriented network services.

MTP was developed before SCCP and it was tailored to the real time needs of telephony applications. Thus a connectionless (datagram) capability was called for which avoids the administration and overhead of virtual circuit networks (one of the disadvantages of CCS6). Later, it became clear that there were other applications that would need additional network services (full OSI Network service capabilities) like an expanded addressing capability and connection-oriented message transfer. SCCP was developed to satisfy this need. The resulting structure, and specifically the splitting of the OSI Network functions into MTP level 3 and SCCP, has certain advantages in the sense that the higher overhead SCCP services can be used only when needed, allowing the more efficient MTP to serve the needs of those applications that can use a connectionless message transfer with limited addressing capability.

Sections II-A and II-B provide an overview of MTP and SCCP, respectively. Section II-C describes the signaling network structures that can be used to implement the Network Services Part.

A. The Message Transfer Part (MTP)

The overall purpose of MTP is to provide a reliable transfer and delivery of signaling information across the signaling network, and to react and take necessary actions in response to system and network failures to ensure that reliable transfer is maintained. Figure 2 illustrates the functions of MTP levels, and their relationship to one another and to the MTP users. These three levels are now described.

1) Signaling Data Link Functions (Level 1): A *Signaling Data Link* is a bidirectional transmission path for signaling, consisting of two data channels operating together in opposite directions at the same data rate. It fully complies with the OSI's definition of the physical layer (layer 1). Transmission channels can be either digital or analog, terrestrial or satellite.

For digital signaling data links, the recommended bit rate for the ANSI standard is 56 kb/s, and for the CCITT International Standard it is 64 kb/s. Lower bit rates may be used, but the message delay requirements of the User Parts must be taken into consideration. The minimum bit rate allowed for telephone call control applications is 4.8 kb/s. In the future, bit rates higher than 64 kb/s may be required (e.g., 1.544 Mb/s in North America and 2.048 Mb/s elsewhere), but further study is needed before these rates can be standardized.

2) Signaling Link Functions (Level 2): The Signaling Link functions correspond to the OSI's data link layer (layer 2). Together with a signaling data link, the signaling link functions provide a *signaling link* for the reliable transfer of signaling messages between two directly connected signaling points. Signaling messages are transferred over the signaling link in variable length messages called *signal units*. There are three types of signal units, differentiated by the length indicator field contained in each, and their formats are shown in Fig. 3. The Signaling Information Field (SIF) in a Message Signal Unit (MSU) must have a length less than or equal to 272 octets. This limitation is imposed to control the delay a message can impose on other messages due to its emission time (which is limited by the maximum standardized link speed of 64 kb/s).

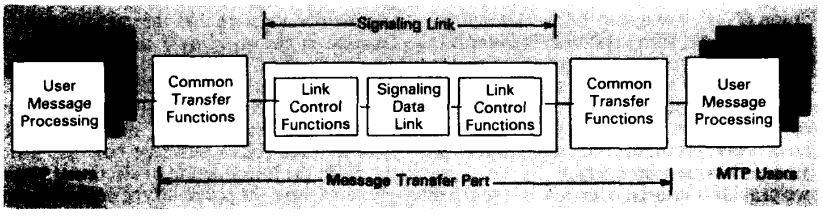


Fig. 2. MTP functional diagram.

The SS7 link functions show a strong similarity to typical data network bit-oriented link protocols (e.g., HDLC, SDLC, LAP-B), but there are some important differences. These differences arise from the performance needs of signaling (e.g., lost messages, excessive delays, out-of-sequence messages) that require the network to respond quickly to system or component failure events. The standard flag (01111110) is used to open and close signal units, and the standard CCITT 16-bit CRC checksum is used for error detection. However, when there is no message traffic, Fill-In Signal Units (FISU's) are sent rather than flags, as is done in other data link protocols. The reason for this is to allow for a consistent error monitoring method (described below) so that faulty links can be quickly detected and removed from service even when traffic is low.

a) Error correction: Two forms of error correction are specified in the signaling link procedures. They are the *Basic Method* and the *Preventive Cyclic Retransmission (PCR) Method.* In both methods only errored MSU's and Link Status Signal Units (LSSU's) are corrected, while errors in FISU's are detected but not corrected. Both methods are also designed to avoid out-of-sequence and duplicated messages when error correction takes place. The PCR method is used when the propagation delay is large (e.g., with satellite transmission).

The Basic Method of error correction is a non-compelled positive/negative acknowledgment retransmission error correction system. It uses the "go-back-N" technique of retransmission used in many other protocols. If a negative acknowledgment is received, the transmitting terminal stops sending new MSU's, rolls back to the MSU received in error, and retransmits everything from that point before resuming transmission of new MSU's. Positive acknowledgments are used to indicate correct reception of MSU's, and as an indication that the positively acknowledged buffered MSU's can be discarded at the transmitting end. For sequence control, each signal unit is assigned forward and backward sequence numbers and forward and backward indicator bits (see Fig. 3). The sequence numbers are seven bits long, which means at most 127 messages can be transmitted without receiving a positive acknowledgment.

The PCR method is a non-compelled positive acknowledgment cyclic retransmission, forward error correction system. A copy of a transmitted MSU is retained at the transmitting terminal until a positive acknowledgment for that MSU is received. When there are no new MSU's to be sent, all MSU's not positively acknowledged are retransmitted cyclically. When the number of unacknowledged MSU's (either the number of messages or the number of octets) exceeds certain thresholds, it is an indication that error correction is not getting done by cyclic retransmission. This would occur, for example, if the traffic level was high, which causes the retransmission rate to be low. In this situation a *forced retransmission* procedure is invoked. In this procedure new MSU transmission is stopped and all unacknowledged MSU's are retransmitted. This forced retransmission continues until the unacknowledged message and octet counts are below specified threshold values. These threshold values must be chosen carefully, for if they are set too low, and the link utilization is large enough, the link will become unstable (i.e., once a forced retransmission starts, the link continues to cycle in and out of forced retransmission [2]).

b) Error monitoring: Two types of signaling link error rate monitoring are provided. A *signal unit error rate monitor* is used while a signaling link is in service, and it provides the criteria for taking a signaling link out of service due to an excessively high error rate. An *alignment error rate monitor* is used while a signaling link is in the proving state of the initial alignment procedure, and it provides the criteria for rejecting a signaling link for service during the initial alignment due to too high an error rate.

The signal unit error rate monitor is based on a signal unit (including FISU) error count, incremented and decremented using the "leaky bucket" algorithm. For each errored signal unit the count is increased by one, and for each 256 signal units received (errored or not), a positive count is decremented by one (a zero count is left at zero). When the count reaches 64, an excessive error rate indication is sent to level 3, and the signaling link is put in the out of service state. When loss of alignment occurs (a loss of alignment occurs when more than six consecutive 1s are received or the maximum length of a signal unit is exceeded), the error rate monitor changes to an octet counting mode. In this mode it increments the counter for every 16 octets received. Octet counting is stopped when the first correctly-checking signal unit is detected.

The alignment error rate monitor is a linear counter that is operated during alignment proving periods. The counter is started at zero at the start of a proving period, and the count is incremented by one for each signal unit received in error (or for each 16 octets received if in the octet counting mode). A proving period is aborted if the threshold for the

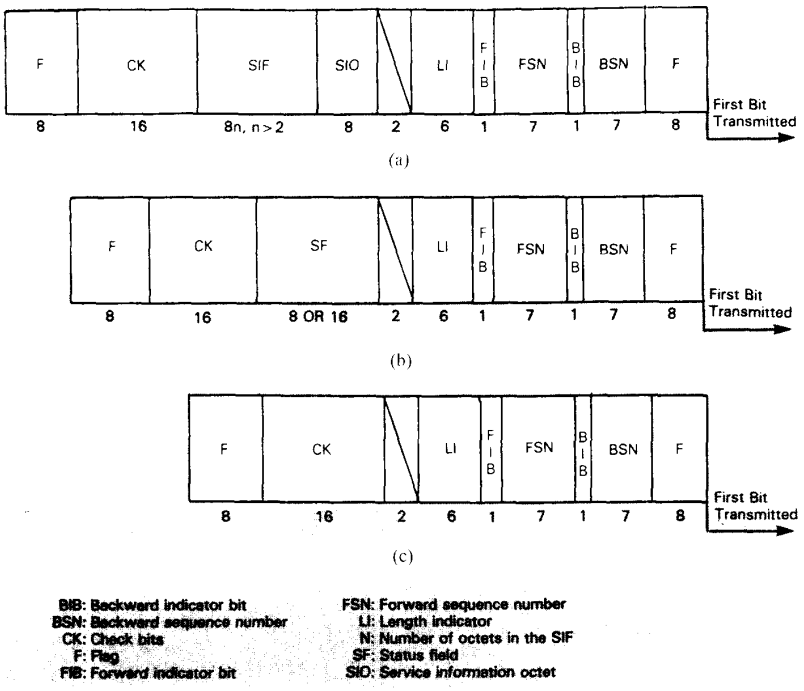


Fig. 3. Signal unit formats.

alignment error rate monitor count is exceeded before the proving period timer expires.

c) Flow control: The flow control procedure is initiated when congestion is detected at the receiving end of the signaling link. The congested receiving end notifies the transmitting end of its congestion with a link status signal unit (LSSU) indicating busy, and withholds acknowledgment of all incoming signal units. This action stops the transmitting end from failing the link due to a time-out on acknowledgment. However, if the congestion condition lasts too long (3–6 s), the transmitting end will fail the link.

A processor outage condition indication is sent by level 2, called signaling indication processor outage (SIPO), whenever an explicit indication is sent to level 2 from level 3 or when level 2 recognizes a failure of level 3. This indicates to the far end that signaling messages cannot be transferred to level 3 or above. The far-end level 2 responds by sending fill-in signal units and informing its level 3 of the SIPO condition. The far-end level 3 will reroute traffic in accordance with the signaling network management procedures described as follows.

3) Signaling Network Functions (Level 3): The signaling network functions correspond to the lower half of the OSI's Network layer, and they provide the functions and procedures for the transfer of messages between signaling points, which are the nodes of the signaling network. The signaling network functions can be divided into two basic categories: *signaling message handling* and *signaling network management.* The breakdown of these functions and their interrelationship is illustrated in Fig. 4.

a) Signaling message handling: Signaling message handling consists of message routing, discrimination, and distribution functions. These functions are performed at each signaling point in a signaling network, and they are based on the part of the message called the *routing label,* and the Service Information Octet (SIO) shown in Fig. 3. The routing label is illustrated in Fig. 5 and consists of the Destination Point Code (DPC), the Origination Point Code (OPC), and the Signaling Link Selection (SLS) field. In the international standard the DPC and OPC are 14 bits each, while the SLS field is 4 bits long. For ANSI, the OPC and DPC are each 24 bits (to accommodate larger networks), while the SLS field has 5 bits, and there are 3 spare bits in the routing label. The routing label is placed at the beginning of the Signaling Information Field, and it is the common part of the label that is defined for each MTP user.

When a message comes from a level 3 user, or originates at level 3, the choice of the particular signaling link on which it is to be sent is made by the message routing function. When a message is received from level 2, the discrimination function is activated, and it determines if it is addressed to another signaling point or to itself based on the DPC in the message. If the received message is addressed to another signaling point, and the receiving signaling point has the transfer capability, i.e., the Signal Transfer Point (STP) function, the message is sent to the message routing function. If the received message is addressed to the

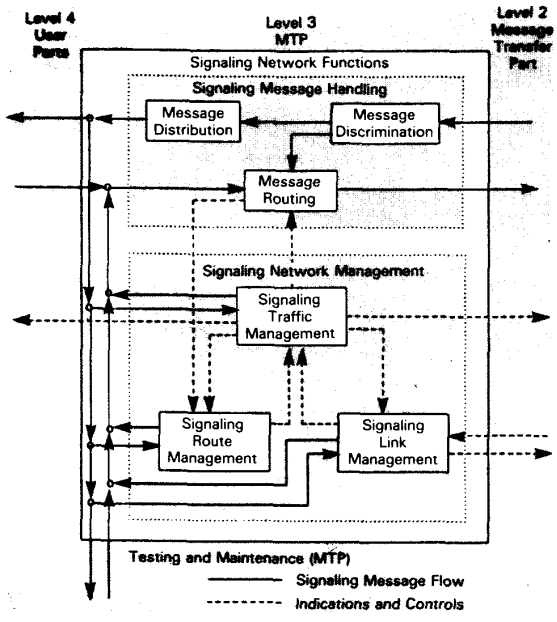


Fig. 4. Signaling network functions.

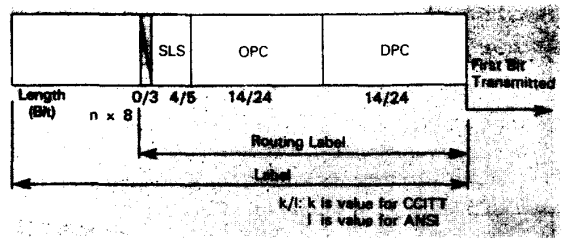


Fig. 5. Routing label structure.

receiving signaling point, the message distribution function is activated, and it delivers the message to the appropriate MTP user or MTP level 3 function based on the service indicator, a sub-field of the SIO field. Message routing is based on the DPC and the SLS in almost all cases. In some circumstances the SIO, or parts of it (the service indicator and network indicator), may need to be used.

Generally, more than one signaling link can be used to route a message to a particular DPC. The selection of the particular link to use is made using the SLS field. This is called load sharing. A set of links between two signaling points is called a *link set*, and load sharing can be done over links in the same link set or over links not belonging to the same link set. A load sharing collection of one or more link sets is called a *combined link set.*

The objective of load sharing is to keep the load as evenly balanced as possible on the signaling links within a combined link set. For messages that should be kept in sequence, the same SLS code is used so that such messages take the same path. For example, for trunk signaling with ISUP (see Section IV-A) the same SLS code is used for all messages related to a particular trunk. In order to ensure proper load balance using SLS fields, it is critical that the SLS codes are assigned such that the load is shared evenly across all the SLS codes. Even then, the SLS load sharing method does not provide a fully balanced loading of signaling links in all cases. For example, if there are six signaling links in a combined link set, the 16 SLS codes would be assigned so that four signaling links would each carry three SLS codes and two of the signaling links would each carry only two SLS codes.

b) Signaling network management: The purpose of the signaling network management functions is to provide reconfiguration of the signaling network in the case of signaling link or signaling point failures, and to control traffic in the case of congestion or blockage. The objective is that, when a failure occurs, the reconfigurations be carried out so messages are not lost, duplicated, or put out of sequence, and that message delays do not become excessive. As shown in Fig. 4, signaling network management consists of three functions: signaling traffic management, signaling route management, and signaling link management. Whenever a change in the status of a signaling link, signaling route or signaling point occurs, these three functions are activated as summarized below.

The *signaling traffic management* procedures are used to divert signaling traffic, without causing message loss, missequencing, or duplication, from unavailable signaling links or routes to one or more alternative signaling links or routes, and to reduce traffic in the case of congestion. When a signaling link becomes unavailable, a *changeover* procedure is used to divert signaling traffic to one or more alternative signaling links, as well as to retrieve for retransmission messages that have not been positively acknowledged. When a signaling link becomes available, a *changeback* procedure is used to reestablish signaling traffic on the signaling link made available. When signaling routes (succession of links from the origination to the destination signaling point) become unavailable or available, *forced rerouting* and *controlled rerouting* procedures are used, respectively, to divert the traffic to alternative routes or to the route made available. Controlled rerouting is also used to divert traffic to an alternate (more efficient) route when the original route becomes restricted (i.e., less efficient because of additional transfer points in the path). When a signaling point becomes available after having been down for some time, the *signaling point restart* procedure is used to update the network routing status and control when signaling traffic is diverted to (or through) the point made available.

The *signaling route management* procedures are used to distribute information about the signaling network status in order to block or unblock signaling routes. The following procedures are defined to take care of different situations. The *transfer-controlled* procedure is performed at a signaling transfer point in the case of signaling link congestion. In this procedure, for every message received having a congestion priority less than the congestion level of the signaling link, a control message is sent to the

OPC of the message asking it to stop sending traffic that has a congestion priority less than the congestion level of the signaling link to the DPC of the message. In ANSI Standards four congestion message priorities are used; in international networks only one is used. The *transfer-prohibited* procedure is performed at a Signal Transfer Point to inform adjacent signaling points that they must no longer route to a DPC via that STP. This procedure would be invoked, for example, if the STP had no available routes to a particular destination. The *transfer-restricted* procedure is performed at a Signal Transfer Point to inform adjacent signaling points that, if possible, they should no longer route messages to a DPC via that STP. The *transfer-allowed* procedure is used to inform adjacent signaling points that routing to a DPC through that STP is now normal. In the ANSI standards, the above procedures are also specified on a cluster basis (a cluster being a collection of signaling points), which significantly reduces the number of network management messages and related processing required when there is a cluster failure or recovery event. The *signaling-route-set-test* procedure is used by the signaling points receiving transfer prohibited and transfer restricted messages in order to recover the signaling route availability information that may not have been received due to some failure. Finally, in ANSI standards the *signaling-route-set-congestion-test* procedure is used to update the congestion status associated with a route toward a particular destination.

The *signaling link management* function is used to restore failed signaling links, to activate new signaling links, and to deactivate aligned signaling links. There is a basic set of signaling link management procedures, and this set of procedures are provided for any international or national signaling system. Two optional sets of signaling link management procedures are also provided, which allow for a more efficient use of signaling equipment when signaling terminal devices have switched access to signaling data links. The basic set of procedures are *signaling link activation* (used for signaling links that have never been put into service, or that have been taken out of service), *signaling link restoration* (used for active signaling links that have failed), *signaling link deactivation*, and *signaling link set activation*. The optional sets of procedures address automatic allocation of signaling terminals, and automatic allocation of data links and signaling terminals.

B. The Signaling Connection Control Part (SCCP)

SCCP enhances the services of the MTP to provide the functional equivalent of OSI's Network layer (layer 3). The addressing capability of MTP is limited to delivering a message to a node and using a four bit service indicator (a sub-field of the SIO) to distribute messages within the node. SCCP supplements this capability by providing an addressing capability that uses DPC's plus Subsystem Numbers (SSN's). The SSN is local addressing information used by SCCP to identify each of the SCCP users at a node. Another addressing enhancement to MTP provided by SCCP is the ability to address messages with global titles, addresses (such as dialed 800 or free phone numbers) that are not directly usable for routing by MTP. For global titles a translation capability is required in SCCP to translate the global title to a DPC + SSN. This translation function can be performed at the originating point of the message, or at another signaling point in the network (e.g., at an STP).

In addition to enhanced addressing capability, SCCP provides four classes of service, two connectionless and two connection-oriented. The four classes are:

- Class 0: Basic connectionless class;
- Class 1: Sequenced (MTP) connectionless class;
- Class 2: Basic connection-oriented class;
- Class 3: Flow control connection-oriented class.

In Class 0 service, a user-to-user information block, called a Network Service Data Unit (NSDU), is passed by higher layers to SCCP in the node of origin; it is transported to the SCCP function in the destination node in the user field of a *Unitdata* message; at the destination node it is delivered by SCCP to higher layers. The NSDU's are transported independently and may be delivered out of sequence, so this class of service is purely connection-less.

In Class 1, the features of Class 0 are provided with an additional feature that allows the higher layer to indicate to SCCP that a particular stream of NSDU's should be delivered in sequence. SCCP does this by associating the stream members with a sequence control parameter and giving all messages in the stream the same SLS code.

In Class 2, a bidirectional transfer of NSDU's is performed by setting up a temporary or permanent signaling connection (a virtual channel through the signaling network). Messages belonging to the same signaling connection are given the same SLS code to ensure sequencing. In addition, this service class provides a segmentation and reassembly capability. With this capability, if an NSDU is longer than 255 octets, it is split into multiple segments at the originating node, each segment is transported to the destination node in the user field of a *Data* message, and at the destination node SCCP reassembles the original NSDU.

In Class 3, the capabilities of Class 2 are provided with the addition of flow control. Also the detection of message loss and missequencing is provided. In the event of lost or missequenced messages, the signaling connection is reset and notification is given to the higher layers.

The structure of SCCP is illustrated in Fig. 6, and consists of four functional blocks. The SCCP connection-oriented control block controls the establishment and release of signaling connections and provides for data transfer on signaling connections. The SCCP connectionless control block provides for the connectionless transfer of data units. The SCCP management block provides capabilities beyond those of MTP to handle the congestion or failure of either the SCCP user or the signaling route to the SCCP user. With this capability, SCCP can route messages to backup systems in the event failures prevent routing to the primary system. The SCCP routing block takes messages received from MTP or other SCCP functional blocks and performs the

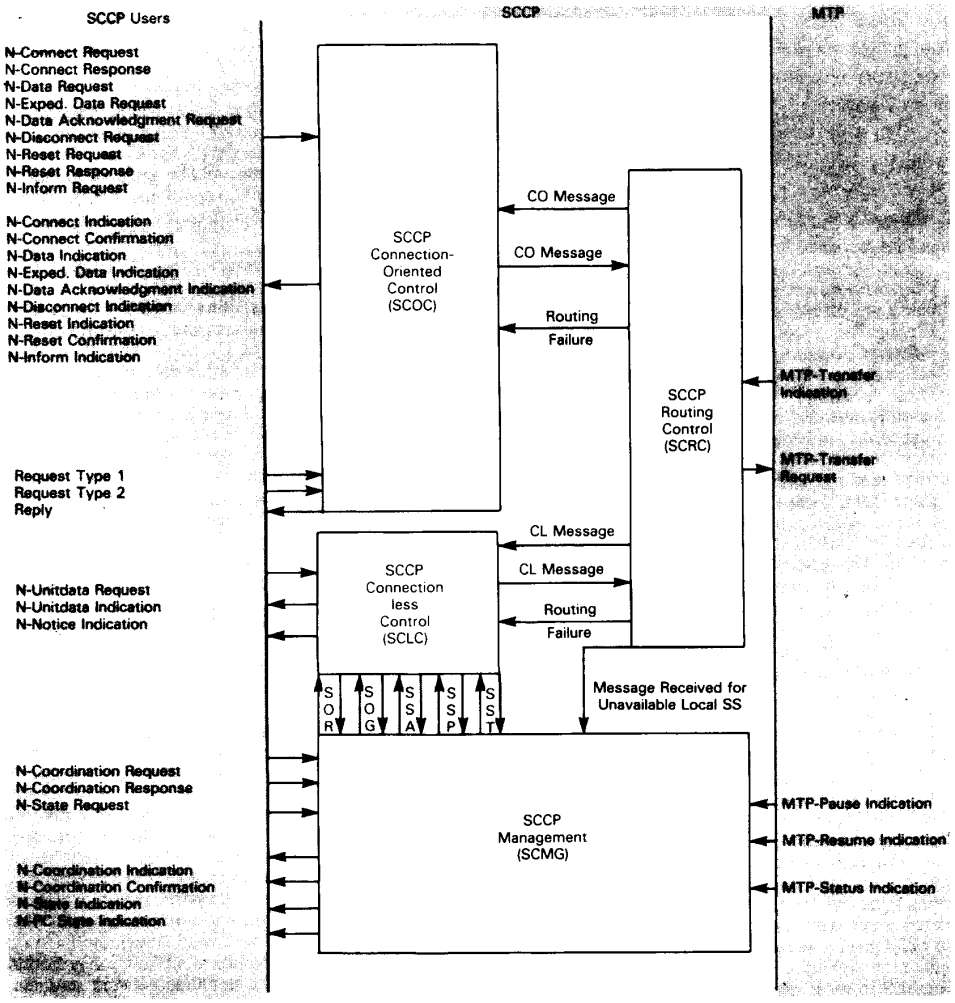


Fig. 6. SCCP functions.

necessary routing functions to either forward the message to MTP for transfer or pass the message to other SCCP functional blocks.

C. Signaling Network Structures

Signaling networks consist of signaling points and signaling links connecting the signaling points together. As alluded to earlier, a signaling point that transfers messages from one signaling link to another at level 3 is said to be a Signal Transfer Point (STP). Signaling points that are STP's can also provide functions higher than level 3, such as SCCP and other level 4 functions like ISUP (see Section IV-A). When a signaling point has an STP capability and also provides level 4 functions like ISUP, it is commonly said to have an *integrated* STP functionality. When the signaling point provides only STP capability, or STP and SCCP capabilities, it is commonly called a *standalone* STP. Signaling links, STP's (stand alone and integrated), and signaling points with level 4 protocol functionality can be configured in many different ways to form a signaling network. The Signaling System No. 7 protocol is specified independent of the underlying signaling network structure. However, to meet the stringent availability requirements given in Section V-A (e.g., signaling relation unavailability not to exceed 10 min/year), it is clear that any network structure must provide redundancies for the signaling links, which by themselves have unavailabilities measured in many hours per year. In most cases the STP's must also have backups.

The worldwide signaling network is intended to be structured into two functionally independent levels: the national and the international levels. This allows numbering plans and network management of the international and the different national networks to be independent of one another. A signaling point can be either a national signaling point, an international signaling point, or both. If it serves as both, it is identified by a specific signaling point code in each of the signaling networks.

Administrations and Exchange Carriers can form agreements to interconnect their signaling networks, as is currently being done in North America [3], as well as internationally [4]. When this is done, it is desirable for security reasons to place restrictions on the signaling messages authorized to go from one network to another. To ensure these restrictions are complied with, screening procedures should be provided at the network interconnection points (gateways).

1) Types of Signaling Network Structures: In the Signaling System No. 7 terminology, when two nodes are capable of exchanging signaling messages between themselves through the signaling network, a *signaling relation* is said to exist between them. Signaling networks can use three different signaling *modes*, where mode refers to the association between the path taken by the signaling message and its corresponding signaling relation. In the *associated* mode of signaling, the messages corresponding to a signaling relation between two points are conveyed over a link set directly interconnecting those two signaling points. In the *non-associated* mode of signaling, a message corresponding to a signaling relation between two points is conveyed over two or more link sets in tandem passing through one or more signaling points other than the origin and the destination of the message. The *quasi-associated* mode of signaling is a non-associated signaling mode where the path taken by the message through the signaling network is predetermined and fixed, except for the rerouting caused by failure and recovery events. Signaling System No. 7 is specified for use with all modes of signaling.

A familiar signaling network structure is the *mesh* structure illustrated in Fig. 7(a). This is also known as the *quad* structure. The STP's in this structure are "mated" on a pairwise basis. This is the type of structure used in North America [3].

The mesh network has 100% redundancy; that is, for any single point of failure, the traffic can be diverted to alternate paths that do not increase the number of transfer points. The network must be engineered so that each component under normal conditions can handle twice its peak load. Also, facility diversity requirements must be placed on the signaling links to meet availability requirements. For example, the two-way diversity rules state that the signaling links in each of the pairs $(AB, AC), (BD, BE), (CD, CE)$, etc., must be realized on physically diverse transmission facilities. For the signaling link quads (e.g., BD, BE, CE, CD), an additional requirement is usually required that either the pair (BD, CE) or the pair (CD, BE) be on diverse facilities (three-way diversity rule).

A routing example is shown in Fig. 7(b) that points out a difficulty that occurs with SLS code assignments when routing through multiple STP's (as in the interconnection of signaling networks in North America). In the example, the switching offices route to the STP's based on the least significant SLS bit and the STP's route to the next pair of STP's based on the second least significant bit. What happens is that the aggregate traffic to D from the pair (B, C) have SLS codes $XX1X$, and E has aggregate

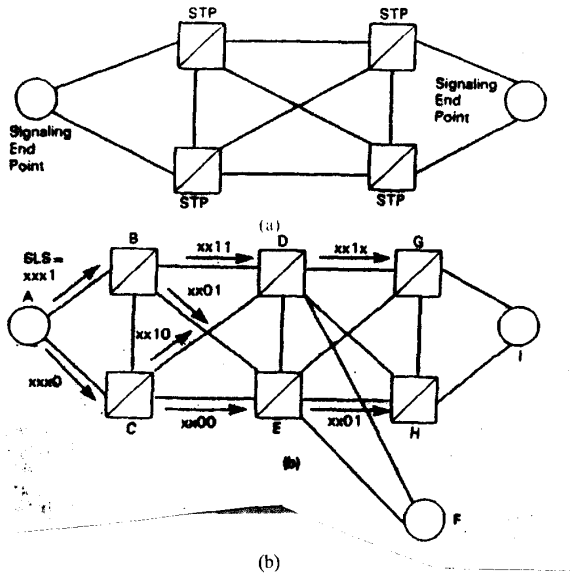
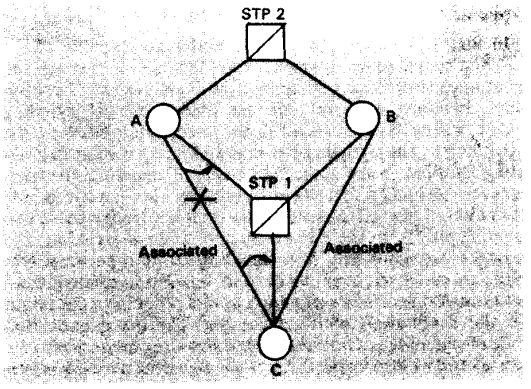


Fig. 7. Signaling network structure (ANSI).

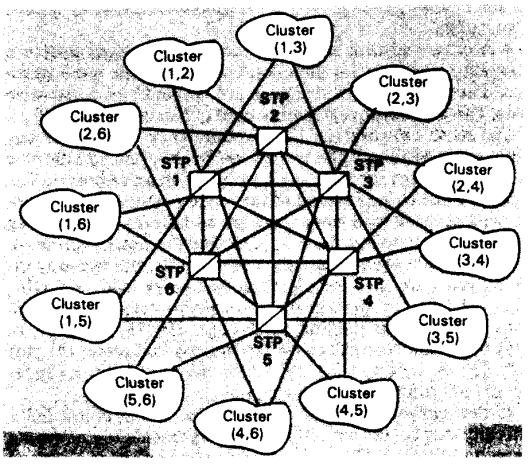
SLS codes $XX0X$ from the pair (B, C). The second least significant bit has become a dependent bit because B and C used that bit for routing to D and E. As a result, if D routes to the pair (G, H) based on the second least significant bit, a major traffic imbalance results since all of the through traffic from pair (B, C) will go on signaling link DG, for example. To avoid this load balancing problem in the North American network, the ANSI standards specify a 5-bit SLS field, the least significant bit indicates the STP to route to, and the SLS code is "rotated" after routing so that the dependent bit is moved out of the least four significant bits, which are used for routing at the STP's.

Other signaling network structures using a mix of associated and quasi-associated signaling modes are possible. For example the associated signaling mode can be used as a first choice route between two signaling points, and a quasi-associated route through an STP as backup in case the associated path fails. This configuration is illustrated in Fig. 8(a). Here, associated mode signaling is used between nodes A and C, as well as nodes B and C, with quasi-associated mode through STP 1 as backup. Nodes A and B, however, load-share the routes through STP 1 and STP 2, thus using only a quasi-associated mode. When the STP's used in this structure are integrated with the switches at a relatively large percentage of nodes (perhaps more than 50%), the resulting *distributed* signaling transport architecture provides for more robustness and survivability due to the distribution of signaling transfer function throughout the network.

Another possibility is a generalization of the mesh network. In this case there is a backbone network of fully interconnected STP's, and different clusters of offices are



(a)



(b)

Fig. 8. Alternate signaling network structures.

homed on different pairs of STP's. This configuration is illustrated in Fig. 8(b). An advantage of this structure is that when an STP fails its load is shed to a number of alternate STP's, and not just one mate as in the mesh network. Similar load shedding properties occur when backbone signaling links fail. The amount of reserve capacity that is needed to accommodate component failures is less in this type of network structure compared to the mated structure [5].

III. Signaling System No. 7 User Parts

In this section we briefly describe three major Signaling System No. 7 User Parts that use the transport services provided by the MTP and the SCCP: the Integrated Services Digital Network User Part (ISUP), the Transaction Capabilities Application Part (TCAP), and the Operations, Maintenance, and Administration Part (OMAP). Other User Parts like the Telephone User Part (TUP) and the Data User Part (DUP) will not be covered as their functionalities are provided in the ISUP protocol.

A. The Integrated Services Digital Network User Part (ISUP)

The ISDN User Part of the Signaling System No. 7 protocol provides the signaling functions that are needed to support the basic bearer service, as well as supplementary services, for switched voice and non-voice (e.g., data) applications in an ISDN environment. Prior to ISUP, another user part called the Telephone User Part (TUP) was specified that provides the signaling functions to support control of telephone calls on national and/or international connections. ISUP, however, provides all the functions provided by TUP plus additional functions in support of non-voice calls and advanced ISDN and Intelligent Network (IN) services. The following summary is based on the 1988 Blue Book version of ISUP.

Services supported by the ISDN User Part include the basic bearer service, and a number of ISDN supplementary services. ISUP uses the services of the MTP for reliable in-sequence transport of signaling messages between exchanges. It can also use some services of SCCP as one method of end-to-end signaling. Figure 1 shows the relationship of ISUP with the other parts of the SS7 protocol. In accordance with the OSI model, the information exchange between ISUP and MTP (or SCCP) takes place through the use of parameters carried by inter-layer service primitives. The ISDN User Part message structure is shown in Fig. 9. As seen in this figure, ISUP messages have variable lengths (an ISUP message can consist of up to 272 octets including MTP level 3 headers). All ISUP messages include a routing label identifying the origin and destination of the message[1], a circuit identification code (CIC), and a message-type code that uniquely defines the function and format of each ISUP message.[2] In the mandatory fixed part of the message, the position, length, and order of parameters is uniquely determined by the message type. The mandatory fixed part of an ISUP message is followed by a series of pointers that point to mandatory (and possibly optional) variable length parameters. Mandatory variable-length parameters are specified by a length field and a parameter-value field, while optional variable-length parameters are specified by a parameter name field, a length field, and a parameter-value field.

1) Basic Bearer Service: The basic service offered by ISUP is the control of circuit-switched network connections between exchange terminations. Figure 10 shows a typical basic call setup and release procedure between an originating exchange and a destination exchange, through an intermediate (or transit) exchange, using ISUP messages. The user-to-network or access signaling in this figure is performed by use of the DSS1 protocol (Q.931) on the D-channel. Thus in response to a Q.931 setup message, the

[1] As discussed in Section III-A, the routing label is actually an MTP Level 3 header (see Fig. 5), and not an ISUP header. It is shown in Fig. 9 primarily to highlight the fact that i) the ISUP fields are preceded by the routing label and ii) for each individual circuit connection the same routing label must be used in all messages associated with that connection.

[2] It is the CIC that provides the identification that relates all ISUP messages for a given *call*.

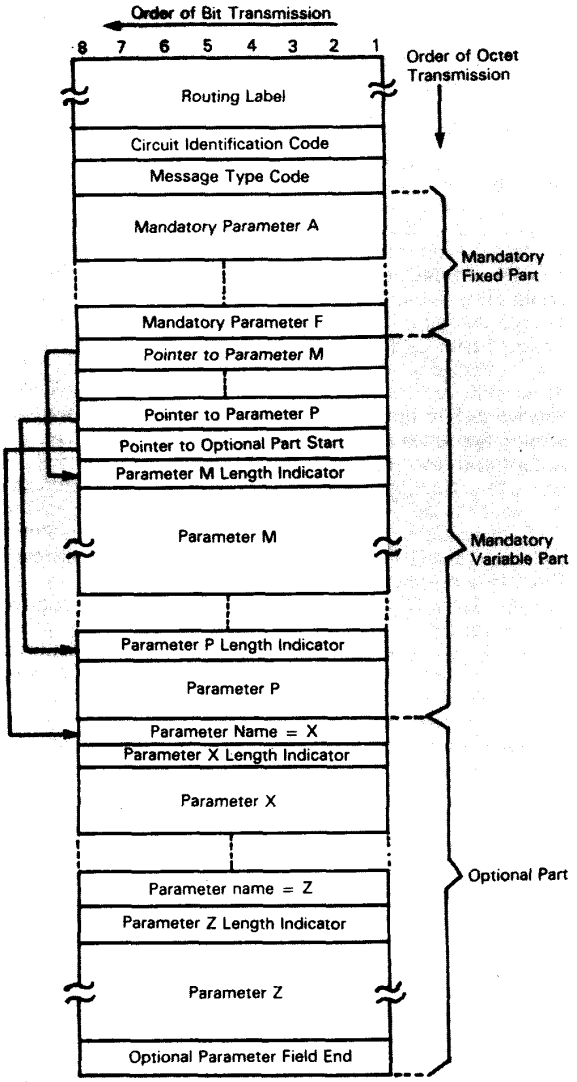


Fig. 9. ISDN-UP message structure.

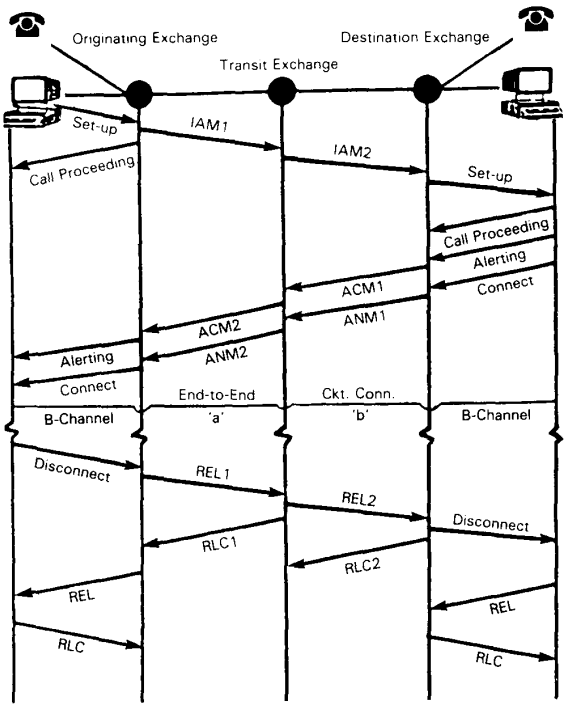


Fig. 10. ISDN-UP call setup example.

originating exchange launches an Initial Address Message (IAM1) toward the transit exchange for the purpose of setting up trunk *a*. The transit exchange processes IAM1, sets up trunk *a*, and launches another Initial Address Message (IAM2) toward the destination exchange requesting use of trunk *b*. As the destination exchange sets up trunk *b*, an ISUP-Q.931 interworking takes place in this exchange, resulting in transmission of a Q.931 setup message on the D-channel toward the subscriber. After the subscriber has been alerted, an Address Complete Message (ACM) is sent by the destination exchange to the transit exchange, which processes it and generates and launches another ACM toward the originating exchange. A Q.931 alerting message is then generated by the originating exchange and sent to the calling station. When the called party answers, a Q.931 Connect message received on the D-channel at the destination exchange causes an ISUP Answer Message (ANM) to be sent to the originating exchange, which sends a Q.931 Connect message to the calling station. The circuit-switched path between the calling and the called stations now consists of the cascade of a B-channel on the originating side, trunk *a*, trunk *b*, and a B-channel on the terminating side. Call tear-down is effected by use of ISUP messages Release (REL) and Release Complete (RLC), as shown in Fig. 10. Many other ISUP messages have been defined in support of the basic service for all contingencies that can arise on call setup, during the call, or on call tear-down, and for maintenance of associated circuits [6].

2) Supplementary Services: Supplementary services supported by ISUP include user-to-user signaling, closed-user group, calling line identification, call forwarding, etc. A short description of these services follows.

a) User-to-user signaling: The user-to-user signaling supplementary services provide a means of communication between two end users through the signaling network for the purpose of exchanging information of end-to-end significance. In order for these services to be possible, they need to be supported by the access protocol as well. Some of the end-to-end signaling methods use the services of SCCP while others use the services of MTP only. Three user-to-user signaling supplementary services have been defined. Service 1 allows the user-to-user signaling information to be included in the ISUP messages during call setup and clearing phases. Service 2 allows up to two ISUP User-to-User Information (UUI) messages to be transferred in each direction between end users during the

call setup phase. Service 3 provides for exchange of any number of ISUP-UUI messages between end users during the active phase of a call. All services can be implemented using the pass-along method. In this method, ISUP uses only the services of MTP. No processing of the pass-along information takes place at the transit exchanges. Services 2 and 3 can also be implemented using SCCP Connectionless (CL-SCCP) or Connection-Oriented (CO-SCCP) services.

b) Closed-user group: The closed user group (CUG) service allows a group of stations to communicate among each another only, with the option of some stations having incoming/outgoing access to users outside the group. A user can be a member of multiple CUG's. Each CUG is given an interlock code and this interlock code is assigned to all facilities associated with stations that belong to that CUG. When such a station initiates a call, a validation check is performed to verify that both the calling and the called stations belong to the CUG associated with the interlock code. The validation data can be in the exchange to which the station is connected (decentralized administration), or in a network database (centralized administration). In the centralized administration, the originating exchange sends a TCAP query to the appropriate network database and receives a response that determines the disposition of the call. Call setup can proceed normally from this point on if the call is permitted. Refer to Section III-B for more detail on TCAP interactions. This capability is the basis of a number of important virtual private network (VPN) services currently offered by a number of carriers.

In the decentralized administration of the CUG service, after the validation check is performed at the originating exchange, the interlock code of the selected CUG is transmitted in an IAM message to the transit exchange together with an indication on whether the calling station has outgoing access privileges. The transit exchange transmits this information to the succeeding exchanges. At the destination exchange, another validation check is performed to verify that the called party belongs to the CUG indicated by the interlock code. The call setup continues only if the information received checks with the information stored at the destination exchange.

c) Calling line identification: Another important supplementary service supported by ISUP is the calling line identity presentation and restriction. The calling line identity presentation (CLIP) service is used to present the calling party's number to the called party possibly with additional sub-address information. The calling party may have the option of activating the calling line identity restriction (CLIR) facility which would prevent the calling party's number from being presented to the called party. The transmission of the calling party's number to the destination exchange can be effected by either the originating exchange including it in the IAM, or by the destination exchange requesting it from the originating exchange through an ISUP Information Request Message. If CLIR is activated by the calling party, the originating exchange will provide the destination exchange with an indication in the IAM that the calling party's number is not to be presented to the called station.

d) Call forwarding: The call forwarding service provides for the redirection of a call from the destination originally intended to a different destination. Three types of call forwarding service have been defined: call forwarding unconditional, call forwarding busy, and call forwarding no reply. By requesting the call forwarding unconditional service, a subscriber is able to have the network redirect all calls, or just calls associated with a basic service, originally intended for a user's number to another number. The call forwarding busy service allows the user to do the same but only if the original destination is busy. The call forwarding no reply works in a similar way to the call forwarding unconditional but only after allowing the original destination to be alerted for a specified length of time before redirecting the unanswered call.

Upon receipt of an IAM with a called party number for which call forwarding is in effect, the destination exchange determines if the redirection number is in the same exchange. If so, it alerts that station and sends back an Address Complete Message containing the redirection number to the originating exchange. If the redirection number is in another exchange, an IAM that contains the redirection number as the called party number is sent from the original destination exchange to the exchange with the redirection number. The latter exchange then sends an Address Complete Message containing the redirection number to the originating exchange.

B. The Transaction Capabilities Application Part (TCAP)

Transaction Capabilities (TC) refer to the set of protocols and functions used by a set of widely distributed applications in a network to communicate with one another. In the SS7 terminology, TC refers to the application-layer protocols, called Transaction Capabilities Application Part (TCAP), plus any Transport, Session, and Presentation layer services and protocols that support it. For all SS7 applications that have been designed thus far, TCAP directly uses the services of SCCP, which in turn uses the services of MTP, with Transport, Session, and Presentation layers being null-layers. In this context, then, the terms TC and TCAP are synonymous (see Fig. 1).

Essentially, TCAP provides a set of tools in a connectionless environment that can be used by an application at one node to invoke execution of a procedure at another node, and exchange the results of such invocation. As such, it includes protocols and services to perform remote operations. It is closely related and aligned (except for one extension[3]) with the OSI Remote Operation protocol (ROSE) specified in Recommendations X.219 and X.229 [7]. In the telecommunications network, the distributed applications that use TCAP can reside in exchanges and in network databases. The primary use of TCAP in these networks is for invoking remote procedures in support of

[3]This extension is the Return Result-Not Last (RR-NL) component whose purpose is to carry the segments of a result that would otherwise be longer than the maximum allowed message size.

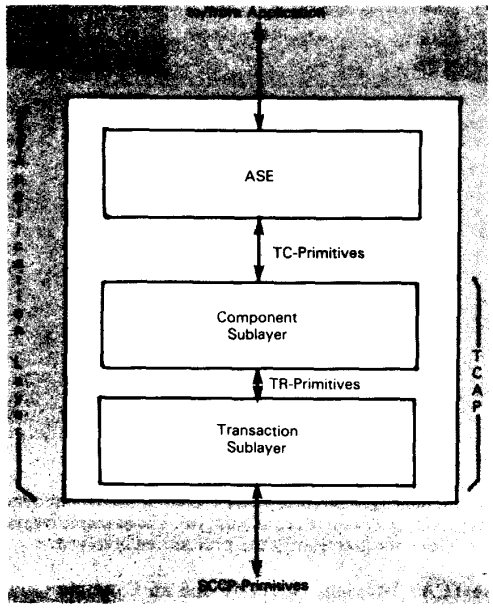


Fig. 11. Application layer structure.

Intelligent Network services like 800-service (free-phone). The application layer structure including TCAP is shown in Fig. 11. A TC-user Application Service Element (ASE) provides the *specific* information that a particular application needs (e.g., information for querying a remote database to convert an 800 number into a network-routable telephone number). TCAP provides the tools needed by *all* applications that require remote operation. TCAP itself is divided into two sub-layers: the component sub-layer and the transaction sub-layer. The component sub-layer involves exchange of "components" (the equivalent of Protocol Data Units or PDU's in ROSE) between TC-users. These components contain either requests for action at the remote end (e.g., invoking a process), or data indicating the response to the requested operation. The transaction sub-layer deals with exchange of messages that contain such components. This involves establishment and management of a dialogue (transaction) between the TC-users. A simplified discussion of the two TCAP sub-layers and an example are now given.

1) The Transaction Sub-Layer: A transaction (or dialogue)[4] defines the *context* within which a complete remote operation involving, for example, exchange of queries and responses between two TC-users, is executed. The transaction sub-layer is responsible for management of such a dialogue.[5] Two kinds of dialogues can take place between peer Transaction sub-layers: unstructured dialogue and structured dialogue. In the unstructured dialogue service, the Transaction sub-layer provides a means for a TC-user to send to its remote peer one or more components that do not require any responses. These components are received by the Transaction sub-layer from the TC-user (through the intervening Component sub-layer), and are packaged and sent to the remote Transaction sub-layer in a Unidirectional message. There is no explicit "association" established between peer Transaction sub-layers for this service.

The second kind of dialogue is the structured dialogue. Here, the TC-user issues a TC-BEGIN primitive containing a unique dialogue ID to the Component sub-layer. All the components that the TC-user sends within this dialogue would contain the same dialogue ID. The Component sub-layer maps this TC-BEGIN primitive into a TR-BEGIN primitive containing a *transaction* ID and issues it to the underlying Transaction sub-layer. There is a one-to-one correspondence between transaction ID's and dialogue ID's. Multiple components received by the Component sub-layer from the TC-user with the same dialogue ID can be grouped into a single TR-BEGIN message containing the appropriate transaction ID. The Transaction sub-layer manages each transaction (identified by its unique Transaction ID), groups components belonging to the same transaction into appropriate BEGIN, CONTINUE, END, and ABORT messages, and transmits them to its peer at the remote end (see Figs. 11 and 12).

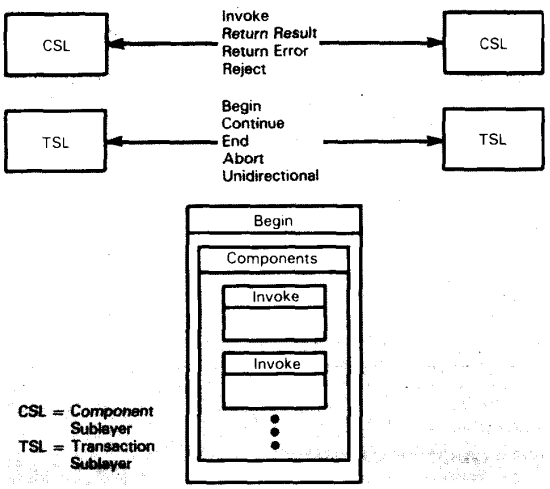


Fig. 12. TCAP sub-layer messages.

In short, the overall purpose of the Transaction sub-layer is to provide an efficient end-to-end connection between two TC-users over which they can exchange components related to one particular invocation of a distributed processing application. It avoids the OSI connection establishment and release overheads by packaging components in the 'connec-

[4]A dialogue refers to an explicit "association" between two TC-users, whereas a transaction refers to an explicit "association" between two peer transaction sub-layers. In TCAP, there is a one-to-one correspondence between dialogues and transactions, except in the case of an unstructured dialogue which does not use a transaction ID.

[5]The Transaction sub-layer is designed to provide an efficient *end-to-end* connection for exchange of "components" using the connectionless services of SCCP. In this role, it may be said to provide a very "skinny" and hybrid substitute for the missing layers (and specifically for the OSI Transport layer).

tion' setup and release messages. To reduce the number of signaling messages, it also supports a prearranged release facility where the peer Transaction sub-layers release their transaction resources related to a "dialogue" after a fixed period of time without the exchange of an END message.

2) The Component Sub-Layer: As alluded to earlier, a component consists of either a request to perform a remote operation, or a reply. Only one response may be sent to an operation request (which, however, could be segmented). The originating TC-user may send several components to the Component sub-layer before the Component sub-layer transmits them in a single message to its peer at the remote end. Components in a message are delivered individually to the TC-user at the remote end, and in the same order in which they were provided at the originating interface. Successive components exchanged between two TC-users for the purpose of executing an application constitute a dialogue. The Component sub-layer allows several dialogues to be run concurrently between two TC-users. Such dialogues can be unstructured or structured.

In the context of a structured dialogue, the Component sub-layer provides the function of associating replies with operations as well as handling abnormal situations. Associated with any invocation of an operation is a unique component ID. This allows several invocations of the same remote operation to be active simultaneously. The value of the invoke ID identifies an invocation of an operation unambiguously, and is returned in any reply to that operation. The Component sub-layer allows for four classes of remote operations. In class 1, both success and failure in performing the remote operation are reported. In class 2, only failure is reported, and in class 3 only success is reported. In class 4, neither failure nor success is reported. The replies to an operation could consist of one of the following components: Return Result (Last), Return Error, or Reject depending, respectively, on whether a result, error, or notification of syntax error in performing the remote operation is being provided (Fig. 12). Also, due to the signaling message size limitation, the segmentation of a successful result can be provided by the non-ROSE component Return Result-Not Last (RR-NL). In addition, any number of linked operations may be invoked prior to transmission of the reply to the original operation.

The reader is referred to CCITT Blue Book Recommendations Q.771–Q.774 for more details [8]. The example in the next sub-section can help clarify the procedures involved.

3) The TCAP Message Structure and an Example: The overall TCAP message structure is shown in Fig. 13. The encoding is according to Recommendation X.209, where every information element is coded as Name (TAG), Length, and Value. The transaction portion of the message specifies, through the use of a message type identifier (TAG), whether the message is a Unidirectional message, a Begin message, a Continue message, an End message, or an Abort message. It also specifies the total length of the message. This is followed, except in the case of a Unidirectional message, by the Name, Length, Value of the Transaction ID(s). The component portion of the message has a component portion Name and a component portion Length field, followed by the individual components. Each individual component in turn has a component type Name, a component Length field, and an information field specifying the required parameters for that component. The component type Name specifies whether the component is an Invoke, a Return Result, a Return Error, or a Reject component.

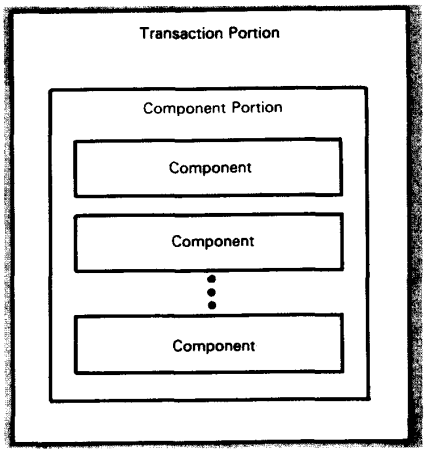


Fig. 13. TCAP message structure.

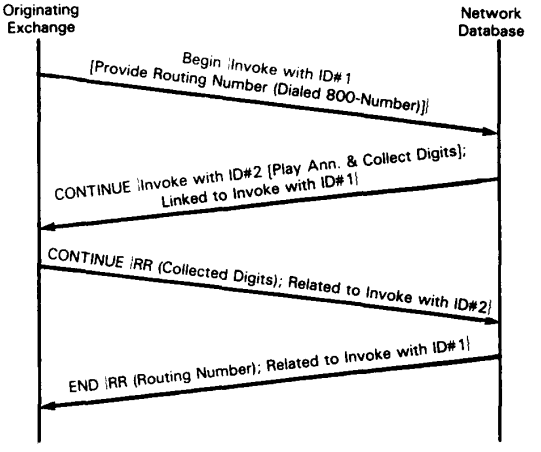


Fig. 14. 800 Service example using TCAP.

By way of a simplified example, consider an interactive 800 (free-phone) service and one way that it may be implemented using TCAP. Figure 14 shows the flow of TCAP messages between the originating exchange that has received the 800 call and the network database that contains the information for routing the call. In order for the database to provide the routing number, it is necessary for it to ask the exchange to play a certain announcement to the calling party, collect some more digits, and pass it on to the database.

The first TCAP message sent by the exchange is a BEGIN message that establishes a structured dialogue (and its associated transaction) between the exchange and the database in order to execute this application. Within the BEGIN message, a process that provides the routing number for the 800-service is invoked and given an invocation ID #1. The dialed 800 number is included as a parameter in the Invoke component. The database sends back a CONTINUE message as part of the structured dialogue in which it invokes a 'linked' operation that requires the exchange to play a certain announcement to the calling party and collects some digits. This invocation has Invoke ID #2 and is linked (related) to the invocation with Invoke ID #1. The exchange performs the required action and sends a CONTINUE message to the database. Within this CONTINUE, a Return Result component with invocation ID #2 is included containing the collected digits. Upon receipt and processing of this message, the database sends an END message to the originating exchange terminating the dialogue. Within the END message, however, a Return Result component is included with invocation ID #1, with the final routing number contained in it as a parameter.

C. The Operation, Maintenance, and Administration Part (OMAP)

The Operation, Maintenance, and Administration part (OMAP) of the Signaling System No. 7 provides the application protocols and procedures to monitor, coordinate, and control all the network resources that make communication based on Signaling System No. 7 possible. OMAP is specified in CCITT Blue Book Recommendation Q.791 [8].

The position of OMAP with respect to other parts of the Signaling System No. 7 is shown in Fig. 15. The collection of all the monitoring, control, and coordination functions above the application layer is known as the Systems Management Application Process (SMAP). All the management data that can be transferred or affected is contained in the Management Information Base (MIB). This data is gathered by the interaction of the MIB with the protocol entities at each layer through the Layer Management Interfaces (LMI). SMAP uses the services of TCAP through the OMAP Application Service Element (OMAP-ASE). OMAP-ASE sits on top of the TCAP Component sub-layer. An example of an OMAP-ASE is the ASE for MTP Routing Verification Test (MRVT) which uses the connectionless services of TCAP. MRVT is an important function of OMAP and is briefly described below and illustrated with an example.

The MRVT procedure tests MTP routing between any two signaling points in the network. A signaling point initiates an MRVT procedure for a given destination by sending MRVT messages to all its appropriate adjacent signaling points. Each node that receives this message processes it and "routes" it toward the given destination by sending an appropriate MRVT message to its adjacent nodes. This process is continued until the message reaches its destination. The destination node in response sends an acknowledgment message (MRVA) to the initiator of the MRVT in the same fashion. The example shown in Fig. 16

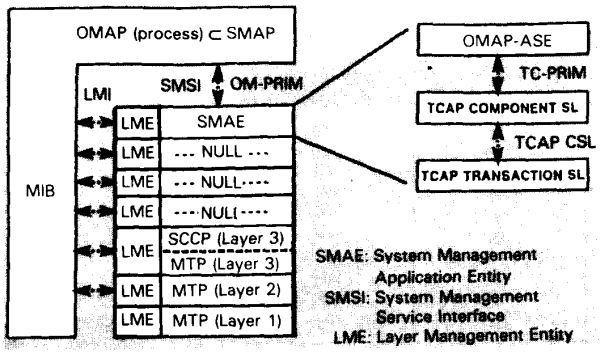


Fig. 15. SS7 management model.

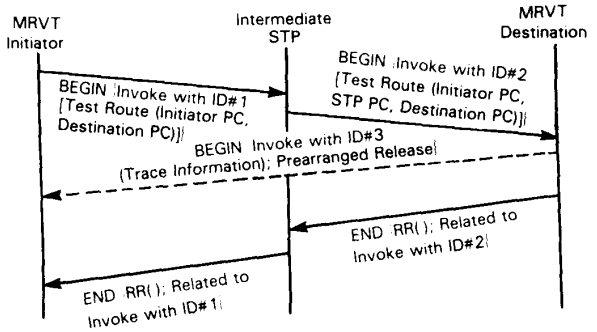


Fig. 16. MTP routing verification test with TCAP.

illustrates the procedure for a successful test of a message route that involves one intermediate STP. Notice that it is actually the BEGIN message here that has been *designated* as the MRVT message, and the END message that has been designated as the MRVA message. The MTP Routing Verification Result (MRVR) message is an optional message that the destination sends to the initiator of the MRVT procedure if the latter had asked for a trace in its original MRVT message.[6] This message is MTP-routed from the destination node to the initiator of the MRVT procedure. It is also an example of a prearranged release facility, alluded to in Section IV-B1), that would not require an END for release of transaction resources.

The MRVT procedure is designed to detect loops, excessive length of routes, excessive delays, inaccessibility of signaling points, and some other anomalies. In addition to MRVT, the OMAP procedures can be used for verifying SCCP routing and Global Title Translation process integrity (SRVT), management of routing data in different entities, circuit validation tests, link equipment failure management, link fault sectionalization, routine or on-occurrence measurement data collection, real time control and similar functions.

[6]The MRVR message is a particularly useful message when the MRVT test fails, as it can provide more detailed information that may be needed to trace the failure.

IV. Signaling System No. 7 Performance Objectives

There are three major categories for signaling network performance objectives: *Availability*, *Dependability*, and *Delay*. This section provides a summary of the current recommendations of CCITT, as well as those of ANSI. If an item is not directly identified as an ANSI or CCITT specification, it is common between both.

A. Availability Objectives

The unavailability of a signaling route set is determined by the unavailability of the components of the signaling network and the network structure. The *unavailability* of a system is defined as the probability that the system is in a failed state at a random point in time. This can be expressed in terms of the system mean-time-to-failure (MTTF) and mean-time-to-repair (MTTR) as $Unavailability = MTTR/(MTTF + MTTR)$. In applying this definition of unavailability to a pair of signaling network users, the system is defined to be the collection of all hardware and software in the user nodes, STP's, and the signaling links that are required to permit message transport between the pair of users.

MTP Unavailability Objective:

The unavailability of an MTP signaling route set should not exceed 1.9×10^{-5}, which equates to an expected downtime of less than 10 min/year.

From this objective and the network structure, availability objectives for the network components must be determined. Component objectives are not provided in the standards, since they depend on the network structure.

SCCP Relay Point Unavailability Objective:

The unavailability of an SCCP relay point should not exceed 10^{-4}, which equates to an expected downtime of less than 53 min/year.

B. Dependability Objectives

Dependability objectives relate to the ability of the network to reliably transport messages and not cause malfunctions. For the MTP there are four objectives:

Undetected Errors:

On each signaling link, not more than one in 10^{10} of all signal unit errors should be undetected by the MTP.

Lost Messages:

Not more than one in 10^7 messages should be lost due to failure of the MTP.

Messages Out-of-Sequence:

Not more than one in 10^{10} messages should be delivered out-of-sequence to the User Parts due to failure in the MTP. This includes duplicated messages.

Transmission Error Rate:

The signaling data link shall have a long term bit error rate that does not exceed 10^{-6}.

The ISUP dependability objectives are:

Probability of False Operation:

Not more than one in 10^8 of all signal units transmitted should be accepted and, due to errors, causes false operation.

Probability of Signaling Malfunction:

Unsuccessful calls can be caused by undetected errors, loss of messages, or messages delivered out-of-sequence. No more than 2 in 10^5 of all ISDN calls should be unsuccessful due to signaling malfunction. No more than 1 in 10^5 of all ISDN circuit connections should be unsuccessful due to signaling malfunction.

C. Delay Objectives

Delay performance is a very important attribute of signaling networks. There are no objectives given in the standards for MTP or User Part end-to-end signaling delays. Formulas are provided in Q.706 for signaling link queueing delays, which are based on M/G/1 queueing models. However, no recommendations are given in the CCITT standards on what delay criteria should be used in the traffic engineering and dimensioning of signaling networks. Work along these lines is currently being done in CCITT Study Group II. Reference [9] gives a good overview and list of references for performance modeling and delay considerations in Signaling System No. 7 networks.

At present, the standards provide some objectives on cross-office delays for STP's, SCCP message relay points, and ISUP switching exchanges. The cross-office transfer times that have been specified by CCITT and ANSI are given in Table 1. The cross-office transfer time is defined as the time from when the received message is delivered to level 2 to the time the corresponding outgoing message is delivered to level 1. Note that this delay includes the outgoing link queueing delay and emission time. No transmission system propagation times are included. Because queueing delays and emission times depend on the message length distribution, another useful office transit delay measure is *processor handling time*, which is the cross-office transit time less the outgoing link queueing and emission times. Work is in progress in the standards bodies to incorporate this measure into the recommendations.

V. Evolution of ISDN Signaling: Looking to the Future

In this section, we sketch a broad picture of the likely evolution of signaling systems in the remaining years of this decade based on trends that have emerged or are emerging in the global information age network. The discussion is divided into two parts: the first part is concerned with short-term trends (1992–1996), and the second part addresses the longer term trends (1996–2000 and beyond).

A. Evolution of Signaling Systems (1992–1996)

The short-term evolution of ISDN signaling systems is likely to be in the following directions.

- Widespread deployment of Signaling System No. 7 in public networks by various Administrations and Ex-

Table 1 Signaling Engineering and Performance Specifications

Signaling Point	Message Type	Load	Cross Office Transfer Time (ms)	
			Mean	95%
STP (ANSI)	All	Normal	45	80
		2×Normal	55	90
STP (CCITT)	All	Normal	20	40
		+15%	40	80
		+30%	100	200
ISDN Exchange	Simple	Normal	110	220
		+15%	165	330
		+30%	275	550
	Processing Intensive	Normal	180	360
		+15%	270	540
		+30%	450	900
SCCP Relay Point	Unit Data and Call Request	Normal	50–155	100–310
		+15%	100–233	200–465
		+30%	250–388	500–775
	Data and Conn. Conf.	Normal	30–110	60–220
		+15%	60–165	120–330
		+30%	150–275	300–550

Note: Assumes 64 kb/s emission and 15 octet average message length.

change Carriers. This will result in rapid proliferation of Intelligent Network capabilities and services within each carrier's SS7 network.

- Rapid growth of DSS1 signaling systems (both Basic and Primary Rate Interfaces) for user-to-network out-of-band signaling, and penetration of greater levels of intelligence to the edges of the network and to customer premise's equipment (e.g., very sophisticated PBX's as well as advanced ISDN workstations).
- Widespread interconnection of SS7 signaling networks between different Exchange Carriers (North America), and between different Administrations worldwide. This ushers in advanced ISDN and Intelligent Network services on an inter-network, international, end-to-end basis.
- Intensification and acceleration of work on signaling standards in support of multimedia and broadband applications as well as advanced Intelligent Network capabilities (see Section V-B).

B. Evolution of Signaling Systems (1996–2000)

In the closing years of this century, Broadband ISDN (B-ISDN) capabilities are likely to emerge in the network. B-ISDN provides a cell-based[7] network infrastructure with extremely high-speed switching and transmission capabilities. It will provide a unified transport infrastructure that can be used for *all* kinds of traffic: voice, data, image, video, signaling, OA&M, etc. Although the extent of penetration of B-ISDN services in the telecommunication market place may be unclear, it is quite clear that broadband technologies will become commercially available and some new services using these capabilities will be offered. Given that the B-ISDN network will be based on Asynchronous Transfer Mode (ATM) switching/transmission principles, and implemented on a ubiquitous optical fiber facility infrastructure, use of enormous bandwidth on demand will become not only technologically possible, but also economically feasible. Truly integrated multimedia services involving voice, high-speed data, image, and video will emerge and penetrate the business and residential markets with a potential to profoundly impact and transform the very fabric of those markets and the nature of the "work place".

In the B-ISDN environment of tomorrow, signaling will play a crucial role. In the context of multimedia services, call control and connection control functions will have to be separated. A call is an end-to-end entity whereas a connection (bearer) may have only a link-by-link significance. In its duration, a multimedia multipoint "call" may require the capability to add/drop a number of connections and/or legs involving widely different bandwidths. Evolution of signaling to the broadband era has two major components, a transport component and a user-part component.

Work has been underway in CCITT for some time on signaling user part evolution to accommodate multimedia, broadband, and intelligent network needs. It has its origin in the effort to extend signaling capabilities for separation of call control from connection control. Initially, the call control and connection control functions of ISUP were identified and conceptually separated. This led to the concept of "separated ISUP" (in contrast to the monolithic ISUP). Subsequently, the term *ISDN Signaling Control Part (ISCP)* was coined to designate a new user part of the Signaling System No. 7 protocol in which such separation is built in from the outset. According to the ISCP Baseline Document [10], the principles and premises on which the ISCP effort is based include the following.

- The ISCP architecture should be based as much as possible on the OSI Application Layer Structure (OSI-ALS).
- ISCP should be viewed as both the network signaling protocol and the access signaling protocol for B-ISDN (in contrast to the current state of affairs where two different protocols with similar functionalities but different origins are used).
- ISCP functions and capabilities should be developed in the context of a B-ISDN environment to provide

[7] A cell is a fixed-size packet of 48 octets of information and 5 octets of control overhead, for a total of 53 octets.

for separation of bearer control from call control, to support supplementary services, and to be applicable to Intelligent Network services.

- To enhance portability, ISCP should be positioned to use an OSI Network Service for transport. If MTP/SCCP is to provide such a transport, work is needed to further align it with OSI.

ISCP is supposed to modularize the communication capabilities needed for its signaling applications into ASE's. Depending on the nature of the "call"/"connection" requested (the ISCP "Application Context"), appropriate ASE's will be dynamically combined to provide the required protocols. The nature of these ASE's and the various ISCP "Application Contexts" are the subject of current study in CCITT.

The signaling transport evolution work in CCITT, on the other hand, started quite some time later [11].

The issues to be addressed here relate to the kind of signaling transport architecture and protocol that can be used in an ATM environment to provide the reliability that signaling transport needs while making efficient use of the enormous broadband capabilities of ATM networks in support of new and vastly expanded signaling applications. Here, a number of alternatives, ranging from retention of MTP to fully associated signaling mode using signaling permanent virtual circuits and a skinny part of MTP Level 3, have been identified. Although it is generally agreed that MTP Levels 1 and 2 are going to be replaced by the Physical and ATM layers in the B-ISDN protocol model, questions related to the more complex evolution of MTP Level 3 and SCCP to the broadband environment are only beginning to be studied [12], [13].

The work currently underway on ISCP should be influenced appropriately with the work on signaling message transport issues.

REFERENCES

[1] A. R. Modarressi and R. A. Skoog, "Signaling System No. 7: A tutotial," *IEEE Commun. Mag.*, vol. 28, pp. 19–35, July 1990.

[2] R. A. Skoog, "Engineering common channel signaling networks for ISDN," *Proc. ITC-12,* Torino, Italy, June 1988.

[3] R. R. Goldberg and D. C. Shrader, "Common channel signaling interface for local exchange carrier to interexchange carrier Interconnection," *IEEE Commun. Mag.*, vol. 28, pp. 64–71, July 1990.

[4] J. J. Lawser, J. Matsumoto, and J. M. Pigott, " Common channel signaling for international service applications," *IEEE Commun. Mag.*, vol. 28, pp. 89–92, July 1990.

[5] R. A. Skoog, H. Ahmadi, and S. Boyles, "Network architecture planning for common channel signaling networks," in *Proc. 2nd Ann. Int. Symp. on Network Planning,* Brighton, UK, Mar. 1983.

[6] CCITT Study Group XI, "Specifications of Signaling System No. 7," Blue Book, vol. VI — Fascicle VI.8, Geneva 1989.

[7] CCITT Study Group VI, " Data communication networks," CCITT Blue Books , vol. VIII, Fascicles VIII.4 and VIII.5, Geneva, Switzerland, 1989.

[8] Study Group XI, "Specifications of Signaling System No. 7," Blue Book, vol.VI, Fascicle VI.9, Geneva, Switzerland, 1989.

[9] G. Willman and P. J. Kuhn, "Performance modeling of Signaling System No. 7," *IEEE Communications Mag.*, pp. 44–56, July 1990.

[10] ISCP Baseline Document, Temporary Document 699-E, CCITT Working Party XI/6 (Question 11/XI), Ottawa, Canada, Oct. 1989.

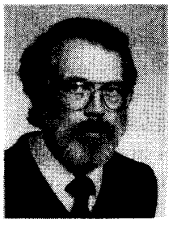
[11] R. A. Skoog and A. R. Modarressi, "Alternatives and issues for network signaling transport in a broadband environment," *Computer Networks and ISDN Systems,* vol. 20, pp 361–368, 1990.

[12] A. R. Modarressi and M. Veeraraghavan, "Network signaling architectures for broadband ISDN," in *Proc. Forum 91 Technical Symp.*, pt. 2, vol. III, pp. 193–197, Geneva, Switzerland, Oct. 1991.

[13] G. I. Stassinopoulos and I. S. Venieris, "ATM adaptation layer protocols for signaling," *Computer Networks and ISDN Systems*, vol. 23, pp. 287–304, 1992.

Abdi R. Modarressi (Member, IEEE) received the B.E.E. degree from the American University of Beirut, Lebanon, in 1969, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1973 and 1976, respectively.

In 1976, he joined AT&T Bell Laboratories and worked on new algorithms for optimization of traffic routing in the national toll network. Between 1978 and 1982 he taught control and communications courses at the National University of Iran. In 1983 he directed the Network Planning Group at the Telecommunications Company of Iran. He rejoined AT&T Bell Laboratories in 1984, where he has been involved in performance studies of common channel signaling networks, conformance testing in these networks, signaling architecture studies in narrowband and broadband environment, and applications of object-oriented methods to implementation of SS7 networks. He is currently a Distinguished Member of Technical Staff in the Signaling Platforms Department at AT&T. He has authored several journal and conference papers on multidimensional system theory and applications, congestion control performance in data networks, signaling technology, and signaling architecture in broadband networks. While at the National University of Iran, he also coauthored a two-volume text on the "Theory and Applications of Control Systems."



Ronald A. Skoog (Member, IEEE) received the B.S. degree from Oregon State University in 1964, and the M.S. and Ph.D. degrees from M.I.T., Cambridge, in 1965 and 1969, respectively, all in electrical engineering.

From 1969 to 1971 he taught in the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He joined AT&T Bell Laboratories in 1971, where he worked in the area of transmission facility network planning and optimization. He became a supervisor in 1974. In 1981 he began working in the area of common channel signaling networks and supervised work on new signaling network architecture studies, performance, and simulation studies of signaling neworks, establishing signaling network performance objectives, and developing signaling network engineering and dimensioning methodologies. He has written a number of journal and conference papers in the areas of non-linear systems, control theory, optimization, signaling network design, performance analysis of real-time systems, and performance of signaling networks.

Dr. Skoog is a member of Sigma Xi.