Capacity Planning and Performance Analysis of Signalling Networks for the Support of Mobile Communication and Intelligent Networks

M. Bafutto *, M. Schopp, P.J. Kühn

University of Stuttgart
Institute of Communications Switching and Data Technics
Seidenstr. 36, 70174 Stuttgart, Germany
Phone: +49 711 121 2481, Fax: +49 711 121 2477
E-Mail: schopp@ind.e-technik.uni-stuttgart.de

Connection management in the ISDN and B-ISDN is based on a common channel signalling network which provides fast and reliable transfer of signalling messages. With the introduction of new service features such as mobility management in mobile communication networks, authentication procedures for UPT or Intelligent Network services, the signalling traffic will drastically increase. Signalling channels and protocol processing functions have to be properly engineered to meet the load and network performance requirements. Due to the layered protocol architecture, the various signalling scenarios for different services, and the spacial extension of signalling networks, engineering and analysis issues turn out to be rather complex.

A methodology is proposed by which the physical structure, the protocol architecture and the signalling traffic is modelled directly from the physical network structure, the standardized protocol specification and signalling traffic parameters. A generic modelling technique is proposed which leads to hierarchical queueing networks including multi-class traffic streams, priorities, message routing and protocol functions.

Two types of analysis are applied: traffic flow analysis and traffic performance analysis which form the basis for a capacity and QOS related sizing of signalling network resources. The traffic flow analysis yields the resulting loading factors of all transmission and processing resources. For the performance analysis, a decomposition approach is encountered. Results of that analysis are directed to performance figures such as connection setup delay, handover delay or post-selection delay when IN functions lead to database accesses and which are directly visible to the user or the network operator. The contribution concludes with a case example which underlines the feasibility of the methods.

1 Introduction

The load of the signalling network is being increased through the introduction of a number of new services represented by supplementary services, IN applications, mobile communication services, UPT, etc. This additional load may lead to a performance degradation of the signalling network, which, in turn, affects not only the quality of the new services, but also the services already offered by the network.

---

* Under Contract with Telegoiás and supported by CAPES Brazil, grant 9532/88-15

In this paper, a generic modelling methodology for the signalling load as a result of the various communication services is presented ([10]). The model is obtained considering the signalling system No.7 protocol functions specified by the ITU-T ([5]), as well as the information flow through these functions. The approach allows the derivation of virtual processor models, which in a further step are mapped onto the physical implementation according to individual architectures. This allows the analysis of signalling networks in a multivendor environment.

Using these principles, a signalling network planning tool concept has been developed ([1]). Two types of analysis are applied: traffic flow analysis and traffic performance analysis which form the basis for a capacity and QOS related sizing of signalling network resources. The traffic flow analysis yields the resulting loading factors of all transmission and processing resources. For the performance analysis, a decomposition approach is encountered([3]). Results of that analysis are directed to performance figures such as connection setup delay, handover delay or post-selection delay ([4]) when IN functions lead to database accesses and which are directly visible to the user or the network operator. The contribution concludes with a case example which underlines the feasibility of the methods.

2 Modelling of the Protocol Architecture

The analysis of a real network initially requires a modelling methodology able to represent the internal processes of the protocols. The consideration of all mechanism of a complex protocol may create mathematical difficulties making the resulting model intractable. The modelling methodology adopted here, which uses decomposition and aggregation techniques, follows [10]. Each submodel is derived directly from the ITU-T functional specifications [5] including internal mechanisms such as segmenting/forking of messages, thus reflecting the internal behaviour of the underlying functional blocks. The basic idea is the observation that, in the 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. The principles of this methodology can be briefly explained using the TCAP (Transaction Capabilities Application Part) block as an example. According to [5], the TCAP is composed of two subblocks: the Transaction Sub-Layer (TSL) and the Component Sub-Layer (CSL). The CSL consists of the Dialogue 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, i.e., different message chains. As an example, the message chain corresponding to an outgoing Dialogue Begin message containing two Invoke components is described bellow and depicted in Figure 1.

- The primitive *TC-BEGIN req* received from the TC User is processed by DHA. Any Invoke components with the same Dialogue ID (in this case two components) are then requested from the CCO through a *request components* signal;

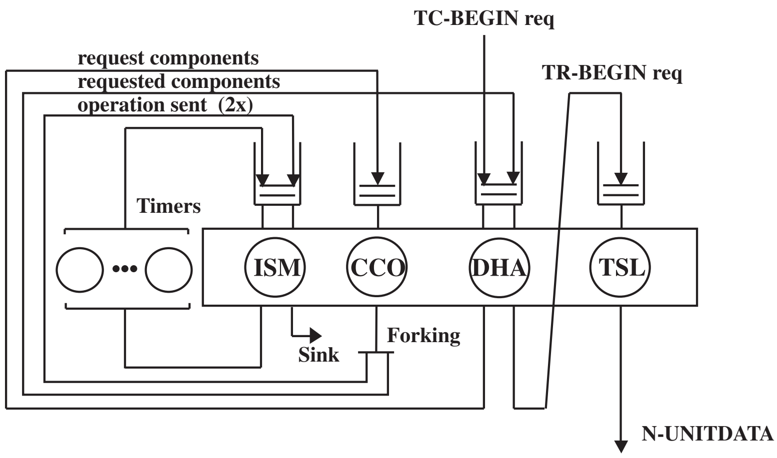

Figure 1: Message chain for the *Dialogue Begin* message containing two *Invoke* components

- 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 time-out can be modelled by a message branching with the branching probability given by the time-out probability;
- When the DHA receives 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 Signalling Connection Control Part (SCCP) through an *N-UNITDATA* primitive.

The TCAP functional block is not restricted to the Dialogue Begin message, and 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 depicted in Figure 1 through additional chains. The full model, which is depicted in Figure 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 signalling network protocol architecture (MTP Level 3, SCCP, and ISUP) are obtained in the same way. The 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. Instead the CCITT (ITU-T) approach and the corresponding queueing delay formulas are adopted ([5]).

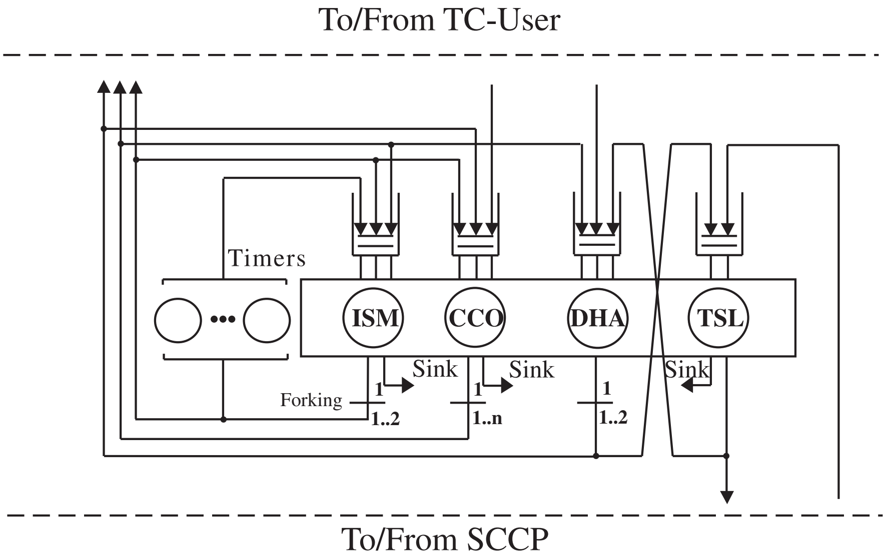

Figure 2: Generic model for the TCAP block

In order to provide an overview of a complete SP model, the full model of an SP is depicted in a reduced form in Figure 2. A detailed description of all submodels may be found in [10].

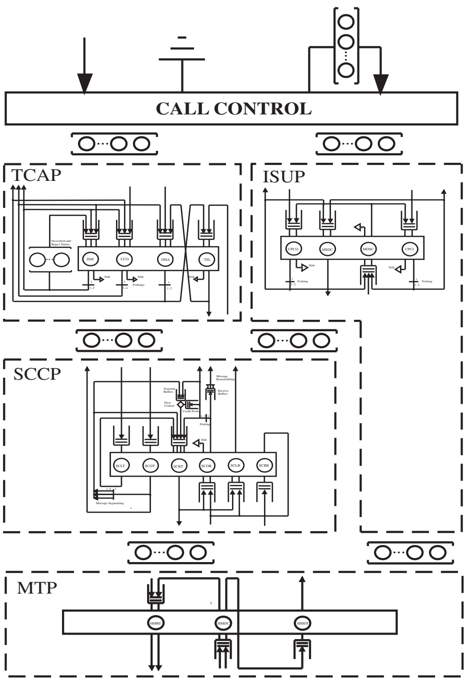

Figure 3: Generic model of a signalling point

3 Implementation Architecture

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 centralized one, various functional blocks might share the same processor. Therefore, in order to account for vendorspecific 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 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 Figure 4.

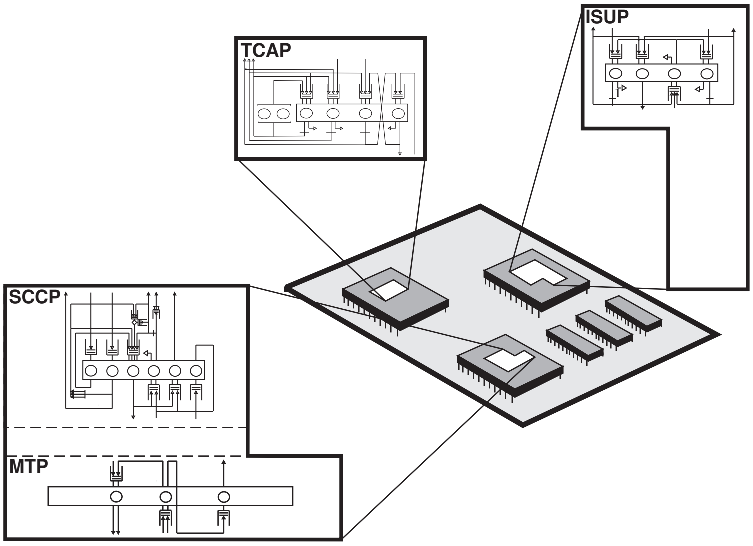

Figure 4: Mapping onto a hypothetical implementation: ISUP and TCAP in two different processors, MTP and SCCP together in another processor.

4 Analysis

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 ITU-T 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 [9,7]. An algorithm considering also bulk arrivals, forking and branching of messages, and different pre-emption strategies is available in [8]. 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 Case Study

A simplified case study ([2]) is provided to demonstrate the capabilities of the planning tool. The topology of Figure 5 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 5ms. The routing is assumed to be strictly hierarchical.

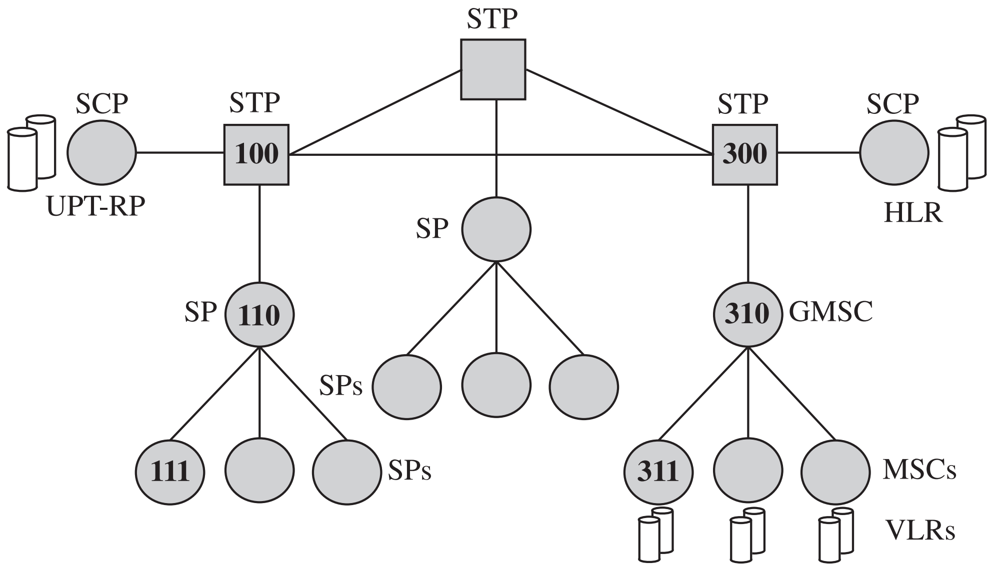

Figure 5: UPT case study: example network

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 Visitor 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.

Several signalling scenarios have been considered: call set-up, call release, location update, handover, registration and authentication. Call set-ups are further distinguished in *PLMN originated* / *destinated*, *UPT user originated* / *destinated*, and *successful* / *subscriber busy* / *no answer*. The assumed scenario rates, the traffic matrices, the message types and their lengths, the mapping onto the physical implementation and the parameters of the physical implementation (priority assignment and processing times) can be found in [2].

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 6. 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 results corresponding to the connection setup time for a call originated from an UPT user registered in the ISDN to a PLMN user is shown in Figure 7.

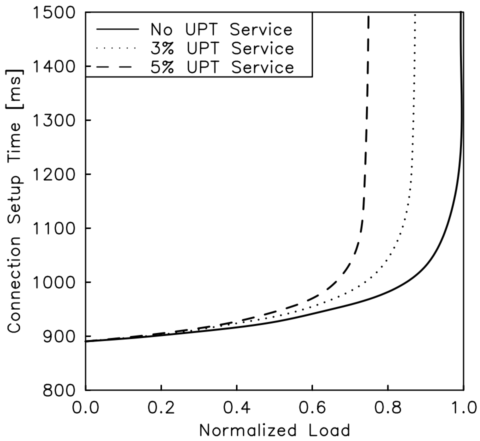

Figure 6: Influence of the UPT service introduction on the ISDN-PLMN connection set-up delay

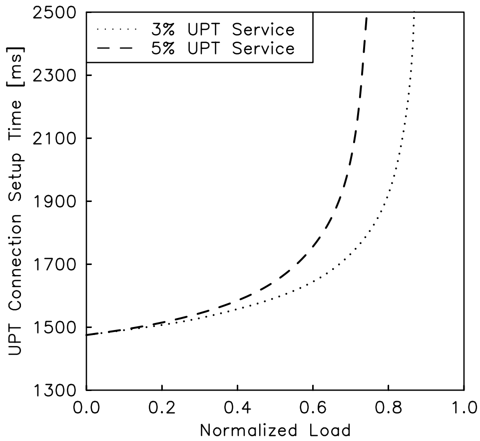

Figure 7: Connection set-up delay from an UPT user in ISDN to a PLMN user

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.

6 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.

7 References

[1] Bafutto, M., Kühn, P. J., Willmann, G., Zepf, J.: A Capacity and Performance Planning Tool for Signalling Networks based on CCITT Signalling System No. 7, *Intelligent Networks - The Path to Global Networking*, pp. 368-379, P. W. Bayliss (Editor), IOS Press, Washington, 1992

[2] Bafutto, M., Kühn, P. J.: Capacity and Performance Analysis for Signalling Networks Supporting UPT, *Proc. of the 8th ITC Specialist Seminar on Universal Personal Telecommunication*, Santa Margherita Ligure (Genova), Italy, pp. 201-213, October 1992

[3] Bafutto, M., Kühn, P. J., Willmann, G.: Capacity and Performance Analysis of Signaling Networks in Multivendor Environments, *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 3, pp. 490-500, April 1994

[4] ITU-T *Recommendations E.750 Series: Traffic Engineering Aspects of Mobile Networks*, International Telecommunication Union, Geneva, 1993

[5] ITU-T *Recommendations Q.700 Series: Specifications of Signalling System No. 7*, International Telecommunication Union, Geneva, 1993

[6] ITU-T *Recommendations Q.1200 Series: Intelligent Network*, International Telecommunication Union, Geneva, 1993

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

[8] Paterok, M.: Warteschlangensysteme mit Rückkopplungen und Prioritäten, *Arbeitsberichte des Instituts für Mathematische Maschinen und Datenverarbeitung*, Band 23, Nummer 12, Friedrich Alexander Universität Erlangen Nürnberg, Germany, October 1990

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

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