Copyright Notice

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

Capacity and Performance Analysis of Signaling Networks in Multivendor Environments

Marcos Bafutto, Paul J. Kühn, and Gert Willmann, *Member, IEEE*

Abstract—The load of common channel signaling networks is being increased through the introduction of new services such as supplementary services or mobile communication services. This may lead to a performance degradation of the signaling network, which affects both the quality of the new services and of the services already offered by the network.

In this paper, a generic modeling methodology for the signaling load and the signaling network performance as a result of the various communication services is extended in order to include certain implementation-dependent particularities. The models are obtained by considering the protocol functions of Signaling System No. 7 as specified by the CCITT, as well as the information flows through these functions. With this approach, virtual processor models are derived which can be mapped onto particular implementations. This allows the analysis of signaling networks in a multivendor environment.

Using these principles, a signaling network planning tool concept has been developed which provides the distinct loading of hardware and software signaling network resources, and on which hierarchical performance analysis and planning procedures are based. This allows to support the planning of signaling networks according to given service, load, and grade-of-service figures.

A simple case study outlines the application of the tool concept to a network supporting Freephone, Credit Card, and ISDN voice services.

I. Introduction

MODERN telecommunication networks can be characterized as large and complex distributed systems. The information flows necessary for the coordination of processes related to call and connection control, distributed application processing, and network management are transported through the signaling network. A good introduction to signaling networks can be found, for example, in [5], [25], [32], [36], and in the introductory paper in this issue.

The protocol architecture of the signaling network was the subject of international standardization by the CCITT [8] (Comité Consultatif International Télégraphique et Téléphonique), and a set of recommendations is available since the last decade. The national recommendations generally reflect regional particularities and are specified by national institutions, for example, as in the United States [1], or by public telecommunication operating companies, as in Germany [17].

The early applications supported by the signaling network were mostly circuit-related, that is, voice services. These services are characterized by intense signaling activities only in the connection setup and release phases. In this case, the availability of trunks in the transport network represents an upper bound for the signaling network traffic, and the number of circuits handled by a single signaling channel is typically in the range of 2000 [32].

Since then, the signaling network load has been increased through the introduction of new services that require signaling activities not related to the setup of connections over the transport network, like database transactions or remote procedure invocations. These signaling activities may be present in any phase of a call or even when there is no call active, for example, for keeping track of a mobile subscriber location. In a study on Personal Communication Networks (PCN) [33], it was concluded that the load of the signaling network will be 4 to 11 times greater for cellular than for ISDN (Integrated Services Digital Network), and 3 to 4 times greater for PCN than for cellular. It was also observed that these results are very sensitive to the model assumptions.

The additional load caused by services such as supplementary services, Intellignet Network (IN) applications, mobile communication services, and Universal Personal Telecommunication (UPT) may lead to performance degradation of the signaling network and, therefore, affect the Quality of Service (QOS) of new services as well as of the services already offered by the network.

In this paper, a generic modeling approach is used to develop a tool concept which is suitable to support the planning of new signaling networks according to given service, load, and grade-of-service figures, or to detect bottlenecks or possible deficiencies in case of resource outages.

II. Signaling Network Planning Problems

The problems faced by the network planner in a medium term is the balance between required capacity, expected traffic, available budget, and target grade-of-service to be achieved. In a medium term, the nonstationary behavior of the network, for example, due to overload control actions, is not primarily subject of the planning process, and the traffic variables may be considered as mean values of a busy-hour call attempt. The reader interested in more information on signaling network congestion and flow control is referred, for example, to the references given in [39] and [52].

Manuscript received June 21, 1993; revised September 10, 1993. This work was supported by CAPES Brazil under Grant 9532/88-15, and by the Deutsche Forschungsgemeinschaft (DFG).

M. Bafutto is with Telecommunications of Goiás S. A. (Telegoiás), Goiânia, Brazil, on leave at the Institute of Communications Switching and Data Technics, University of Stuttgart, Stuttgart, Germany.

P. J. Kühn is with the Institute of Communications Switching and Data Technics, University of Stuttgart, Stuttgart, Germany.

G. Willmann is with Alcatel SEL AG, Stuttgart, Germany.

IEEE Log Number 9214767.

For an already-existing network, additional problems arise from the introduction of new services. The user behavior with respect to new services may be different from that for the existing ones, and patterns of traffic variation as well as grade-of-service (GOS) concepts may also be different. In this case, a careful analysis must be carried out, directed to the new service as well as to its impact on the currently supported services.

Within the signaling network domain, a telecommunication service is characterized by a number of signaling messages exchanged between the involved nodes. According to the service type, different network capabilities may be required. As a result, changes of the network load and, therefore, of the QOS parameters may be caused by the introduction of a new service, by variations in the traffic intensity of a particular service, or by temporary outages. The various network components are typically affected in the following way:

- *Links:* The offered signaling link load per call may change drastically with the introduction of new service concepts.
- *Signaling Points (SP's) and Signaling Transfer Points (STP's):* The increasing signaling traffic load requires more processing capacity; since this additional load may not be homogeneously distributed over all processes within an SP or STP, only some processors may become overloaded.
- *Service Control Points (SCP's):* Since the capacity of an SCP is determined by the number and by the types of transactions as well as by the offered functionality, the effect of the introduction of a new service can be estimated by the resource capacity required to process the related transactions.

The complexity of the signaling network environment and its importance to the network have motivated the development of various signaling network planning tools (e.g., see [2]–[4]). In contrast to this paper, however, most tools found in the literature deal with planning aspects which are not primarily related to performance in terms of delay.

III. Modeling Framework

The analysis of a real network initially requires a modeling methodology which is able to represent the internal processes of the protocols. However, the consideration of all mechanisms of a complex protocol may create mathematical difficulties making the resulting models intractable. Nevertheless, methodologies to determine the throughput and delay behavior of communication architectures analytically can be found in the literature (see [13], [14], [26], [40], [51], [52]). In [13], [14], [26], a method based on multiple-chain product-form queueing networks is proposed. Another approach based on decomposition and aggregation techniques is presented in [51], [52]. A comparison between some of these methodologies can be found in [15].

A. Modeling Methodology

The modeling methodology adpoted here, which uses decomposition and aggregation techniques, follows [51] and [52]. All submodels are derived directly from the CCITT functional specifications [8], including internal mechanisms such as segmenting/forking of messages, thus reflecting the internal behavior of the underlying functional blocks. The basic idea is the observation that, in the CCITT specifications, both the set of functional entities and the distinct information flows through these entities are precisely defined. From this, it is possible to construct "virtual" processor models.

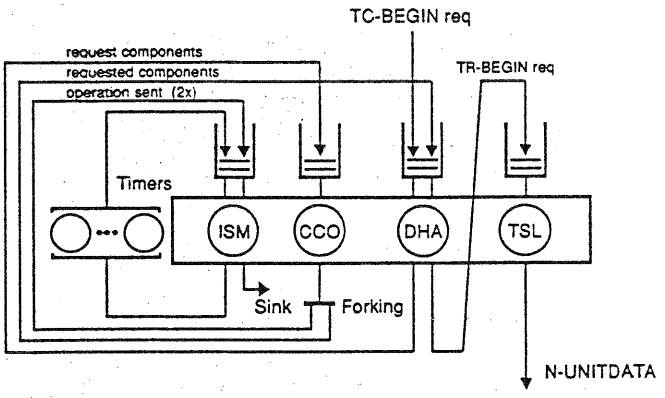

Fig. 1. TCAP functional block: the message chain for the Dialog Begin message containing two Invoke components.

The principles of this methodology can be briefly explained using the TCAP (Transaction Capabilities Application Part) block as an example. According to [8, Figure A-2a/Q.774], the TCAP is composed of two subblocks: the Transaction Sublayer (TSL) and the Component Sublayer (CSL). The CSL consists of the Dialog Handling (DHA) and the Component Handling (CHA). The CHA is further subdivided into the Component Coordinator (CCO) and the Invocation State Machine (ISM). From this, a virtual processor model comprising four distinct processing phases (TSL, DHA, CCO, and ISM) and four message input queues is derived. Inside this submodel, there are different message routing paths, that is, different message chains. As an example, the message chain corresponding to an outgoing Dialog Begin message containing two Invoke components is described below and depicted in Fig. 1.

- The primitive "TC-BEGIN req" received from the TC User is processed by DHA. Any Invoke components with the same Dialog ID (in this case, two components) are then requested from the CCO through a "request components" signal.
- The CCO processes the "request components" signal and generates three outputs signal (fork): two "operation sent" signals to the ISM (one for each Invoke component) and then one "requested components" signal to the DHA.
- Under reception of each of the two "operation sent" signals, the ISM starts an invocation timer. No output is generated (sink). The case of timeout can be modeled by a message branching with the branching probability given by the timeout probability.
- When the DHA received the "requested components" signal, it composes a "TR-BEGIN req" primitive to the TSL.
- The TSL processes the "TR-BEGIN req" and requests the service of the Signaling Connection Control Part (SCCP) through an "N-UNITDATA" primitive.

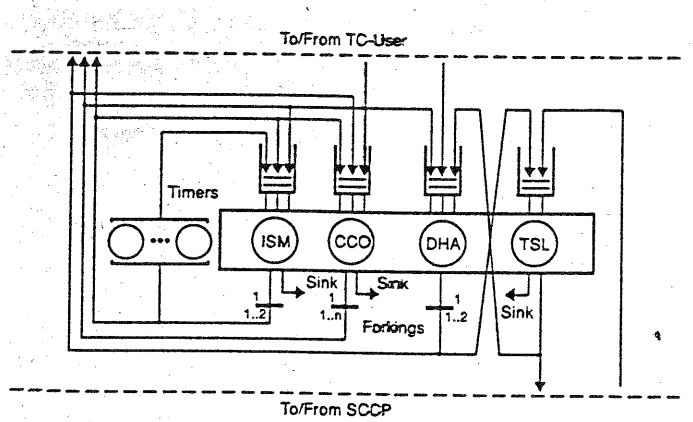

Fig. 2. TCAP functional block: the generic submodel.

The TCAP functional block is not restricted to the Dialog Begin message; it comprises a set of messages consisting of a combination of all possible types for the transaction portion and component portion. The representation of all these message chains requires the extension of the model shown in Fig. 1 through additional chains. The full model, which is depicted in Fig. 2, is obtained by considering the set of all possible message chains for the TCAP functional block.

The models for the functional blocks of the levels 3 and 4 of the signaling network protocol architecture (MTP Level 3, SCCP, and ISUP) are obtained in the same way.

It should be noted that real implementations need not necessarily follow the structure used by the CCITT functional specifications. In a particular product, it is possible that processing phases are combined in a completely different way, or even subdivided into additional phases.

In order to embed all these models in a sufficiently realistic environment, it is necessary to extend the models in such a way that aspects related to the call control and to the interprocessor communication are included.

The call control itself is a complex software system which is strongly coupled to the product arcitecture. For simplicity, it is modeled by traffic sources and traffic sinks representing the traffic generated by and destined to the users, and by infinite servers representing general delays such as, for example, exchange and user delays or database access delays. The reader interested in more detailed switching system control modeling is referred, for example, to [21], [27]. The interprocessor communication delays are dependent on the particular communication protocols [31], [48], and the corresponding delay characteristics are also approximated by infinite servers.

The full model of an SP is depicted in a reduced form in Fig. 3. A detailed description of all submodels is given in [52].

The lower levels are represented by MTP Level 1 and MTP Level 2. MTP Lever 1 is simply modeled as an infinite server with a service time representing the signaling link propagation delay. The modeling approach described before cannot be applied to the MTP Level 2 entities, because they are closely coupled via the error correction and flow control mechanisms on level 2. The performance analysis of signaling links under certain simplifying assumptions can be found in [7], [10], [18], [24], [35], [44]–[46], [50]. A collection of formulas extracted

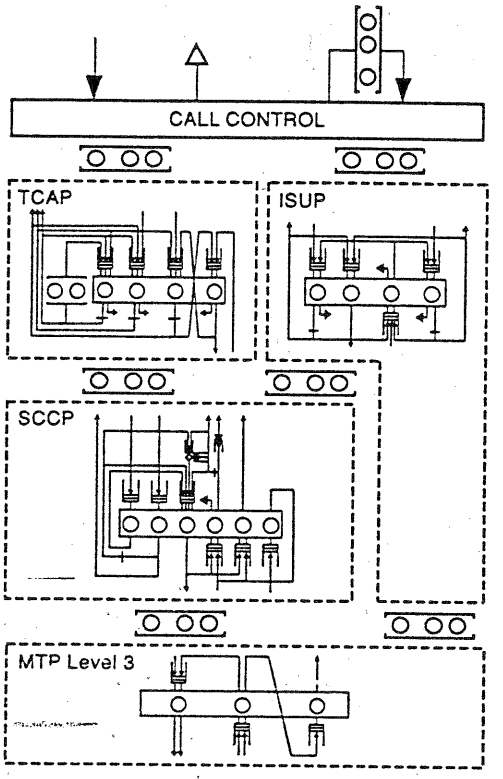

Fig. 3. Generic model of a signaling point.

from some of these publications were incorporated into CCITT Recommendation Q.706 [8]. In this work, the CCITT approach and the corresponding queueing delay formulas are adopted.

B. Consideration of Implementation Architectures

The derived submodels were designated as "virtual" in the sense that they can be implemented in different ways. For example, in a distributed SP architecture, each functional block (MTP Level 3, SCCP, TCAP, and ISUP) might be implemented in a separated processor; while in a contralized one, various functional blocks might share the same processor. Therefore, in order to account for vendor-specific particularities, the protocol model must be mapped onto the specific implementation according to its architecture.

The distribution of the functional blocks through the hardware support is strongly coupled with the underlying switching system architecture. In most cases, a decentralization tendency may be observed in the functions related to links and subscribers, that is, MTP Level 2 and ISUP.

Examples for real implementations are the AT&T U.S. 5ESS [6] switch from AT&T, ALCATEL 1000 S12 [9] and ALCATEL 1000 E10 [12] from Alcatel, System X [19] from British Telecom, FETEX-150 [22] from Fujitsu, MS7 [23] from GTE, System 8300 [28] from Alcatel, DMS [29] from Northern Telecom, NEAX61 [34] from NEC, AXE10 [41] from L. M. Ericsson, the STP No. 2 [42] from AT&T, and EWSD [47] from Siemens. The referred list is not exhaustive and should be considered as a sample of a larger universe of products; it does not represent any restriction by the authors with respect to any omitted product.

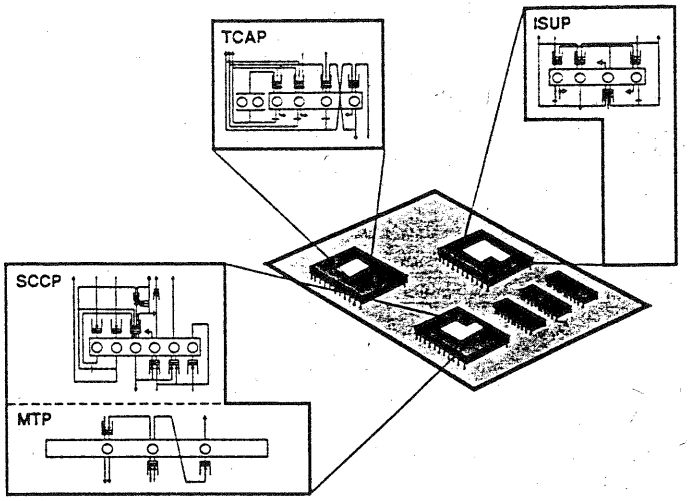

Fig. 4. Mapping of the signaling point model onto a hypothetical implementation.

In order to provide a better overview, a complete model of the levels 3 and 4 and its mapping onto a hypothetical implementation are depicted in Fig. 4. In this case, it is assumed that the ISUP and the TCAP are implemented in isolated processors, while the MTP Level 3 and SCCP functions are deployed in a common processor. However, Fig. 4 shall be understood in a didactic context; in reality, the functional blocks are implemented in processor boards and not in single chips as illustrated.

IV. Analysis Outline

The complete model for the entire signaling network includes extended queueing network elements such as, for example, full-duplex flow-controlled links, priority processors, multiple-chain multiple-class traffic streams, segmenting and reassembling of messages, etc. The exact analysis of such a large and complex system is far beyond the current knowledge, and an approximation based on a combined decomposition and aggregation technique is used.

A. Decomposition and Aggregation

The principle of decomposition is to break up a complex system into its subsystems in order to achieve a reduction in the complexity of the whole system. This approximation is valid if the system may be classified as *nearly decomposable* [16], [30], that is, if the interactions between the subsystems are largely dominated by the local interactions inside each subsystem. It is possible that some mapping of the functional blocks on a particular product may violate this assumption. However, in current implementations, dedicated processors are usually not assigned to single subprocesses of the higher-level functional blocks of the protocol, which guarantees at least some local interactions within each subsystem.

The network is decomposed into link sets, SP's, and STP's. Then, in the second decomposition step, the link sets are further decomposed into single signaling links, whereas the SP's and STP's are decomposed into their submodels according to the mapping onto the particular implementation, that is, the distribution of the functional blocks among the processors.

The basic idea behind the signaling traffic aggregation is the observation that a particular subsystem is shared by a large number of connections and, because of this, message streams belonging to individual connections need not be distinguished from the corresponding aggregate traffic streams. The aggregate arrival processes to each subsystem are approximated by Poisson processes.

The output processes of such models are no longer Poisson. Since many different traffic streams are usually superimposed in the subsequent queueing models, however, the Poisson approximation can still be used and leads to sufficiently reliable results in most practical cases.

B. Analysis of the Submodels

The above assumptions allow the approximate analysis of the decomposed systems in isolation.

The performance analysis of signaling links is primarily based on so-called M/GI/1 priority queueing models (see [7], [10], [18], [24], [35], [44]–[46], [50]). The formulas summarized in the CCITT recommendations give closed-form expressions for the mean and for the standard deviation of the message queueing time in MTP Level 2 under certain simplifying assumptions [8]. The total transfer time across a signaling link is given by the sum of the queueing time, the emission time, the channel propagation delay, and the processing times in the signaling terminals.

In the models for the remaining functional blocks (MTP Level 3, SCCP, TCAP, and ISUP), a message departing from one processing phase can be fed back to another process in the same processor or even to the same process. In addition, one message can be forked or segmented upon feedback. This functionality is implemented in the switching system software, where priority-based strategies are often used to schedule the process execution. These assumptions lead to the classification of the corresponding models into a rather general class of M/GI/1 priority systems with feedback.

The approximate evaluation of the response time distribution for such systems has been first presented in [20]; further work in this direction can be found in [49]. In the approach followed there, however, only the chain to be evaluated is modeled in detail, while the remainder of the processor is only modeled by chains with probabilistic feedback.

Another approach, which is based on [11], is the so-called method of moments to derive mean performance measures. A methodic treatment of this approach has been given in [43] and extended in [37], [38] to consider branching, forking, and more sophisticated scheduling strategies. However, since all message chains visiting a model must be distinguished, the analysis tends to become rather complex, but nevertheless is not time-consuming. In principle, the mean sojourn times for all message chains are obtained from the solution of a system of linear equations. An advantage of this method is the capability to be implemented in a general aglorithmic form, yielding the analysis of general models.

The available analysis techniques allow the consideration of most of the internal mechanisms occurring in the generic submodels, but there are some restrictions concerning the adopted approximation.

The decomposition approach, that is, the analysis of the submodels in isolation, does not allow the inclusion of the window flow control mechanism of SCCP protocol class 3; however, this protocol class is rarely used in current applications. The processor overhead, which is typically between 5 and 20%, is subdivided into context switching and processsor job management overhead. The effect of the context switching overhead may be included through a service time inflation of the processing phases, whereas the processor job management overhead can only be taken into consideration via an isolated high-priority processing phase with the preemptive or nonpreemptive interruptions approximated by Poisson processes. The tight functional connection of call control and ISUP may, in some cases, complicate the mapping of the virutal model onto the implementation. And, finally, the weak correlations between messages belonging tc the same scenario, analyzed in [46], are considered as negligible.

V. A Planning Tool Concept

Using the described methodology, a conceptual signaling network planning tool has been implemented. This planning tool takes into consideration the network configuration, the routing plan, the impact of the individual messages on the network components, the mix of scenarios, and the traffic matrix.

The network configuration and the routing plan are network-specific information and include aspects related to the topology, equipment type, and routing strategy. The impact of the individual messages on the network components is obtained by a detailed description of the unitary message load characteristics. The contribution of a message to the network load comprises the message length and the sequence of visited processes in the network elements on the path between its origin and destination. The information about the functional blocks visited by a message is obtained from the SDL (Functional Specification and Description Language) diagrams of the CCITT specifications, as shown in Section III for the TCAP Dialog Begin message.

A catalog with the description for a large number of messages is available, and the user just has to provide the message length and the processing time on each process of the underlying implementation. The messages are used to compose the mix of scenarios representing the services supported by the network. It should be noted that even a single service may generate a set of subscenarios; hence, a normal ISDN voice call consists of at least three subscenarios, that is, a successful call, no answer, and destination subscriber. Each one of these subscenarios corresponds to a distinct message exchange over the signaling network and must be considered individually. The traffic matrix gives the amount of call attempts per second between either two local exchanges.

TABLE I
Basic Steps of the Planning Process Supported by the Tool Concept

Step	Description
1	Identification of network structure and elements, routing strategy, service types, message scenarios per service type, message types and parameters, and traffic matrix;
2	Message flow analysis;
3	Decomposition and component analysis;
4	Calculation of global performance measures.

The first step is to perform a message flow analysis. This yields the message flow rates through all transmission and processing resources with respect to all message types. From the unitary message load characteristics, the resource utilizations follow straightforwardly. In order to provide a better survey on the impact of the introduction of a new service in an existing network, the load information is computed on a per-subscenario basis.

The performance evaluation phase starts with the application of the decomposition approach to consider the signaling links and the SP's and STP's in isolation. The SP's and STP's are further decomposed according to the particular architectures, that is, the mapping of functional blocks onto physical processors.

The analysis of the processor models requires the priority assignments to the processing phases within each processor. This priority assignment represents one of the parameters that determine the sojourn time for a particular message type. Depending on the priority assignment and traffic load, the differences between the sojourn times of distinct message types can be significant. Furthermore, the processor can be overloaded for some message types, but still be able to process higher-priority messages; such a behavior is typical for priority systems (see [11]).

The message arrival rates at the decomposed subsystems are obtained from the results of the message flow analysis. When there are several identical units available, it is assumed that the incoming traffic is homogeneously distributed between these units. The mean link delays are calculated from the CCITT formulas [8]. The SP and STP submodels are analyzed in isolation, based on the corresponding M/GI/1 priority queueing models with feedback. The application of the algorithm described in [38] to each subsystem yields the mean sojourn time for each message chain, that is, for each message type visiting this subsystem.

The global performance measures are obtained by composition of the submodel results. The end-to-end transfer time of a particular message is computed by the summation of the individual transfer times, sojourn times, and other delays along its path through the network.

With the end-to-end transfer time of a particular message sequence, it is possible to evaluate response delay parameters, such as connection setup delays, data transfer delays, or database query delays.

The basic steps of the planning process supported by the tool concept are summarized in Table I; its capabilities are demonstrated by a case study in the next section.

TABLE II
DISTRIBUTION OF FUNCTIONAL BLOCKS AMONG PHYSICAL PROCESSORS AND PROCESS PRIORITIES FOR THE DIFFERENT PRODUCTS (HIGHER PRIORITY VALUES CORRESPOND TO HIGHER PRIORITIES)

Product	MTP-L3			SCCP						ISUP				TCAP			
	HMDT	HMDC	HMRT	SCRC (r)	SCRC (t)	SCLC (r)	SCLC (t)	SCOC (r)	SCOC (t)	MDSC	MSDC	CPCO	CPCI	TSL	DHA	CCO	ISM
"A"	Processor 1			Processor 2						Processor 3				Processor 4			
	2	1	3	3	4	1	2	6	5	4	1	2	3	1	2	3	4
"B"	Processor 1									Processor 2				Processor 3			
	2	1	3	6	7	4	5	9	8	4	1	2	3	1	2	3	4
"C"	Processor 1			Processor 2													
	2	1	3	8	7	5	6	9	10	14	11	12	13	4	3	2	1

VI. A CASE STUDY

The following example is primarily provided to demonstrate the capabilities of the described planning tool concept. In general, a real network is rather complex, comprising a large number of nodes and particularities concerning the topology, routing strategy, traffic matrix, etc. Nevertheless, the analysis can be carried out in the same way as demonstrated in this example.

A. Network Structure and Product Architectures

First, it is necessary to define the architecture of the products deployed in the network, that is, the distribution of the functional blocks among the physical processors for each product. The next step is to assign the product-specific priorities to the processing phases of a functional block in a specific processor. The chosen architectures with the corresponding distribution of the functional blocks among the physical processors as well as the chosen priority of execution of each processing phase are summarized in Table II. As described in [52], the three basic SCCP processes are split into transmitting (t) and receiving (r) subprocesses.

These architectures aim to provide a fair representation concerning the various grades of centralization, varying from a more decentralized architecture (product "A") to a more centralized one (product "C"). It should be noted that, in many practical cases, the product architecture turns out to be a function of the switch size.

The topology of the example network consists of a three-level hierarchical network; it is depicted in Fig. 5. The highest hierarchical level (level 1) comprises 5 fully interconnected STP's and an SCP linked to the STP's with code 200 and 300. The next lower hierarchical level (level 2) contains the transit SP's. The number of transit SP's connected to a specific STP reflects factors such as, for example, area coverage and subscriber density, and is generally variable from STP to STP. This characteristic is included in our example in such a way that there are three transit SP's under each one of the STP's with code 100, 400, and 500, while there are just two transit SP's under the STP's with code 200 and 300. For the lowest hierarchical level (level 3), the same arguments may be applied to justify a heterogeneous number of SP's under each transit SP. Again, it is assumed that the lowest level structure is the same for the areas 100, 400, and 500, where 2, 3, and 4 SP's are connected to the transit SP's, respectively. In the areas 200 and 300, there are 2 and 4 SP's under the transit SP's, respectively.

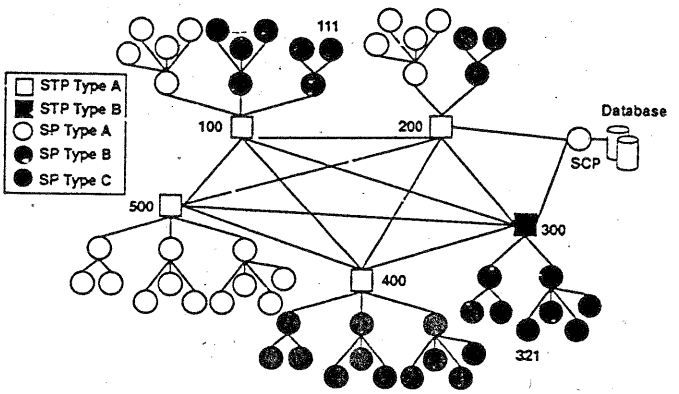

Fig. 5. Topology of the example network.

Each link set consists of 4 links in the highest level and 2 links in the remainder of the network. The transmission capacity of each link is 64 kb/s, and the propagation delay is 5 ms. The links are assumed to operate with the basic error correction method and to be free of disturbances.

In a multivendor environment, it is also necessary to identify the product type of each one of the STP's and SP's. The products deployed in a given area are determined by factors such as, for example, the grade of deregulation, the type of operator (PTT or private), if the network provider also produces equipment, etc. It is assumed that in some areas all STP's and SP's are of the same type, as for the areas 300 (exclusively type "B") and 500 (exclusively type "A"). There are also areas which are dominated by few vendors, for example, area 400 (STP of type "A"; SP's of type "C") and 200 (STP of type "A"; SP's of types "A" and "C"). A fully heterogeneous environment is assumed in area 100 (STP of type "A"; SP's of types "A"; "B"; and "C"). The SCP has a decentralized architecture and is of type "A".

B. Service Mix

The services in operation in the example network are the ISDN voice service, the Freephone service, and the Credit

TABLE III
MESSAGES USED BY THE BASIC SERVICES

Message	Designation	Length
IAM	Initial Address Message	60
ACM	Address Complete Message	20
ANM	Answer Message	15
REL	Release Message	20
RLC	Release Complete Message	15
INV	Invoke Message	60
RES	Response Message	70

Card service. The Freephone service is assumed to be an extension of the ISDN voice service with retrieval of some routing information from the database. The Credit Card service requires two database queries: the first to determine how to prompt the customer, and the second to check the user service access authorization.

The subscenarios for these services may vary according to the service characteristics. The subscenarios for the ISDN voice service are classified into normal call, subscriber busy, and no answer. The Credit Card service includes, in addition to the ISDN subscenarios, the case of service termination due to unsuccessful subscriber authentication. In the case of Freephone, the subscenarios of ISDN theoretically apply. In reality, however, the case of no answer may be considered as negligible, given that the caller generally listens at least to a recorded message.

In this example, it is assumed that the ISDN voice service comprises 70% successful calls, 20% subscriber busy, and 10% no answer, respectively. In 10% of the Credit Card service calls, there are access authentication errors, while the remaining calls are subdivided in the same way as the ISDN voice service. The Freephone service is characterized by 80% successful calls and 20% subscriber busy.

The traffic matrix contains the information about the traffic between the SP's of the network, and the amount of related input data may be large even for networks with few nodes. In real networks, this information should generally be available on magnetic media. In order to avoid excessively long input data files, an automatic procedure for traffic matrix generation is adopted. The generated traffic is considered to be the same for all signaling points and directed as follows: 50% homogeneously distributed between the SP's under the same level 2 transit exchange, 25% to the SP's under the same level 1 STP, and 25% to the remaining SP's, respectively. The calls originated in an SP are assumed to be 80% ISDN, 10% Freephone, and 10% Credit Card calls, respectively.

C. Message Types and Parameters

The message types and their lengths (in octets) generated by the considered scenarios are listed in Table III. In this example—for simplifying the required input data—it is assumed that the processing times in a process are the same for all messages, and that the processing time differences inherent to the distinct products are compensated by a processor speed factor.

TABLE IV
MESSAGES USED BY THE NEW SERVICE

Message	Designation	Length
IAM (CR)	Initial Address Message	90
CC	Connection Confirm Message	20
DT1	Data Form 1 Message	80
REL	SCCP Release Message	20
RLSD	SCCP Released Message	20

The database query is performed using the TC. The processing times in the TCAP block of an SP are assumed to be 2 ms for DHA and CCO, 1 ms for TSL, and 0.5 ms for ISM, respectively. All processes of the SCCP have processing times of 1 ms. For the ISUP, the processing times are 2 ms for CPCI and CPCO, and 0.5 ms for MSDC and MDSC. In MTP Level 3, the processing times are 1 ms for HMDT and HMRT, and 0.5 ms for HMDC, respectively. All processing times are assumed to be constant.

Interprocessor communication delays of 5 ms are assumed for all interfaces between the processors. The call control response delays are 50 ms for circuit establishment, 20 ms for circuit release, and 10 ms for other actions, respectively. Finally, a database query in the SCP takes 200 ms.

D. Introduction of a New Service

In this environment, the introduction of a new service that allows simultaneous exchange of voice and data information is considered. The transfer of User-to-User Information (UUI) elements in either direction during the active phase of a call is performed using UUI Service 3. The transport of the UUI messages is assumed to use SCCP end-to-end signaling connections based on protocol class 2 [8], where the end-user data information is transmitted in Data Form 1 (DT1) messages. The SCCP connection setup request (CR: Connection Request) is embedded in the IAM message, and the successful setup is acknowledged by a Connection Confirm (CC) message. Finally, the connection release procedures are started simultaneously for both the bearer and the signaling connection.

In this new service, it is assumed that 5 DT1 messages are transmitted in each direction. The characteristics of the additional messages of this service are summarized in Table IV. The scenarios are the same as for the ISDN voice service.

The connection setup delay for the deployed services from node 321 to node 111 is considered. The selection of this pair of nodes is justified by the presence of all considered product types along the path between them. The new service is still in the planning phase, and according to the introduction strategy, a substitution of 3 or 5% of the ISDN voice service by the new service is expected. The interest of the network planner is focused on the behavior of the call setup delay of the supported services for the different introduction alternatives. The new service is studied in terms of the end-to-end transfer delay of a DT1 message.

E. Results

Before evaluating the network performance measures, it is interesting to check the accuracy of the adopted modeling approach through a simulation study. A good agreement between

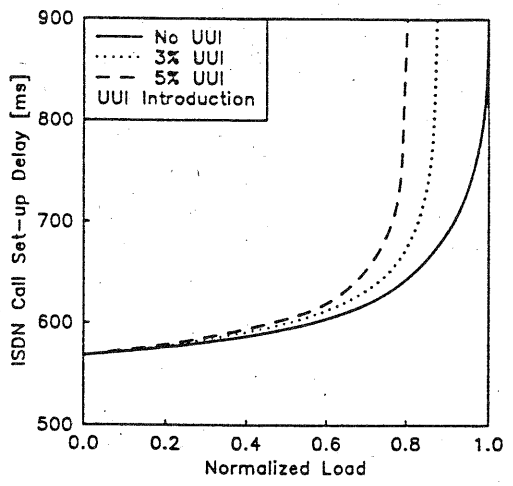

Fig. 6. Impact of the new service on the ISDN voice service.

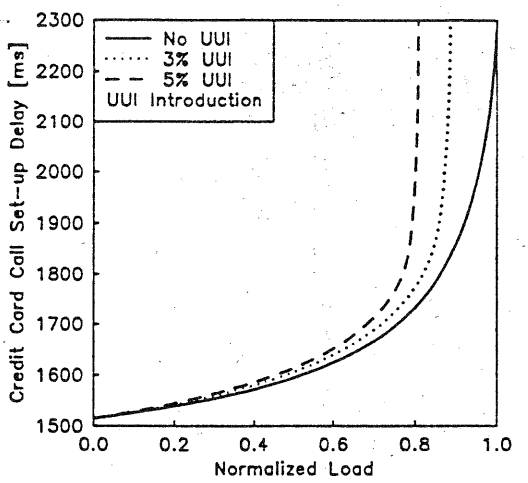

Fig. 8. Impact of the new service on the Credit Card service.

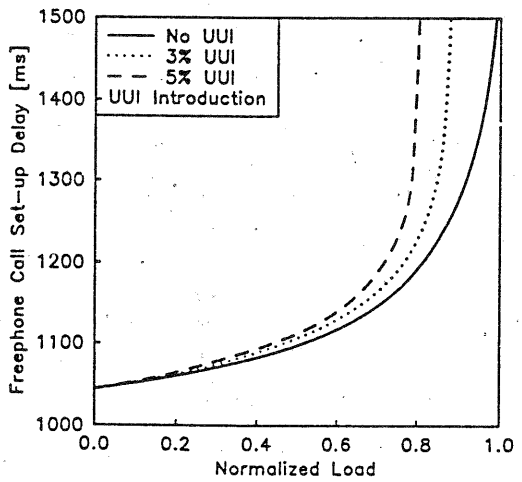

Fig. 7. Impact of the new service on the Freephone service.

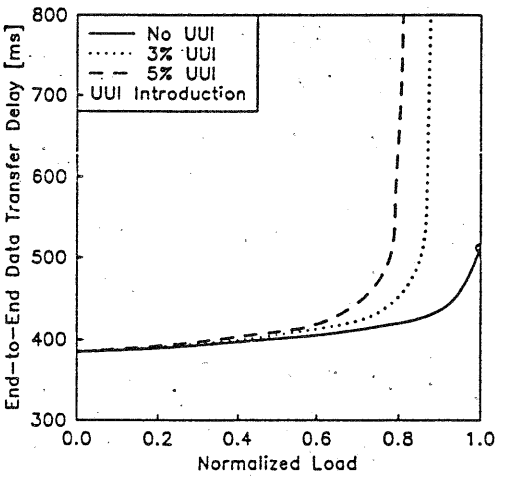

Fig. 9. User-to-user information transfer delay (sum of both directions).

analysis and simulation results suggests that the assumptions of the analysis methodology do not introduce significant errors. Comparisons between analytical and simulation results for this modeling methodology can be found in [3], [52].

All analytical results are depicted with the call attempt loading normalized with respect to the maximum call attempt loading of the network, that is, the traffic value that causes an overload situation in any network resource. The impact of the new service introduction on the ISDN voice service is depicted in Fig. 6. The results for the Freephone and Credit Card services are shown in Figs. 7 and 8, respectively. Due to the unbalanced network traffic, the end-to-end user data transfer delays are not the same for both directions, that is, the results from node 321 to 111 and from 111 to 321 are different. This effect is included in Fig. 9 by representing the end-to-end delay as the sum of the message transfer delays in both directions.

The introduction of the new service and the related transport of user information impacts the services supported by the network. The amount of exchanged information is about 300 octets/call in each direction, which is probably an underestimated value even for applications with a medium level of complexity.

Next, the sensitivity of the results with respect to the priority assignment of the processes in a processor is studied. The priority assignment is a crucial factor in the determination of the system delay characteristics, because any process with priority p in Table II must wait until there are no more processes with priority $> p$ active in this processor. In order to demonstrate the priority assignment effect, the network analysis is repeated with the priority assignment of Table II completely inverted, that is, the highest priorities now become the lowest priorities, and vice versa. The new results for the Credit Card service are shown in Fig. 10; and Fig. 11 depicts the results for the end-to-end data transfer delay.

The changes observed in the corresponding results can be easily explained by the priorities listed in Table II. The common part to the underlying services is mainly concentrated in the SCCP. The Credit Card service uses the TC to retrieve data from the database, and the TC messages are further transported using the SCCP connectionless protocol, that is, the processes SCLC(r) and SCLC(t). In contrast, the user-to-user signaling is supported by the SCCP connection-oriented protocol, that is, it involves the processes SCOC(r) and SCOC(t). Since the inversion of priorities assigns higher priorities to the SCLC processes than to the SCOC processes,

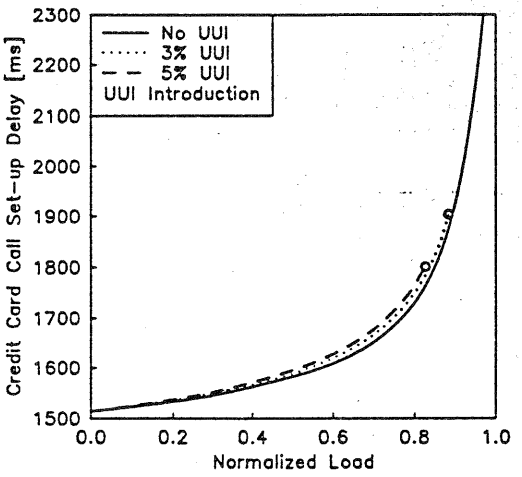

Fig. 10. Impact of the new service on the Credit Card service after priority inversion.

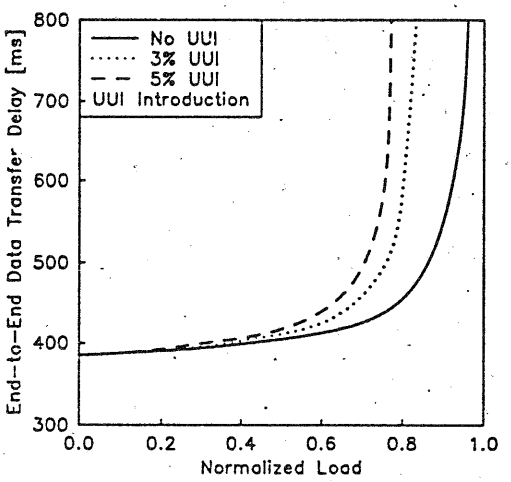

Fig. 11. User-to-user information transfer delay after priority inversion (sum of both directions).

messages related to the Credit Card service are generally processed before those related to UUI Service 3. With these observations, the insensitivity of the Credit Card service with respect to the new service is not surprising; on the other hand, the delay characteristics of UUI Service 3 are significantly affected by the priority inversion, as expected.

In this example, the UUI Service 3 assumes the coupling of multiple connection sections at each intermediate transit node; but this service could also be MTP-routed from the originating to the destination point without coupling at the transit nodes. If this is the case, the corresponding delays may not be very sensitive to the priority inversion/reassignment.

Another particularity of Fig. 10 concerns the discontinuity of the curves after the saturation point of the first processor which is becoming overloaded. It indicates that this processor is only overloaded with respect to the low-priority processes, but it continues to execute higher-priority processes. In this case, the lower-priority processes of this processor are becoming saturated, and the analysis cannot be continued although the processor may still be able to process messages visiting only the higher-priority processes.

In this case study, the system bottleneck is located either in the TCAP processor of the SCP (no UUI) or in the MTP/SCCP processor of the SP with code 320 (3 or 5% UUI). The reasons for this are mainly due to the centralized position of the SCP and to the concentration of functions in product "B" (MTP Level 3 plus SCCP), respectively.

It should be noted, however, that the architectures as well as the values used in this example are hypothetical and arbitrary. Therefore, the particular behavior of the network elements and the presented results should not be taken as a judgment of any implementation or service.

The total CPU time requirements to produce all graphics of this case study were less than one hour on an HP 9000/725 workstation. Each curve depicted in Figs. 6–11 is plotted using spline interpolation on roughly 10 points.

VII. Conclusions

In this paper, an earlier-described modeling approach for signaling networks based on Signaling System No. 7 has been extended in order to account for certain implementation-dependent characteristics of signaling network elements. It has been shown that the consideration of physical implementation aspects, such as the distribution of the processes among processors and the priority assignment to processing phases, allows to cover important aspects present in multivendor environments.

Based on this methodology, a signaling network planning tool concept has been described. This concept yields the distinct loading of many hardware and software signaling network resources, and forms the basis for hierarchical performance analysis and network planning procedures. However, since the underlying models do not generally cover the whole variety of real implementations, this tool concept has to be adapted to particular architectures.

The results of an example case study indicate that QOS parameters of future networks will be strongly influenced by the performance of the signaling network. In particular, the introduction of new services which require database interactions—such as IN applications, mobile communication services, or UPT—or user-to-user signaling connections will have a significant impact on signaling network performance aspects such as, for example, the processing load of signaling points and the end-to-end message transfer delays.

Acknowledgment

The helpful comments of three anonymous referees are greatly appreciated.

References

[1] ANSI, Standards T1.110–T1.116, "American National Standards for Telecommunications—Signalling System No. 7," Amer. Nat. Standards Inst., Bethesda, MD, 1987–1990.

[2] M. N. Antonios and C. E. Perez, "Resource planning in common channel signaling (CCS) networks," in *Proc. 5th Int. Network Plan. Symp. (NETWORKS)*, Kobe, Japan, May 1992, sess. 13, pp. 253–258, Paper 2.

[3] M. Bafutto and P. J. Kühn, "Capacity and performance analysis for signalling networks supporting UPT," in *Proc. 8th ITC Specialist Sem. Universal Personal Telecommun.*, Santa Margherita Ligure (Genova), Italy, sess. VII, Oct. 1992, pp. 201–213, Paper 1.

[4] M. Bafutto, P. J. Kühn, G. Willmann, and J. Zepf, "A capacity and performance planning tool for signalling networks based on CCITT Signalling System No. 7," in *Intelligent Networks—The Path to Global Networking*, P. W. Bayliss, Ed. Washington, DC: IOS Press, 1992, pp. 368–379.

[5] British Telecom, "CCITT Signalling System No. 7," *Brit. Telecommun. Eng.*, vol. 7, part 1, (whole issue), Apr. 1988.

[6] R. B. Brown, C. V. Holmes, M. D. Lanoux, and T. P. Marciani, "Common channel signaling in the AT&T U.S. 5ESS® switch," *Proc. IEEE*, vol. 80, pp. 618–627, Apr. 1992.

[7] M. Buttò, G. Colombo, and A. Tonietti, "Delay distribution in a data communication system with error recovery," in *Proc. Int. Conf. Commun. (ICC)*, Boston, MA, June 1979, vol. 3, sess. 43, Paper 8.

[8] CCITT, *Blue Book, Volume IV, Fascicles VI.7-VI.9*, "Specifications of Signalling System No. 7," Recomm. Q.700–Q.795, Int. Telecommun. Union, Geneva, 1989.

[9] A. Chalet and I. Ess Skinner, "ISDN transaction architecture for the System 12 digital exchange," *Elect. Commun.*, vol. 61, no. 1, pp. 50–56, 1987.

[10] H. K. Cheong, "Signalling System No. 7 link capacity estimation," in *Proc. 1st Australian Teletraffic Res. Sem.*, Clayton, Australia, Nov. 1986, sess. 8, Paper 3.

[11] A. Cobham, "Priority assignment in waiting line problems," *J. Operat. Res. Soc. Amer.*, vol. 2, no. 1, pp. 70–76, Feb. 1954.

[12] P. Collet, J. Craveur, P. Lucas, F. Lanquetot, J. Botherel, and D. Courtel, "Introduction of CCITT No. 7 Signalling System into French exchanges," *Commutat. Transmiss.*, vol. 5, no. 3, pp. 5–24, 1983.

[13] A. E. Conway, "Performance modeling of multi-layered OSI communication architectures," in *Proc. Int. Conf. Commun. (ICC)*, Boston, MA, June 1989, vol. 2, sess. 21, pp. 651–657, Paper 1.

[14] ——, "Queueing network modeling of Signaling System No. 7," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 1990, vol. 1, sess. 406, pp. 552–558, Paper 2.

[15] ——, "A perspective on the analytical performance evaluation of multilayered communication protocol architectures," *IEEE J. Select. Areas in Commun.*, vol. 9, pp. 4–14, Jan. 1991.

[16] P. J. Courtois, *Decomposability: Queueing and Computer System Applications*. New York: Academic, 1977.

[17] DBP-Telekom, FTZ-Richtlinien zum ZGS Nr. 7, "Richtlinien 163 TR 71-82," Deutsche Bundespost Telekom Forschungs- und Technologiezentrum, Darmstadt, 1993.

[18] V. G. Dedobortsh, G. P. Basharin, M. A. Zharkov, and K. E. Samuilov, "Methods of quality parameter analysis for common channel signalling system," in *Proc. 3rd Int. Sem. Teletraffic Theory*, Moscow, U.S.S.R., June 1984, pp. 77–89, Paper 11.

[19] D. G. Fisher, R. J. Manterfield, R. Bekaert, J. van Goethem, and D. L. Thomas, "Experience in the implementation of the CCITT No. 7 Signalling System," in *Proc. 4th World Telecommun. Forum*, Geneva, Switzerland, 1983, sess. 1.3, Paper 3.

[20] B. Fontana and C. Diaz Berzosa, "M/G/1 queue with N-priorities and feedback: Joint queue-length distributions and response time distribution for any particular sequence," in *Proc. 11th Int. Teletraffic Congr. (ITC)*, Kyoto, Japan, Sept. 1985, vol. 1, sess. 3.3A, Paper 4.

[21] B. Fontana, M. Villén-Altamirano, and G. H. Petit, "Models and tools for evaluating the traffic handling performance of System 12 ISDN exchanges," *Elec. Commun.*, vol. 61, no. 1, pp. 104–109, 1987.

[22] Y. Fujiyama, H. Takeichi, and T. Masuda, "Implementation of the Signalling System CCITT No. 7 in FETEX-150 digital switching system," in *Proc. 11th Int. Switching Symp. (ISS)*, Florence, Italy, May 1984, vol. 3, sess. 31B, Paper 7.

[23] B. Gobbi, A. Lazzari, and G. Premoli, "MS7: A switching node for common channel signalling networks," in *Proc. 12th Int. Switching Symp. (ISS)*, Phoenix, AZ, Mar. 1987, vol. 1, sess. B2, Paper 1.

[24] G. Hebuterne, "Evaluation en trafic d'une procédure de ligne: Le système de signalisation CCITT no 7," *Ann. Télécommun.*, vol. 36, nos. 5–6, pp. 315–327, May–June 1981.

[25] B. Jabbari, "Common Channel Signalling System Number 7 for ISDN and intelligent networks," *Proc. IEEE*, vol. 79, pp. 155–169, Feb. 1991.

[26] P. S. Kritzinger, "A performance model of the OSI communication architecture," *IEEE Trans. Commun.*, vol. COM-34, pp. 554–563, June 1986.

[27] P. J. Kühn, "Analysis of switching system control structures by decomposition," in *Proc. 9th Int. Teletraffic Cong. (ITC)*, Torremolinos, Spain, Oct. 1979, vol. 2, sess. 52, Paper 4.

[28] J. Lamy, J. P. Olivier, and J. C. Pennanec'h, "CCS No 7 transfer point," in *Proc. 12th Int. Switching Symp (ISS)*, Phoenix, AZ, Mar. 1987, vol. 1, sess. B2, Paper 5.

[29] M. Langlois and B. Sembi, "DMS SuperNode: The cornerstone of a CCS7 network," *Telsis*, vol. 15, no. 2, pp. 16–27, 1988.

[30] S. S. Lavenberg, Ed., *Computer Performance Modeling Handbook*. New York: Academic, 1983.

[31] D. Manfield, P. Tran-Gia, and H. Jans, "Modelling and performance analysis of inter-processor messaging in distributed systems," *Perform. Eval.*, vol. 9, no. 2, pp. 83–91, Apr. 1989.

[32] R. Manterfield, *Common-Channel Signalling*, IEE Telecommunications Series 26. London: Peter Peregrinus, 1991.

[33] K. S. Meier-Hellstern, E. Alonso, and D. R. O'Neil, "The use of SS7 and GSM to support high density personal communications," in *Proc. 3rd Rutgers WINLAB Workshop Wireless Inform. Networks*, East Brunswick, NJ, Apr. 1992.

[34] H. Miyoshi, H. Manabe, M. Makishi, J. Ishihara, C. Iwata, H. Kikushi, and M. Kawakami, "Commercial application of CCITT Signalling System No. 7 in the NEAX61," *NEC Res. Devel.*, no. 81, pp. 61–67, Apr. 1986.

[35] A. R. Modarressi and R. A. Skoog, "Performance considerations of signaling networks in an ISDN environment," in *Proc. Brussels Special. Sem. ISDN Traffic Issues*, Brussels, Belgium, May 1986, sess. 3, Paper 2.

[36] ——, "Signaling System No. 7: A tutorial," *IEEE Commun. Mag.*, vol. 28, pp. 19–35, July 1990.

[37] M. Paterok and O. Fischer, "Feedback queues with preemption-distance priorities," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 17, no. 1, pp. 136–145, May 1989.

[38] M. Paterok, *Warteschlangensysteme mit Rückkopplung und Prioritäten*, Arbeitsberichte Inst. Mathematische Maschinen Datenverarbeitung, Friedrich Alexander Univ. Erlangen-Nürnberg, Band 23, Nr. 12, Erlangen, 1990.

[39] V. Ramaswami, "Analysis of the link error monitoring protocols in the common channel signaling network," *IEEE/ACM Trans. Network.*, vol. 1, no. 1, pp. 31–47, Feb. 1993.

[40] M. Reiser, "Communication-system models embedded in the OSI-reference model, A survey," in *Computer Networking and Performance Evaluation*, T. Hasegawa, H. Takagi, and Y. Takahashi, Eds. Amsterdam, The Netherlands: North-Holland, 1986, pp. 85–111.

[41] J. D. Rietz and H. Giertz, "CCITT Signalling System No. 7 in AXE 10," *Ericsson Rev.*, vol. 59, no. 2, pp. 100–105, 1982.

[42] D. M. Rouse, R. J. Spires, and R. E. Wallace, "The Number 2 signal transfer point: An overview of the AT&T common channel signaling packet switch," in *Proc. 8th Int. Conf. Comput. Commun. (ICCC)*, P. J. Kühn, Ed., Munich, Germany, Sept. 1986, sess. B7, pp. 370–374, Paper 4.

[43] B. Simon, "Priority queues with feedback," *J. ACM*, vol. 31, no. 1, pp. 134–149, Jan. 1984.

[44] R. A. Skoog, "Performance and engineering of common channel signaling networks supporting ISDN," in *Traffic Engineering for ISDN Design and Planning*, M. Bonatti and M. Decina, Eds. Amsterdam, The Netherlands: Elsevier Science, 1988, pp. 415–424.

[45] ——, "Engineering common channel signaling networks for ISDN," in *Traffic Science for New Cost-Effective Systems, Networks and Services, Part 2 (ITC-12)*, M. Bonatti, Ed. Amsterdam, The Netherlands: Elsevier Science, 1989, pp. 915–921.

[46] ——, "Study of clustered arrival processes and signaling link delays," in *Teletraffic and Datatraffic in a Period of Change (ITC-13)*, A. Jensen and V. B. Iversen, Eds. Amsterdam, The Netherlands: Elsevier Science, 1991, pp. 61–66.

[47] A. Stoll and G. Wenzel, "Realization of Signalling System No. 7 in an ISDN," in *Proc. 11th Int. Switching Symp. (ISS)*, Florence, Italy, May 1984, vol. 3, sess. 31B, Paper 4.

[48] S. Sumita, "Performance analysis of interprocessor communications in an electronic switching system with distributed control," *Perform. Eval.*, vol. 9, no. 2, pp. 83–91, Apr. 1989.

[49] M. Villén Altamirano and B. Fontana, "Models to evaluate response times in single-processor systems and their application to a multiprocessor system," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services, Part 1 (ITC-12)*, M. Bonatti, Ed. Amsterdam, The Netherlands: Elsevier Science, 1989, pp. 402–411.

[50] Y. Watanabe and Y. Ikeda, "Traffic characteristics of PCR method for CCITT Signalling System No. 7," in *Proc. 10th Int. Teletraffic Cong. (ITC)*, Montréal, Canada, June 1983, vol. 1, sess. 3.3, Paper 3.

[51] G. Willmann, "Modelling and performance evaluation of multi-layered signalling networks based on the CCITT No. 7 specification," in *Teletraffic Science for New Cost-Effective Systems, Networks and Ser-*

vices. Part 2 (ITC-12), M. Bonatti, Ed. Amsterdam, The Netherlands: Elsevier Science, 1989, pp. 930–940.
[12] G. Willmann and P. J. Kühn, "Performance modeling of Signaling System No. 7," *IEEE Commun. Mag.*, vol. 28, pp. 44–56, July 1990.

Marcos Bafutto was born in Rubiataba, Brazil, in 1963. He received the B.E. degree in electrical engineering from the Federal University of Goiás in 1984, and the M.Sc. degree in electrical engineering from the Federal University of Uberlândia in 1989.

He joined Telecommunications of Goiás S. A. (Telegoiás), Goiânia, in 1984. Since 1989 he has been on leave at the Institute of Communications Switching and Data Technics of the University of Stuttgart working towards a Dr.-Ing. degree. His interests are the modeling and performance evaluation of signaling networks, mobile communications, and Intelligent Networks.

Paul J. Kühn (F'89) , for photograph and biography, see this issue, p. 378.

Gert Willmann (M'89) was born in Münnerstadt, Germany, in 1954. He received the Dipl.-Ing. degree in electrical engineering from the University of Siegen in 1983.

He joined Standard Elektrik Lorenz AG (SEL) in Stuttgart, where he was involved in the design of communication networks and in software engineering for digital switching systems. From 1985 to 1989 he was a Member of Scientific Staff at the Institute of Communications Switching and Data Technics of the University of Stuttgart, where he worked mainly in the modeling and performance analysis of common channel signaling networks and queueing network analysis. During 1990 and 1991 he worked for a research project at the University of Stuttgart and as a consultant in the area of common channel signaling. In 1992 he joined Alcatel SEL AG, Stuttgart, as a Supervisor in the Public Switching Systems Division. His current activities are in the area of traffic and performance modeling, analysis, and optimization of communication systems, and of communication network planning.