Capacity and Performance Analysis for Signalling Networks Supporting UPT

M. Bafutto *, P.J. Kühn
University of Stuttgart
Institute of Communications Switching and Data Technics
Seidenstraße 36, D-7000 Stuttgart 1, Federal Republic of Germany

Abstract

This work yields a generic modelling approach for the signalling load as a result of the various communications services, especially new services like UPT. A tool has been developed which provides the distinct loading of all hard- and software signalling network resources and on which a hierarchical performance analysis and planning procedures are based. The results of the analysis are used to evaluate the end-to-end performance of the call scenarios under consideration. A numerical example, including ISDN and PLMN scenarios, outlines the application of the tool to a hypothetical network.

1 Introduction

The Universal Personal Telecommunication (UPT) is a new service concept that improves subscriber mobility by deattaching the subscriber identity from the terminal identity. A UPT subscriber is identified by a unique Personal Telecommunication Number (PTN) and can register on specific access points across different networks (PSTN, ISDN, or PLMN). In an access point the UPT subscriber can initiate or receive calls and the charge is provided on the basis of the PTN. A list of services and facilities subscribed is contained in a user service profile.

The efforts to produce international standards for UPT has recently begun in CCITT and also in ETSI. According to the CCITT view [3], the elements of the UPT functional architecture comprises a Registration Point (RP) for storage of dynamic and static data, a Call Routing Point (CRP) for UPT call handling, and an Access Point (AP) through which the subscriber utilize the service.

A possibility for a fast and flexible UPT service deployment is offered by the emerging Intelligent Network concept [4, 5, 14]. The CCITT has already considered this approach and a mapping of the UPT architecture onto Intelligent Network architecture is currently under study.

The implementation of the UPT service depends on the capabilities and restrictions of the support network. In the signalling domain, for example, some aspects of the user and network signalling may be considered. The fixed wired networks have different user

---

*Under contract with Telegoias and supported by CAPES Brazil, grant 9532/88-15.

signalling structure, e.g., the "in-band" user signalling of PSTN, the "out-band" ISDN user signalling. In the mobile networks a user signalling protocol adjustment is necessary, and in some mobile networks, like GSM, the use of smart card as Subscriber Identity Module requires new terminal functions [8]. On the network signalling plane, it was concluded that the Mobile Application Part (MAP) of CCITT Signalling System No. 7 was not in line with current MAP specifications in practical use, thus representing an obstacle for the UPT service implementation, this fact led to a proposal of updating or even deleting the recommendation CCITT Q.1051 [3].

Despite of the support network, the deployment of the UPT procedures related to mobility, call handling and service management require additional information to be transported by the signalling network. The introduction of the UPT service increases the signalling network load, and may lead to a performance degradation of the signalling network, therefore affecting not only the UPT quality of service parameters, but also the services already offered by the network. This work uses a modelling approach to derive a tool concept for the planning of the signalling networks supporting UPT.

2 The Signalling Network

A common channel signalling network [2, 9, 10] is composed of signalling points (SP), signalling control points (SCP) and signalling transfer points (STP) interconnected by links, and supports the exchange of information related to call/connection control, distributed application processing, and network management.

In the signalling network domain, a telecommunication service can be viewed as a number of packets transmitted to and from the related nodes. Changes in the traffic intensity of a particular service, introduction of new services or temporary outages usually cause variations of the network load and, therefore, of the quality of service parameters. The various network components are typically affected in the following way:

- *Links:* The offered signalling link load per call may change drastically with the introduction of new services.
- *Signalling Points:* The increasing signalling traffic load requires more processing capacity from the network elements. Since this additional load may not be homogeneously distributed over all processes within an SP, only some processors may become overloaded.
- *Service Control Points:* Since the capacity of an SCP is determined by the number and types of transactions and 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 functional structure of Signalling System No. 7 is divided into the Network Service Part (NSP) and various User Parts (UPs). The NSP consists of the Message Transfer Part (MTP) and the Signalling Connection Control Part (SCCP), while the UPs are the Telephone User Part (TUP), the Integrated Services Digital Network User Part (ISUP), the Data User Part (DUP), and the Transaction Capabilities (TC). The TC can be further subdivided into the Transaction Capabilities Application Part (TCAP) and the Intermediate Service Part (ISP), which is still empty.

The MTP provides a simple datagram service, and its addressing capabilities are limited to the identification of a certain UP in one specific node. The SCCP extends the MTP capabilities in order to identify a subsystem of a UP or to translate an address that does not contain MTP routing information (e.g., a Global Title).

The ISUP offers the signalling functions that are necessary to support voice and non-voice applications in an ISDN network. The TUP and DUP functions are assumed as part of the ISUP and are not considered in the rest of this work. Finally, the TC provide a set of capabilities in a transaction-based and connectionless environment that support applications which require remote procedure calls or database queries.

3 Modelling Framework

The modelling methodology presented in [15] is used as baseline. Each submodel is derived directly from the CCITT functional specifications with consideration of internal mechanisms such as segmenting/forking of messages and scheduling strategies, thus, reflecting the internal behavior of the underlying blocks. Hence, this approach is relatively independent of specific implementations leading to a generic model of the signalling network composed of generic submodels. The generic submodels are finally transformed into realistic models where the actual assignment of functional processes to real processors are taken into account.

The principles of this methodology can be briefly explained using the TCAP block as an example of derivation of a generic submodel. According to [2, Figure A-2a/Q.774], the TCAP is composed of two subblocks: the Transaction Sub-layer (TSL) and the Component Sub-layer. The Component Sub-layer consists of the Dialog Handling (DHA) and Component Handling. The Component Handling is further subdivided into the Component Coordinator (CCO) and the Invocation State Machine (ISM). From this, a 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, i.e., different message chains. As an example, the message chain corresponding to an outgoing Dialog Begin message containing an Invoke component is listed bellow:

- A primitive "TC-BEGIN-req" received from TC-User is processed by DHA. The invoke components with same Dialog ID are requested from CCO through a "Component-req" indication.
- The CCO processes the "Component req" indication and two outputs are generated (fork): the "Operation Sent" is directed to ISM and the "Requested Components" to DHA.
- The DHA receives the "Requested Components" and composes a "TR-BEGIN-req" primitive to TSL.
- Under reception of the "Operation Sent", the ISM starts an invocation timer. No output is generated (Sink). The case of time-out can be modelled as a message branching with the branching probability given by the time-out probability.
- The TSL processes the "TR-BEGIN-req" and requests the service of the Signalling Connection Control Part (SCCP) through a "N-Unidata" primitive.

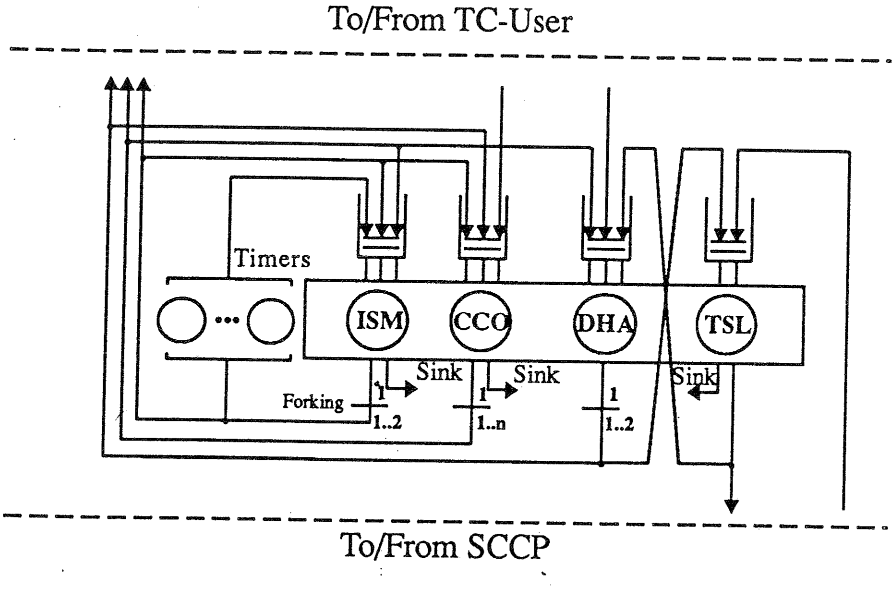

Figure 1: Generic submodel for the TCAP block

The final model, shown in Figure 1, is obtained considering the set of all possible message chains for the TCAP block. The models for the another functional blocks can be obtained in the same way.

MTP Level 1 is simply modelled as an infinite server with a service time representing the signalling link propagation delay. The modelling 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. Therefore, the approach included in the CCITT recommendations is adopted, where the corresponding queueing delay formulas are given explicitly.

A better overview of a complete model for the Levels 3 and 4 can be obtained by representing the individual submodels in a reduced form, as depicted in Figure 2. In that case it was assumed that each functional block is implemented in a single real processor. In the MTP the identified processes are the Message Discrimination (HMDC), Message Distribution (HMDT), and Message Routing (HMRT). The ISUP contains the Call Processing Control Incoming (CPCI), Call Processing Control Outgoing (CPCO), Message Distribution Control (MDSC), and Message Sending Control (MSDC). The processes of the SCCP are the Connection Oriented Control, Connectionless Control and Routing Control and since they are used in both directions, transmitting and receiving, they were split in one processing phase for each direction. For more details about the individual models the reader is referred to [15]. The interfaces between the functional blocks are represented by the interprocessor communication subsystems and are modelled as infinite servers.

In order to embed all these models in a realistic environment, it is necessary to extend the model by adding traffic sources representing the traffic generated by the users, traffic sinks, response times of exchanges and users, database access delays, etc.

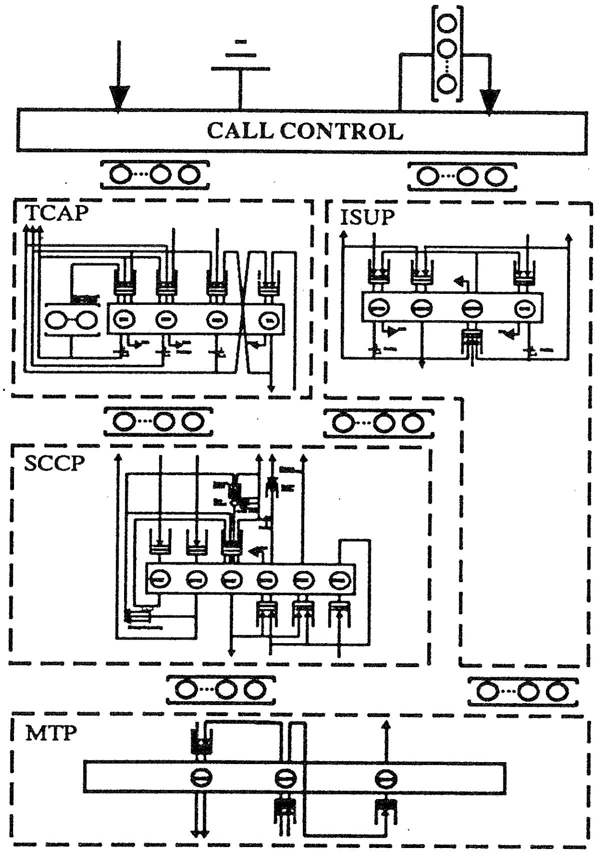

Figure 2: Submodels for the functional blocks in a signalling point

4 Analysis Through Decomposition and Traffic Aggregation

For the analysis of the signalling network, we just start with a message flow analysis from given data about the network configuration, routing plan, mix of services types, and a traffic origination-destination matrix. This message flow analysis yields the message flow rates on each transmission or processing resource partitioned to all message types. From message length and processing time distributions, the resource utilization follow straightforwardly.

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 assumption is valid if the interactions between the subsystems are largely dominated by the local interactions inside each subsystem. This can be assumed to be sufficiently satisfied in our case. The network is decomposed into link sets, SPs, and STPs. Then, in the second decomposition step, the link sets are further decomposed into single signalling links, whereas the SPs and STPs are decomposed into their submodels, i.e., MTP Level 3, SCCP, ISUP, and TCAP.

The basic idea behind the signalling 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 suitable point processes, e.g., Poisson processes.

The application of the principles above allow the analysis of each system in isolation. For the MTP Level 2 the formulas given in the CCITT recommendations are used. The performance of models for the upper Levels depends on additional factors like, e.g., the priority and the distribution of the message processing times of each process.

Although an exact analysis of the models of Level 3 and 4 may become quite complex, a mean value analysis is still possible in most cases. Previous work has already been done in the analysis of priority queue systems in which a message may feedback and change priority after having been served [6, 11, 13]. An algorithm considering also bulk arrivals, forking and branching of messages, and different preemption strategies is available in [12]. The application of this algorithm results in a system of linear equations for the mean sojourn times of all message chains in each process of the model.

The global performance results, e.g., message transfer time or a call setup delay is obtained by composition of the submodel results.

5 A Planning Tool

Using this methodology, a signalling network planning tool has been implemented. Its input data are the topology, the sequence of processing phases visited by a message when passing through a signalling point (message chains) with the corresponding processing times, the scenario descriptions, the traffic matrix, and the routing strategy. The execution of the tool can be divided into two main steps:

- *Message Flow* – With the input data, a global traffic flow analysis is performed and the mean call attempt loading for each type of scenario is calculated for the Link set, for the SPs and for the STPs. This is performed by routines working with the topology, traffic matrix and routing strategy. At this phase the output information concerning the utilization of the processes is generated. 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 already given with respect to each scenario at this stage of the analysis.
- *Performance Analysis* – The performance analysis for all nodes of a large network may represent a considerable expense of computer resources and the network designer is usually interested in the performance characteristic of some critical paths. With these considerations, the analysis is carried out for chosen paths. Low traffic approximations, such those presented in [1], may be used as an approach for the cases where the results for all nodes of a large network are desired.

  In a multivendor environment, the physical implementation of the functional blocks among the processors may not be the same for all SPs and STPs. For example, in a distributed SP architecture each functional block may be implemented in isolated processors, while in a centralized one, various functional blocks may share the same processor. The tool provides the user with the flexibility of defining his own architecture with arbitrary distribution of functional blocks among processors.

  The priority assignment for the processing phases in a processor is one of the parameters that determine the sojourn time for a message type. According to the priority assignment and traffic load, the differences between the sojourn time for distinct message types can be significant, in such a way that for one of them the processor can be considered as overloaded.

With the remarks above, the user is required to provide for each SP, SCP, and STP along the path the following information:

- The number of processors and the distribution of the functional blocks between them.
- For each processor, the number of units and the priority assignment of the implemented processing phases.

These input data and the individual message arrival rates for each submodel obtained from the traffic flow study are used to carry out a performance analysis for the submodels. In the cases where there are more than one processing unit available, it is assumed that the incoming traffic is homogeneously distributed between the units. The mean sojourn time for the individuals message chain are calculated using the algorithm contained in [12] as baseline.

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 the path through the network. As in the message flow phase; this is again done by routines working with the network topology and routing strategy. With the end-to-end transfer time of a particular message sequence, it is possible to evaluate response delay parameters, such as connection set-up delay, data transfer delay and database query delay.

6 UPT Case Study

A simplified case study is provided to demonstrate the capabilities of the described planning tool. In order to do not increase the complexity and the size of the traffic matrix a network with few nodes was chosen as example. This fact does not represent a hindrance for the analysis of a real network containing a much larger number of nodes, in that case the traffic matrix data could be read from a magnetic media.

The topology of the Figure 3 is considered. On the highest hierarchical level (level 1) there are three interconnected STPs. The next hierarchical level (level 2) contains 2 transit SPs of the ISDN subnetwork and 1 Gateway Mobile Switching Center (GMSC) for the PLMN. Under the transit SPs there are three SPs of the lowest level (level 3). Three Mobile Switching Centers MSCs are linked to the GMSC. 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 kbit/s, and the propagation delay is 5 ms. The routing is assumed to be strictly hierarchical.

The information related to the UPT Registration Point (RP) functionality is stored in an SCP connected to the STP identified by the code "100". For the PLMN, the Visiting Location Register (VLR) is integrated in the MSC and the Home Location Register is centralized in an SCP linked to the STP with code "300".

The ISDN and PLMN voice services are in operation in the example network. The typical scenarios for these services can be classified into three subgroups: normal call, subscriber busy, and no answer. Each service comprises 70% successful calls, 20% subscriber busy, and 10% no answer. The PLMN total traffic corresponds to 10% of the ISDN total traffic.

The composition of the mobility related signalling scenarios in the PLMN requires some assumptions to be made, e.g., at location update the user is identified by his International Mobile Subscriber Identity (IMSI) instead by the Temporary Mobile Subscriber

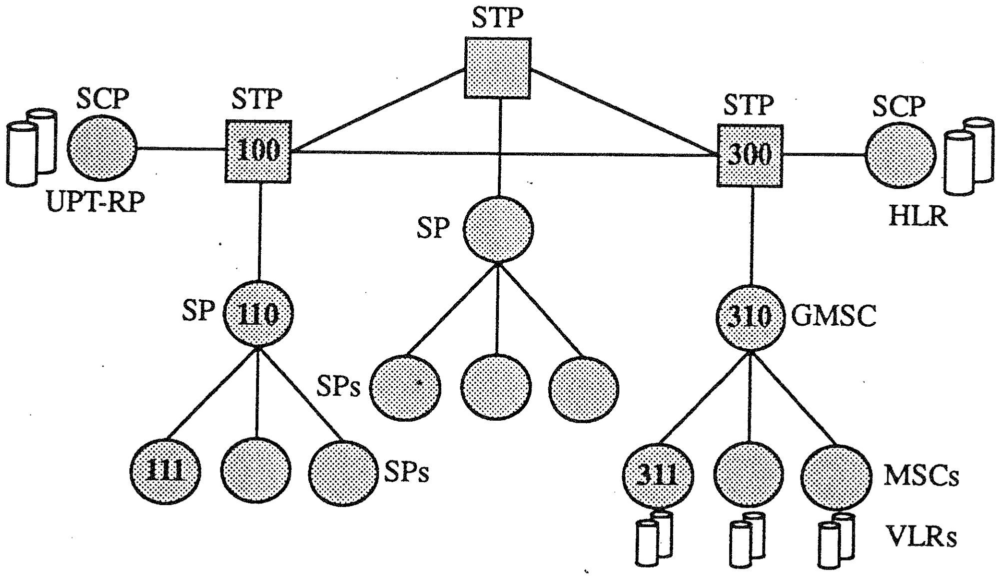

Figure 3: Structure of the example network

Identity (TMSI) avoiding an additional query to the old VLR, the VLR has the necessary information to manipulate an outgoing call, the HLR knows the Mobile Subscriber Roaming Number (MSRN), etc. The consideration of all transactions implies in a spectrum of messages generally with different message lengths. In order to simplify the example, the transaction related messages were divided into two subgroups and according to the amount of contained information the messages are classified into short and long length. A database query in one of the SCPs consumes 200 ms. The message types and their corresponding lengths are listed in Table 1.

Table 1: Messages considered in the example

Message	Designation	Length [byte]		
		ISUP	PLMN	UPT
IAM	Initial Address Message	59	65	70
ACM	Address Complete Message	17	17	17
ANM	Answer Message	15	15	15
REL	Release Message	19	19	19
RLC	Release Complete Message	14	14	14
INV(A)	Invoke Message (Short)	-	20	25
RES(A)	Response Message (Short)	-	30	35
INV(B)	Invoke Message (Long)	-	40	45
RES(B)	Response Message (Long)	-	60	65

The traffic between the ISDN and PLMN subnetworks is assumed to be unbalanced and characterized as follow: 5% of the ISDN traffic is directed to the PLMN and 95% of the PLMN traffic is homogeneously distributed between the ISDN nodes.

The generated traffic corresponding to the support of the mobility related procedures is a function of some network characteristics, e.g., cell size, ground topology, subscriber moving speed, etc. In this example a simplified stationary approach is considered, with a PLMN subscriber generating a location update request for each call and a handover being performed in 30% of the calls. Subsequent handover are not taken into account.

With the introduction of the UPT concept, it is also necessary to differentiate between the calls destinated to/originated from a UPT user currently registered in an ISDN or PLMN terminal. A percentage of the ISDN or PLMN traffic is substituted by the UPT service. The traffic between UPT users is considered to be negligible.

It is assumed that the UPT functions are merged into normal ISDN/PLMN call setup, e.g. by including new UPT information in the related messages. The user is authenticated on registration, deregistration, incoming and outgoing calls. The storage of the authentication information is centralized in the UPT-SCP.

The comparison between the different possibilities of physical implementation of the SS7 functionalities is beyond the scope of this work. In this example we have arbitrarily chosen an architecture where the MTP and SCCP are implemented in a single processor. The ISUP and TCAP functions are assumed to be performed by isolated processors.

The priority strategy of the processors is non-preemptive. The priority assignment for the processes in a decreasing order of priority is:

- ISUP processor: MSDC, CPCO, CPCI, and MDSC.
- TCAP processor: TSL, DHA, CCO, and ISM.
- MTP/SCCP processor: HMDC, HMDT, HMRT, SCLR, SCLT, SCRR, and SCRT.

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 the DHA, and 1.5 ms, 1 ms, and 0.5 ms for the CCO, TSL and ISM, respectively. All processes of the SCCP have a processing time of 1 ms. For the ISUP, the processing time is 1 ms for the CPCI and CPCO, and 0.5 ms for the MSDC and MDSC. In the MTP, the delay for a message in the HMDT and HMRT is 1 ms and in the HMDC 0.5 ms.

Before the evaluation of the network parameters it is interesting to check the accuracy of the adopted modelling approach through a simulation study. A good agreement between analysis and simulation results suggests that the assumptions of the analysis methodology do not introduce significant errors. Simulations were performed with 10 subsequent replications simulating 90 seconds of a MSC. The warm-up phase comprised 60 seconds. The input data was obtained from the output data of the message flow phase. This information were taken as the boundary conditions for the signalling point analysis, i.e., the arrival rate of all messages destinated to, originated from, and passing through the underlined node. The analysis and simulation results for the sojourn time in a MSC of an incoming TCAP Begin Message with an Invoke Component and an incoming TCAP End Message with a Result Component is shown in Figure 4. The results are depicted with the call attempt loading normalized with respect to the maximum call attempt loading of the underlined MSC.

The difference between the sojourn times of the messages result from the fact that the incoming TCAP Begin Message is not processed by the ISM process. The assignment of the lowest priority to the ISM implies that a message in the ISM must wait until the queues of all the remaining processes are empty, and the probability that it occurs decreases with the increase of the load. This example illustrates a case where a processor is practically

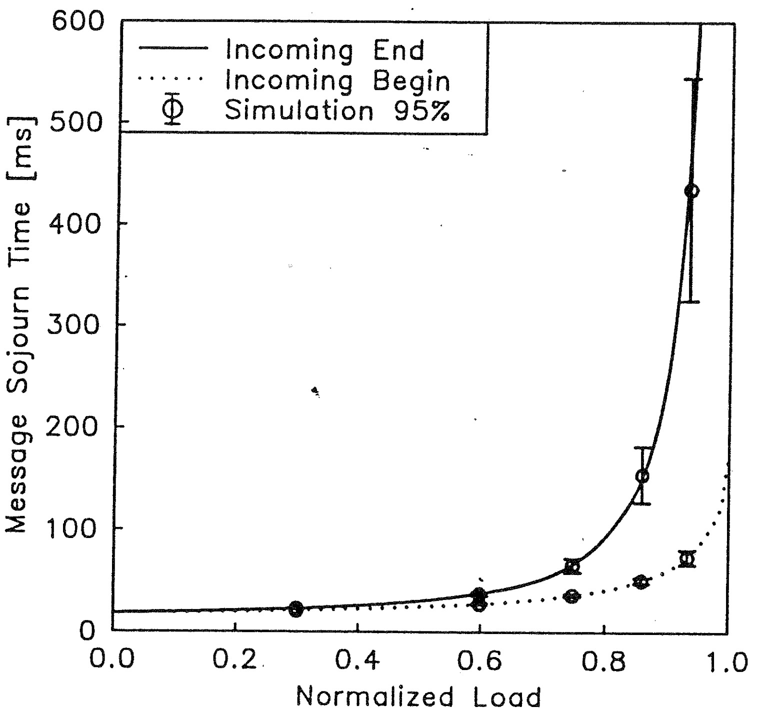

Figure 4: Analysis and simulation results for a MSC

considered as overloaded for a particular message type, while for other message types it continues to operate normally.

The total CPU time for the simulation run of a single load case was about 75 minutes, contrasting with the 50 seconds consumed for the analysis. The computer resources required by the simulation highlights the constraints involved with application of the simulation technique to large signalling networks.

Firstly, the impact of the introduction of the UPT concept on both networks is studied. The UPT concept is introduced in a limited basis and the parameter of interest is the connection setup time between ISDN and PLMN. The results for a substitution of 3% and 5% of the voice service of the ISDN and PLMN are depicted in Figure 5. The call attempt loading is normalized with respect to the maximum call attempt loading of the network carrying only ISDN and PLMN voice services.

The introduction of the UPT concept and the support of the mobility related functions like, e.g., call routing, registration, authentication, etc., represent an additional load for the signalling network in the form of transactions to retrieve the necessary information. This load is not homogeneously distributed on the network and varies with the amount of introduced UPT services. In the cases of no UPT and 3% UPT service, the bottlenecks are located in the ISUP block of the transit ISDN-SP and in the ISUP block of the GMSC, respectively. For an introduction of 5% UPT traffic, however, the overload situation is observed in the TCAP block of the UPT-SCP. The quality of service parameters of the already existing ISDN and PLMN services are also affected by the UPT introduction.

The complexity of these new scenarios and the corresponding effect on the call routing, the number of data base queries, the signalling network load, and the setup times suggest the use of alternative architectures considering factors like, e.g., data base to data base interactions, integration of data bases, allocation of IN functionality into the MSC [7].

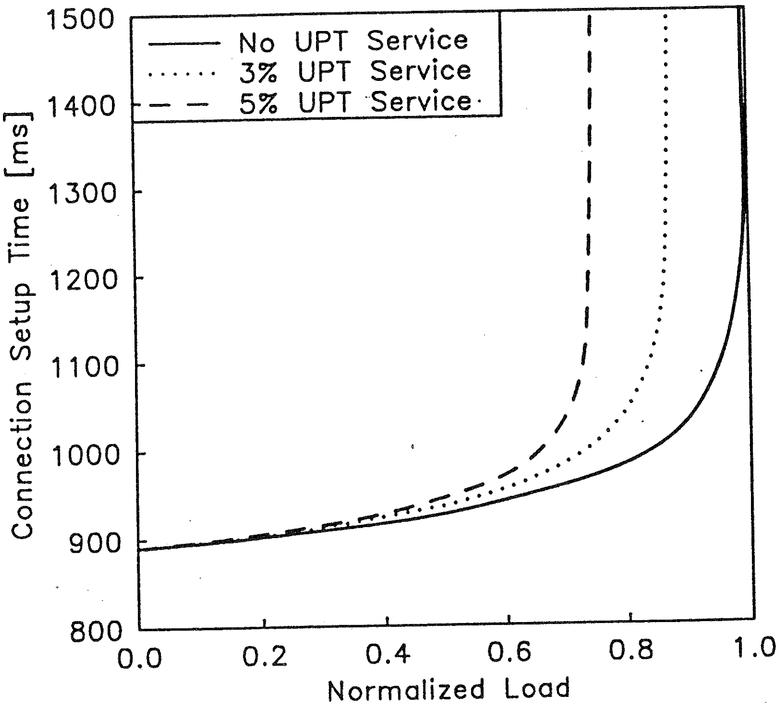

Figure 5: Influence of the UPT concept on the ISDN–PLMN call setup delay

The individual message transfer time between two nodes yields the computation of the delay involved in the exchange of any message sequence. With this information, there are various parameters of interest that can be evaluated. The results corresponding to the connection setup time for a call originated from an UPT user registered in the ISDN to an PLMN user is shown in Figure 6.

The shape of the graphics in Figure 6 differs from those obtained for the connection setup between ISDN and PLMN. This difference is explained by the larger number of network elements involved in the UPT scenario.

7 Conclusion

The results show that quality of service parameters of the future networks will be strongly influenced by the performance of the signalling network. The introduction of new services which require database interactions, such as UPT services and mobile communication services, will have a significant impact on signalling network performance aspects, e.g., the processing load of signalling points and the end-to-end message transfer delays.

The presented modelling approach, with the consideration of physical implementation aspects, like distribution of the processes among processors and priority assignment, cover important characteristics present in a multivendor environment. This modelling framework has been implemented in a tool suitable to support the planning of new signalling networks according to given service, load and grade of service figures, or to detect bottlenecks or possible deficiencies in case of resource outages.

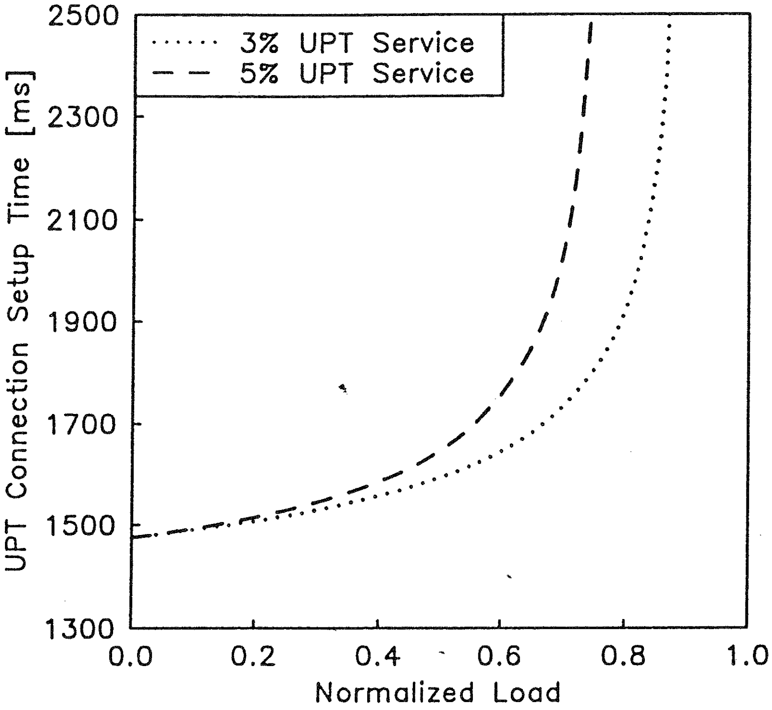

Figure 6: Connection setup delay from an UPT user in ISDN to an PLMN user

8 Acknowledgment

The authors would like to thank Dr. Paterok for providing the algorithm for the M/GI/1 System with Feedback and to Dr. Willmann for the many helpful discussions. The programming efforts of Mr. Kandulski are greatly appreciated.

References

[1] M. Bafutto, P.J. Kühn, G. Willmann, 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*, (Ed. Paul W. Bayliss), IOS Press, Washington, pp. 368–379, 1992.

[2] CCITT, *Blue Book, Volume IV, Fascicles VI.7–VI.9*, "Specifications of Signalling System No. 7", Recommendations Q.700–Q.795, International Telecommunication Union, Geneva, 1989.

[3] CCITT, *Study Group XI, Report COM XI-R 60*, Meeting Held in Geneva on 8–12 April 1991.

[4] M.C. Ciancetta, R. Lavagnolo, T. di Stefano, "An Approach to Universal Personal Telecommunication (UPT) Implementation in Intelligent Networks", *Proceedings of the 1992 International Zurich Seminar on Digital Communications – Intelligent Networks and their Applications*, Zurich, Switzerland, pp. 333–347, March 1992.

[5] J.C. Dang, C. Vernhes, B. Chatras, "IN as Platform for UPT: Constrains and Requirements", *Proceedings of the 2nd International Conference on Intelligence in Networks*, Bordeaux, France, pp. 79–82, March 1992.

[6] D.H.J. Epema, "Mean Waiting Times in a General Feedback Queueing Model with Priorities", *Performance90*, (Editors P.J.B. King, I. Mitrani, and R.J. Pooley), North-Holland Publishing Company, Amsterdam, pp. 221–235, 1990.

[7] E. Guarene, M.C. Ciancetta, T.D. Stefano, "IN and GSM Integration for Personal Communication Services Provision", in *Intelligent Networks – The Path to Global Networking*, (Ed. Paul W. Bayliss), IOS Press, Washington, pp. 290–299, 1992.

[8] H.P.J. Hecker, J. Hegeman, W.R. Mol, M.J.J. van Nielen, "The Application of the IN-concept to Provide Mobility in Underlying Networks", *Proceedings of the 2nd International Conference on Intelligence in Networks*, Bordeaux, France, pp. 95–100, March 1992.

[9] B. Jabbari, "Common Channel Signalling System Number 7 for ISDN and Intelligent Networks", *Proceedings of the IEEE*, vol. 79, no. 2, pp. 155–169, February 1991.

[10] A.R. Modarressi, R.A. Skoog, "Signaling System No. 7: A Tutorial", *IEEE Communications Magazine*, vol. 28, no. 7, pp. 19–35, July 1990.

[11] M. Paterok, O. Fischer, "Feedback Queues with Preemption-Distance Priorities", *ACM SIGMETRICS Performance Evaluation Review*, vol. 17, no. 1, pp. 136–145, May 1989.

[12] M. Paterok, "Warteschlangensysteme mit Rückkopplung und Prioritäten", *Arbeitsberichte des Instituts für Mathematische Maschinen und Datenverarbeitung*, Band 23, no. 12, Erlangen, Germany, October 1990.

[13] B. Simon, "Priority Queues with Feedback", *Journal of the ACM*, vol. 31, no. 1, pp. 134–149, January 1984.

[14] J. Vandenameele, J.B. Thieffry, H. Decuypere, "UPT, a New Dimension in Telecommunications provided by IN", *Proceedings of the 1992 International Zurich Seminar on Digital Communications – Intelligent Networks and their Applications*, Zurich, Switzerland, pp. 41–54, March 1992.

[15] G. Willmann, P.J. Kühn, "Performance Modeling of Signaling System No. 7", *IEEE Communications Magazine*, vol. 28, no. 7, pp. 44–56, July 1990.