Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics

Performance Evaluation, Modelling and Optimisation of Call Processing in Broadband Networks

Sándor Székely

Summary of Ph.D. Dissertation

Advisors:

Sándor Molnár, Ph.D.
Béla Frajka, Ph.D.

High Speed Networks Laboratory
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics

Budapest, Hungary
2004

1. Introduction

The beginning of the 1990's brought new technologies in the telecommunication networks. Asynchronous Transfer Mode (ATM) has been chosen as the transmission technology of broadband networks. The standardisation of ATM is practically finished today, the implementations are ready and the ATM equipments are largely introduced by users and service providers. At the end of the 1990's the Internet Protocol (IP) became the fastest growing network layer protocol that is applicable over any data link layer. The convergence of these two technologies became reality at the beginning of the year 2000, but a couple of months later the telecommunication market entered the deepest recession ever seen. In the current marketplace the service providers must improve network management to reduce their operation costs. While in the last years the ATM network service providers offered only permanent virtual circuit connections to the customers, ***today there is an increasing interest to offer switched virtual circuit connections to end users.*** This later solution is based on the use of *signalling protocols*, and does not need hardware extension (additional cost) of existing equipment. One of the advantages of ATM is the ability to set up and tear down virtual connections dynamically between source and destination. There are many factors that may influence the number of switched connections that a network can accept and the rate at which they can be accepted. Both of these performance criteria are influenced by the signalling performance of the User-Network Interface (UNI).

With the increased capability and complexity of ATM networks, call processing and especially signalling procedures become more and more complex which causes a performance degradation of the signalling networks. Moreover, the size of the signalling messages is larger and larger. Therefore the existing models for SS7 networks, e.g. [Baf93], do not represent correctly the behaviour of the signalling processing in broadband networks. ***There is a need for new signalling models.*** The basic *performance metrics* for ATM signalling are described in [ATMF00], while *performance benchmarking* of ATM signalling software is presented in [Kaus97] and [Nie97], showing the first results obtained on switches from four vendors in a number of hardware, software and network configurations. [Nie97] has some early results on multiple host connections as well, but there is more work to be done. Although the importance of the signalling performance of ATM networks has been recognised as a potential bottleneck in [Gel97], very few papers addressed the congestion situation in switches due to signalling message flows.

The signalling performance of an ATM switch is determined by the speed it can process signalling traffic. In a switch this is limited by its architecture and its call processing capacity. During my Ph.D studies I have investigated the architectures and features of existing signalling protocols and I have collected and processed industrial and academic papers related to this field. Based on these studies *I have concluded that very few research studies are focused on practical experiences with broadband signalling networks, and even those, they do not investigate the basic components of the call establishment and release times or even worse, they deliver ambiguous results due to systematic errors* (see [Mau01], [Far01]). Therefore, the first part of my dissertation focuses on detailed measurements and analysis of these networks.

My dissertation focuses on call processing performance evaluation, optimisation and enhancement of signalling procedures in broadband networks with a special attention to ATM, Voice over DSL, VoIP and third generation of mobile networks.

Section 2 presents the objectives of my research, followed by the methodology in Section 3, where some definitions of performance measures are given as well. Section 4 presents the new results grouped into five theses. Thesis 1 focuses on the performance of the call processing based on the *measurements* of isolated signalling switches. The aim is to identify the main components of call processing in broadband networks and to describe quantitatively the effect of different call profiles.

Point-to-point (p-to-p) single connections, p-to-p multiple connections and point-to-multipoint (p-to-mp) calls are observed and studied. A new way ("population-diagram") of evaluating the results is also presented, thus eliminating the systematic errors of the testers. A guideline for a correct evaluation of signalling performance is inserted at the end of [D-4]. Thesis 2 and Thesis 3 deal with the construction and performance analysis of a new generic call processing model based on these measurement results from Thesis 1. The analysis is extended to network level, case studies for 10-node cascaded, 4-node fully meshed, 7-node, 30-node and 35-node arbitrary networks are investigated. Thesis 4 presents two algorithms to optimise the performance of a signalling node in UMTS networks. This UMTS network node is essentially a re-design of the model described in Thesis 2. Thesis 5 focuses on queueing analysis of wide-band blocked calls at the access point of broadband networks. Finally, some current and possible applications of my theses are described in Section 5, related and cited publications are listed in Section 6, while my own publications are shown in Section 7.

2. Objectives of the research

The objective of my dissertation is to analyse the performance of broadband signalling networks, to construct new methods and models that more accurately describe the behaviour of complex call processing, to develop new algorithms for capacity optimisation of signalling processors and to propose enhancements of existing signalling procedures.

The main focus is on ATM networks, but similar problems may appear in the UMTS Terrestrial Radio Access Networks, where ATM AAL2 has been selected as the transmission technology (see [Ene99], [C-7]). Moreover, the evolution of the Digital Subscriber Line technologies (xDSL), intelligent networks, IP-over-ATM, Voice over ATM (AAL1, AAL2) and WWW applications (RSVP) argue the introduction of the signalling capabilities of ATM networks, and therefore the need of real performance measurements of these networks [Mer00].

For this purpose I have carried out the following studies:

- measurements of the performance of the signalling processor on four commercial ATM switches in order to identify the most significant parameters that influence the message latencies through the switches and thus the call establishment times and release latencies.
- analysis of the complexity of calls in order to detect what is specific to ATM in the signalling flow and which kind of information elements have the strongest influence on latencies in a switch concerning the call establishment and release times;
- development and analysis of the most appropriate call model that reflects accurately all the ATM specific aspects of calls. Both FIFO and priority queues have been investigated;
- comparison of the performance of cascaded vs. arbitrary network topology, and analysis of the signalling load distribution in the network;
- development of two simulation-assisted optimisation algorithms for signalling queues;
- development and performance analysis of an architectural extension of the call model in order to introduce a queueing mechanism for wide-band blocked calls.

3. Methodology of the research

The necessity of performance evaluation studies before a new signalling system is introduced has been widely recognised as a tentative engineering task. The selected methodology should give a clear indication of the signalling network capability to support service requests with acceptable delay. Due to the fact that the most realistic measures are provided by measurements and there are

already available ATM switches on the market with signalling capabilities (there are globally about 10 widely spread ATM switches from different manufacturers), I have decided to carry out ***performance measurements*** for our analyses in Thesis 1. Out of these, four ATM switches were used in our measurements: *GDC APEX DV2* (1995), *Fore ASX200BX* (1997), *Newbridge MSX36170* (1999) and *Seabridge XP140* (2001). This is a good representative sample, covering a range of products from 1995 to 2001. My results are presented in Thesis 1.1 and they are in many cases in accordance to those obtained by other research groups (see [Kaus97], [Nie97], [Nov99]) on five other switches. Divergences are also explained, whenever necessary. Moreover, recent measurement results of 2, 3 and 4 cascaded ATM switches of different types (see [Mau01], [Far01]) confirm the results of my first thesis.

By a detailed analyses of burst arrivals of signalling messages (see Thesis 1.2), I had to conclude that the methodology used by other researchers (e.g., [Pil99], [Mau01]) is not sufficient to analyse burst arrivals of calls and therefore I have introduced a new method of representing the complete set of messages belonging to a burst in a so called *'population-diagram'*, thus opening new dimensions of analysing the structure of such bursts. Basically, this method can be used not only for burst arrivals, but any kind of signalling measurements.

However, while all companies and research groups interested in ATM are united in their interest, there is less consistency among the *evaluation metrics* and *measurement methods* applied. This makes it difficult to compare switches from different vendors, which is particularly important for service providers. I have also pointed out (delivering examples in Section 5.1 in Chapter [D-3]) that there is no consensus in the current research papers for using the same terminology for the evaluation metrics, which makes it quite complex to compare their results. I was trying to explain and adapt all these divergences to the newest standard adopted in 2000 by the ATM Forum [ATMF00]. A couple of new definitions were also necessary for a better characterisation of signalling performance in large networks (see Thesis 1.1 and Thesis 3).

While tests addressing the properties of a single switch are important and clearly relevant to LAN performance, it is obvious that tests addressing larger LAN and much larger WAN configurations are also relevant. But the problem is the unavailability of such large ATM networks with signalling capabilities. To buy a representative number of ATM switches (e.g. 30 nodes) just for testing purposes is too expensive and not arguable. Analytical performance evaluation like flow analysis supplies good results for the mean values only. Emulation of signalling protocols allows a deep insight in the protocol behaviour, but implies limitations in the performance, and is especially difficult to emulate all nodes of different manufacturers. Therefore, I have turned to ***simulations***[1]. This approach supplies results on a higher abstraction level, but the tool is able to simulate all different vendors' nodes. It provides call establishment times, release latencies, bandwidth utilisation of links, call throughput and signalling load of nodes. My investigations include a construction of a new message flow model for one signalling node and then analysis of cascaded and arbitrary large ATM networks (see Thesis 2 and 3), comparison of infinite versus finite buffer solutions, development of an optimisation algorithm and a study of introducing priority handling of certain messages (see Thesis 4). The results are validated by measurements on 2-to-4 cascaded switches and on the 7-node TEN-155 network, see [Nov99].

The proposed generic model is too complex to be solved analytically. Therefore, I have developed a simplified call model as well. Of course, this simplification has its drawbacks, as it is shown in Thesis 2.2. To solve it ***analytically,*** I have used the results of the queueing networks theory, i.e. a special case of the BCMP networks [BCMP75].

[1] The simulation study was carried out based on the simulation software called ACCEPT, developed by the author and his colleague I. Moldován, HSN Laboratory, Hungary.

Finally, in Thesis 5, I have presented a simplified version of the Finite State Machine graph of layer 3 UNI signalling protocols, adding a new extension to it for possible queueing of blocked calls in the access nodes. Then, for the performance evaluation of a queueing mechanism of wide-band blocked calls I have carried out again both *simulation* and *analytical studies*.

A couple of ***definitions*** are given here in order to help understanding the subsequent theses, the rest is given in the *'Terminology'* chapter of my dissertation [D-10].

1. The *call establishment time* is the amount of time that it takes for a signalling system (e.g., ATM) to establish a switched virtual connection between network components.

$$T_C = t_S(Y) - t_S(X), \quad (1)$$

where $t_S(.)$ is the timestamp of an outgoing or incoming message at the source; *X=SETUP, ERQ; Y=CONNECT, ECF* are signalling messages defined in [Q2931], [UNI40], resp. [AALQ99]. This is arguably the most fundamental signalling performance metric.

2. The *call release time* is the amount of time taken to release a connection over one network element:

$$T_R = t_S(Y) - t_S(X), \quad (2)$$

where *X=REL, Y= RLC* are signalling messages described in [Q2931], [UNI40] and [AALQ99].

3. The *call establishment latency* is the difference between the call establishment time and the response time of the destination to a setup message:

$$T_{CN} = T_C - T_{DS}, \quad (3)$$

where T_{DS} is the destination response time on call setup.

4. The *call throughput* (γ_R) of a switch (network) is the *number of successful calls / number of generated calls*.

Most of the call performance metrics I am using throughout this dissertation are depicted in Figure 1 below.

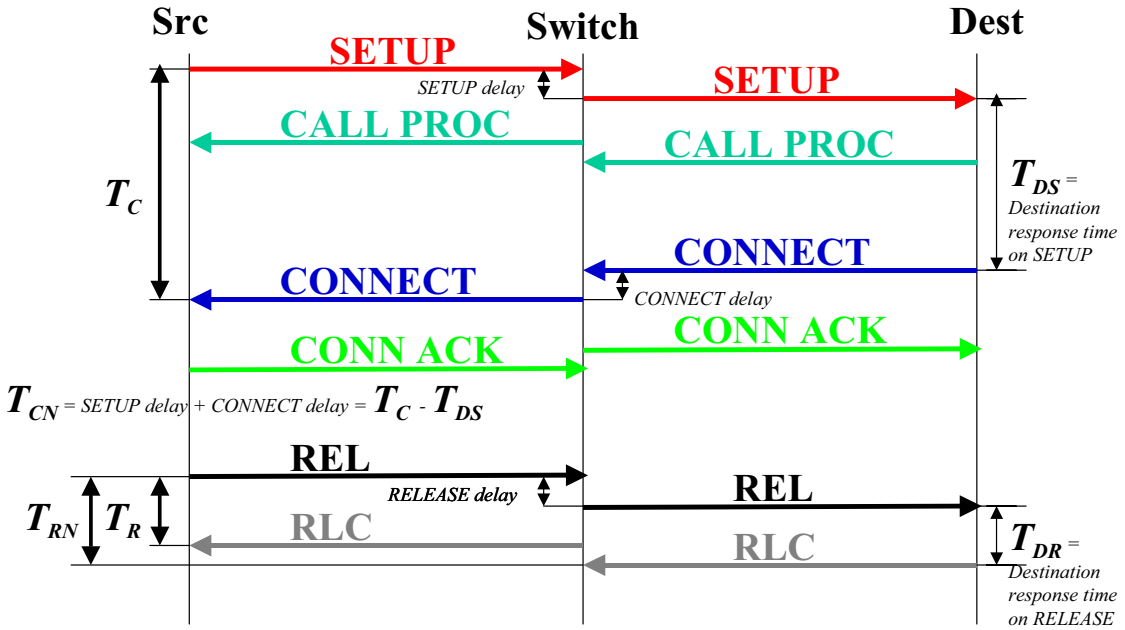

Figure 1 Definition of performance parameters for call establishment and release

4. New results

I have grouped my theses into five groups according to the specific subjects. Each thesis group consists of two theses except Thesis 3 (only one). There are references in each of my thesis to the specific chapter of my dissertation [D-*n*], *n*=1,…,10 for further details and proves.

Intrinsic Properties of Point-to-Point Call Processing in ATM networks

Thesis 1 *Based on the analysis of switch invariant signalling features of ATM switches, I have found a set of intrinsic properties for the steady-state (see Thesis 1.1) and transient-state (see Thesis 1.2) of point-to-point connections. I have introduced a new method ('population-diagram') for the analysis of burst arrivals of calls.*

I have analysed the signalling measurements obtained on the four aforementioned commercial ATM switches. In this thesis I have grouped a set of generalised conclusions, which are *switch invariant features* of the ATM signalling flows and thus *general to a wide range of ATM switches.* The results are in accordance with the results obtained on five other switches by [Nie97], [Nov99], [Far01] and [Mau01].

We have carried out a set of **signalling performance measurements** on *p-to-p single-* and *p-to-p simultaneous* ATM calls. Our *p-to-mp* measurement results are not general enough yet to be included into this dissertation (see [C-8]). The main interest was on the performance of layer 3 call processing in ATM switches. The basic configuration that we have used is the following: one call generator and one receiver connected to one isolated ATM signalling node. In this simple case we had two user-network interfaces (UNI) with STM-1 (155Mbps) optical interfaces.

During our measurements we made some simple assumptions and configured the system as such:

- we assured only successful p-to-p and p-to-mp calls (no errors during call establishment);
- we have assigned very low bandwidth requirements for calls, therefore calls were never rejected due to unavailable bandwidth on links (the user plane was not a bottleneck);
- the input pattern of the *SETUP* messages was set to constant rate or burst arrival (generation of Poisson arrival was not possible with none of our testers: *HP BSTS75000* (1996) and *GNN iW95000* (2000)). However, as shown by the measured results, the offered call arrival rate varied from constant rate to on-off type starting at a certain rate due to some systematic errors of the testers (see Thesis 1.2.1).

We have carried out both steady state and transient state signalling measurements, the results are grouped into two theses according to these two sets.

Thesis 1.1 *I have found a set of intrinsic properties of point-to-point single connections in ATM switches (steady state measurements).* **[C-5], [C-9], [J-4], [H-3], [D-4]**

Table 1 Overview of the Thesis 1.1

1.1.1.	The dominancy of layer 3 processing of messages + new performance measures defined
1.1.2.	Differences between message processing delays
1.1.3.	Dependency of the T_C on routing table, bandwidth allocation and number of active connections
1.1.4.	Dependency of message processing times on different call profiles
1.1.5.	The impact of the call release phase on the T_C
1.1.6.	The impact of the signalling overload on the T_C and T_{RN}
1.1.7.	Estimation of the T_{CN} for cascaded switches

In this thesis I have proven that the standard performance measures (e.g., call establishment time, release time) defined by the standardisation institutes (e.g. [ATMF00]) are not enough to properly characterise the performance of a broadband signalling network. I have shown that the impact of the destination response time, release latency, call throughput of switches, size of the routing table and complexity of call profiles (i.e., kind of information elements in *SETUP* messages) have to be considered as well. Measurements are taken after the steady-state flow of the signalling traffic has been reached. The results are grouped into 7 independent sub-theses, a brief overview of these statements is shown in the Table 1.

I have stated that the call release time T_R does not provide a measurement of the time taken to tear down a connection over a call-path from end-to-end in ATM networks. Therefore, **I have introduced new performance measures** to properly characterise the signalling behaviour in the network side:

Def 1.1 The *call release latency* is the amount of time taken for a release message to travel along the path from end-to-end, followed by an acknowledgement of the destination:

$$T_{RN} = t_D\,(RLC) - t_S\,(REL), \tag{4}$$

where $t_D\,(.)$ is the timestamp of an incoming or outgoing message at the destination.

Def 1.2 The *overall handling time* is the sum of the call establishment time and call release latency:

$$T_H = T_C + T_{RN}\,. \tag{5}$$

1.1.1. Based on measurements and analysis **I have concluded that the dependency of call establishment time T_C on layer 3** (e.g., UNI4.0) **processing is at least one order of magnitude higher than its dependency on lower layer processing.**

The layer 3 processing is software implementation, layer 2 is partially implemented in SW, partially in HW, while layer 1 processing is implemented only in hardware. Measurement results, calculations and a table of comparison are given in [D-4]. As a conclusion of Thesis 1.1.1 in the followings I have investigated only the effect of layer 3 processing on the signalling performance of ATM switches.

1.1.2. **I have found that the following relationship between the minimum message delays is valid for all** (*tested*) **ATM switches regardless of the processing capacity:** *the CONNECT and RELEASE delays are (25...35)% of the SETUP*[2] *delay, respectively***.**

1.1.3. **I have shown that the call establishment time T_C depends *linearly* on the size of the routing table, but it is independent of the size of the allocated bandwidth.** Moreover, **I have shown that the number of active calls in the switch has no influence on the call establishment time of a new call.**

In addition to our results, some more details can be found in [Nie97] and [Far01], e.g. the dependency on the PNNI hierarchy is shown. Note that, when running the tests with different bandwidth sizes, we paid attention to avoid call rejections due to overloaded links.

1.1.4. **I have given a quantitative measure, how the mean delays of *SETUP, CONNECT and RELEASE* messages (and therefore the call establishment time T_C) depend on different call profiles, when processed at an ATM signalling node:**

- The mean *SETUP delay* and the T_C depend on the call profiles as follows:

[2] It represents a SETUP message containing the mandatory Information Elements only, with a length of 3 ATM cells.

Table 2 Effect of IEs on SETUP delay and call establishment time

Adding to the *default SETUP message* (containing only the mandatory IEs)	**Mean SETUP delay**	**Call establishment time**
AAL (1, 3/4 or 5) Parameters IE	1-2% decrease!	0% increase
P2MultiP connection IE (Bearer Capability)	0% increase	0% increase
Called Party Sub-address IE	10-32% increase	10-22% increase
Calling Party Address IE	12-14% increase	6-10% increase
Higher Layer IE	38-60% increase	28–40% increase

This variation of the mean *SETUP delay* is due to the time taken to process call profiles of different complexity (i.e. kind of information elements). There is a large scale from simple voice call to complex multimedia calls. Some examples are given in Table 2. The results are obtained at 1 call/sec signalling load. In general, the increase in the call establishment time due to the complexity of the *SETUP* message can be described as follows:

$$T_C^{CCP} = (1+s) \cdot T_C^{default}, \text{ where } 0 < s < 1. \quad (6)$$

- I have found that in contrast to the mean *SETUP delay*, the mean *CONNECT delay* and mean *RELEASE delay do not increase* when the parameters (*IEs*) from Table 2 are added to the default *SETUP* message. Furthermore, the call release latency of complex calls:

$$T_{RN}^{CCP} = T_{RN}^{default} \quad \forall s,\ 0 < s < 1. \quad (7)$$

- I have shown that the mean call establishment time does not depend on the type of call, e.g.:

$$\overline{T}_C\big|_{VBR} \approx \overline{T}_C\big|_{CBR} \approx \overline{T}_C\big|_{ABR} \approx \overline{T}_C\big|_{UBR}, \quad (8)$$

where $\overline{T}$ denotes the average time (see Table 2, the case when adding AAL1, AAL3/4 or AAL5 parameter IEs). Early papers in the literature expected that the call establishment time of VBR calls will be larger than that of CBR calls (see [Gel97], [Wu97]), but our measured results contradict to these statements. A recent paper [Far01] has shown again by measurements that the T_C of UBR calls is only a bit shorter than that of CBR calls in a 4-node PNNI hierarchical network, which is in line with the equation (8).

1.1.5. I have investigated the *mean call establishment latency* T_{CN} of simple calls (i.e. *default SETUP*) **in two cases: call establishment followed by call release vs. call establishment without release phase and I have found the following relationships:**

- In case of releasing the calls after a given holding time, the mean of the measured call establishment latency is (15...20)% longer compared to the case when calls are established but not released. This property is independent of the call arrival rate (see Figure 2).

- In case of releasing the calls after a given holding time, the call intensity threshold where the call establishment latency starts to increase dramatically is (65...70)% that of the case when calls are established but not released. This threshold is also related to the point where rejected calls start to appear.

- I have shown that the duration of calls has no influence on the call establishment time, except one case, when the holding time is infinitely long (then there is no release phase).

Note: In [Kaus97] it is stated that calls of zero (or non-zero) duration give very similar results to those not tearing down switched virtual connections at all. As seen also in Figure 2, our results contradict to this statement.

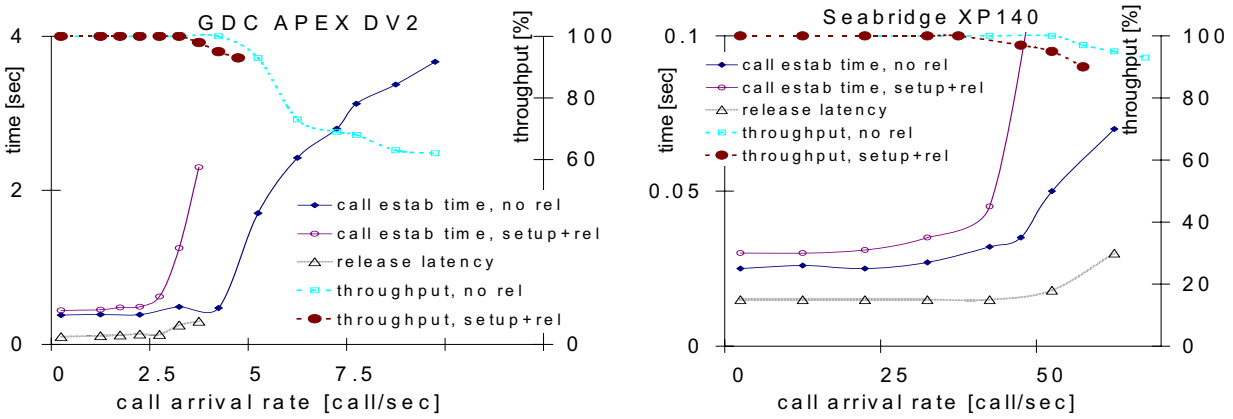

Figure 2 Call establishment time, release latency and throughput vs. call intensity
case a) GDC APEX DV2; case b) Seabridge XP140

1.1.6. In case of a light call overload in the switch (i.e., less than 10% of the calls are rejected), **the call establishment time is dramatically increased, but the slope of the call release latency remains for this range still unchanged** (see Figure 2)**.** Further increasing the call intensity leads to a dramatical increase of the release latency as well.

This may happen when there is a priority mechanism applied or the *SETUP* and *RELEASE* messages visit partly different paths inside the signalling processor (distributed call processing). In none of the 4 switches we have studied is priority applied. The call rejections are due to the buffer overflow in the processor.

1.1.7. I have shown that the *call establishment latency* of calls with the same call profile that go through *'r'* similar cascaded switches (within one PNNI peer group) **satisfies the following inequality:**

$$r \cdot \overline{T}_{CN}\big|_{1switch} \geq \overline{T}_{CN}\big|_{'r'switches}, \quad r = 1,2,...,k\,. \tag{9}$$

This formula can be generalised to a heterogeneous cascaded network (within one PNNI peer group), as follows:

$$\sum_{i=1}^{P} n_i \cdot \overline{T}_{CN}^{\text{type}\,i} \geq \overline{T}_{CN}\big|_{'r'switches}, \quad r = \sum_{i=1}^{P} n_i, \tag{10}$$

where P is the number of different types of ATM switches in the network and n_i is the number of switches of one type. This behaviour is due to message overlapping in the cascaded switches. Measurement results presented by [Mau01] on 3 different types of cascaded switches confirm the equation (10). An approximation formula instead of these inequalities is developed in Thesis 2.2.

Thesis 1.2 *Introducing a new method, I have found a set of intrinsic properties of point-to-point simultaneous calls in ATM switches (transient state measurements) (see Table 3).* **[C-5], [C-9], [T-2], [D-4]**

Table 3 Overview of the Thesis 1.2

1.2.1.	A new method for analysing a signalling burst, called *'population-diagram'*
1.2.2.	Intrinsic properties of simultaneous calls
	1.2.2.1. The first call needs a longer establishment time than the next ones
	1.2.2.2. Change is the slope of curve of the *CONNECT delay* during the burst
	1.2.2.3. Approximation formula for $T_C(b)$
	1.2.2.4. Approximation formula for $T_{RN}(b)$

Performance evaluation of calls with constant arrival rate in steady-state (see Thesis 1.1) does not give a complete characterisation of the exact behaviour of ATM switches. I expected different behaviour when generating p-to-p multi-connection (simultaneous) calls, or at a network node failure automatically followed by a re-establishment of previously existed connections, which lead to burst arrival of signalling messages. Due to a high number of ambiguous published results (see e.g., [Mau01], [Pil99]) regarding burst measurements of signalling messages, I have decided to carry out a detailed analyses of these measurements. The traditional way of representing the flow of signalling messages (using arrows between two nodes) or looking at the call establishment times only does not always help us finding the "hidden" system faults, therefore I have looked for new representation methods. My conclusions are summarised in Thesis 1.2, which consists of two independent sub-theses as follows:

1.2.1. I have introduced a new method, *'population-diagram'* of representing the flow of signalling messages, which reveals the detailed structure of the signalling burst mechanism.

The main idea is (see Figure 3):

- to represent all the (interesting) layer 3 messages of the system related to the observed burst in *one* diagram split into many time axis on the horizontal scale and representing the links instead of nodes in the vertical scale;
- to represent one signalling message with a *plot* at the link (interface between two nodes) instead of an *arrow* between two nodes;
- to create "*m*" time axis at each interface, representing each (interesting) type of message in order of appearance at separate time axis, where *m* is the number of message types at that interface;
- to split the diagram into (*n*+1) parts if the messages visit *n* nodes during a call setup;
- if the calls are generated from both sides of the communication path, then split the "*m*" axes into *"*2x*m"* sub-axes, to be able to distinguish between the two directions (tx; rx) when a certain message is propagated through the link;
- to represent the layer 2 messages (*POLL, STAT, USTAT*) in the same diagram, any time it will be necessary to avoid confusions.

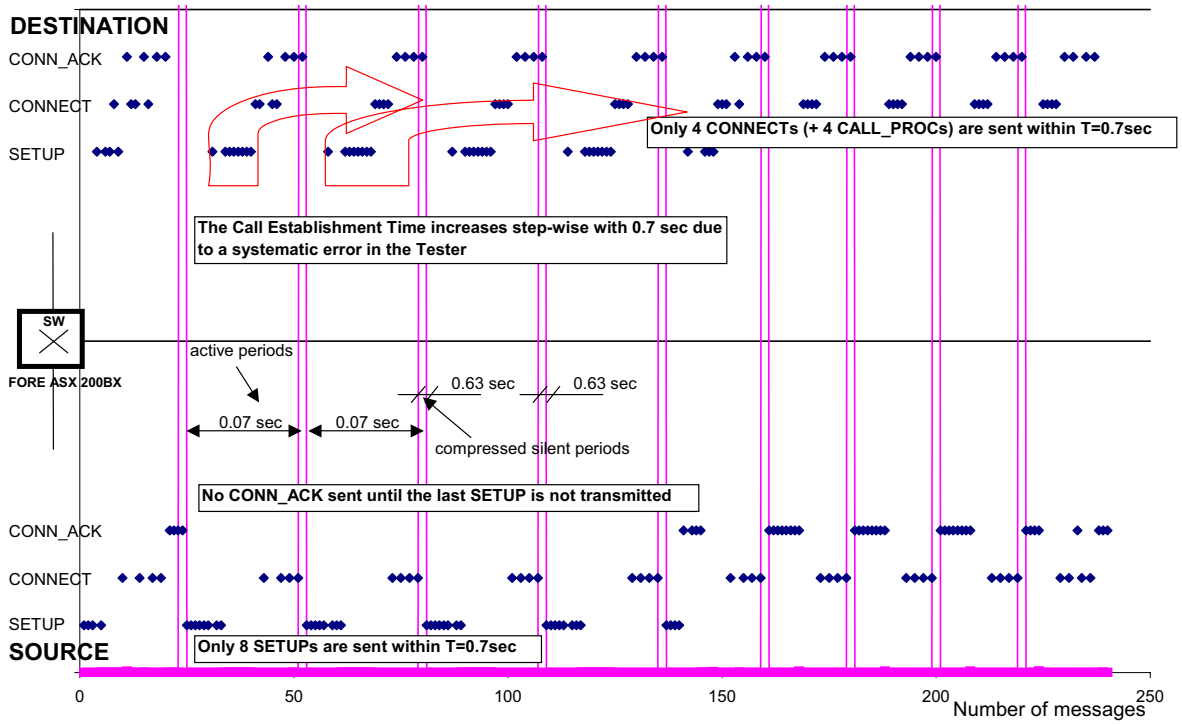

Figure 3 Example of using the "population-diagram" for representing a burst arrival of signalling messages

The lower segment covers the signalling events at the link between the source and network node, while the upper part shows the messages at the link between the node and destination, sorted by the time they have arrived. It is very important that the messages are arranged in a consecutive order defined by standards (e.g., [Q2931]) from the bottom to the top (e.g., *SETUP*→ *CALL_PROC*→ *CONNECT*→ *CONN_ACK*→ *STATUS_ENQ*→ *STATUS*→ *REL*→ *REL_COMP*). In fact, it is not necessary to represent all messages, but it is recommended to use at least m=3 (e.g., Figure 3).

This diagram provides a better global overview than the "arrow-type" diagram used previously in the literature (e.g., Figure 1). As an example, I have represented a scenario with (1 source; 1 node; 1 destination) in Figure 3 (m=3; n=1), when originally 40 *SETUP* messages were generated in a burst.

The x axis representing the time scale can be compressed in order "to make the things visible", e.g., the "active" periods of 70 msec are shown, followed by "silent" periods of 630 msec, which are compressed. Note that this way of representation can be used to steady-state measurements as well.

Def 1.3 I have given a new definition of the signalling burst arrival at a network node, which "looks" at the node that receives a message, instead of looking at the generator. This is in contrast to the classical way of defining a burst arrival, which captures the idea of clustering in time. My definition is based on the observation of a state variable:

It is considered a signalling burst arrival, if at the time a *CONNECT* message arriving to a node founds more than 2 *SETUP* messages in the buffer, i.e.:

$$n[SETUP_{in}(i)] - n[SETUP_{out}(i)]\big|_{\Delta t} > 2 \qquad (11)$$

where the i=1,…,r is a node, $n[X(i)]$ is the number of elements of type X at node i, Δt can be arbitrarily chosen.

In the followings we use the term "signalling burst arrival" according to the definition Def 1.3. Many case studies are given in [D-4] to show the wide range of applicability of the 'population-diagram'. After a detailed analyses of the burst measurement results on four switches, I have collected my conclusions in Thesis 1.2.2, describing *the properties of the signalling burst structure*.

1.2.2. I have found that at burst arrival of *SETUP* messages the following properties can be found for the call establishment and release processes:

1.2.2.1 The first call has always 80-100% longer establishment time than the second one.

This happens because either an address resolution from IP to ATM has to take place before a call can be initiated or an end-to-end signalling path has to be established first. Once this mapping is stored in the cache, the next calls are executed in shorter time interval than the first one, showing then a gradual increase as the burst size increases.

1.2.2.2 Opposite to the steady-state behaviour, in the first-phase of the burst the *SETUP delay* is lower than the *CONNECT delay*, but the gradient ($tg\ \alpha$) of these curves is the same until equation (11) is fulfilled, then the slope of the *CONNECT delay* starts to decrease with $tg(-\alpha)$, $0 < \alpha < \pi/2$.

Due to the fact that all messages have the same priority, before the first *CONNECT* is sent back from the destination, the signalling buffer of the switch is already filled with the subsequent *SETUP* messages, thus the *CONNECT delay* will be longer than the *SETUP delay*. Later, when no more *SETUP* message arrives, then the *CONNECT delay* will gradually decrease.

1.2.2.3 The *call establishment time* T_C is always longer for simultaneous call arrivals (transient state) than in steady-state of deterministic arrivals. Moreover, T_C depends on the size of the burst of the arriving *SETUP* messages, and can be approximated as follows:

If the *burst size (b) ≤ max buffer size (BS)* of the signalling processor, then

$$T_{Ci}(b) = 2 \cdot T_{C2} \quad \text{for } i=1,$$
$$T_{Ci}(b) = T_{C2} + (i-2) \cdot tg\,\alpha(b), \quad \text{for } 1 < i \le b \le BS, \qquad (12)$$

else $T_{Ci}(b) = 0$, if the call is lost,

or $T_{Ci}(b) = T_{C2} + T_{303} + (j-1) \cdot tg\,\alpha(b)$, if the call is repeated,

where:

i = position of the message within the burst, $i = 1,2,...,b$ (for i=1 see statement 1.2.2.1);

$tg\,\alpha(b)$ = is the tangent to the curve of T_C, this gradient will be however different for different burst size;

T_{303} = 4 sec is the retransmission timer, see [Q2931];

T_{C2} = the second (*minimum*) call establishment time;

j = position of the message within the retransmitted burst, $j = 1,2,...,b$.

1.2.2.4 Similarly, the *call release latency* T_{RN} can be approximated as follows:

If the *burst size (b) ≤ max buffer size (BS)* of the signalling processor, then

$$T_{RNi}(b) = T_{R1} + (i-1) \cdot tg\,\beta(b), \qquad (13)$$

else $T_{RNi}(b) = 0$,

where:

i = position of the message within the burst, $i = 1,2,...,b$;

$tg\,\beta(b)$ = is the tangent to the curve T_{RN}.

The tangents to the curves of T_C and T_{RN} ($tg\,\alpha, tg\,\beta$) depend on the following parameters: burst size, message type, call profile, processor capacity. All these dependencies are shown in [D-4]. The burst size is increased gradually from 10 to 50 messages, the message type is either *SETUP* or *RELEASE*, the call profile is changed from simple calls to complex multi-media calls, and the capacity of all four ATM switches has been investigated.

Validation of Thesis 1

The intrinsic properties presented in Thesis 1 represent the results of a very "unpopular" research, the measurements were collected during the last four years by our research team on four different generation of switches manufactured by different vendors. Moreover, the results were obtained with two different testers, and in addition, independent of our work four other research groups (see [Nie97], [Nov99], [Far01] and [Mau01]) obtained very similar results with other types of ATM switches, thus confirming that these statements are general to a wide range of ATM switches. All of our measurements were repeated 10 times, one set containing 60 to 100 calls. The minimum, average, maximum values, the standard deviations and standard errors were evaluated on these sets of measurements.

Representing the measured results in a 'population-diagram' (Thesis 1.2.1) gives us immediately a 'visual inspection', a first-step validation of our results. Secondly, it helps us discovering interdependencies which are not visible measuring the call establishment times and release latencies only (e.g., retransmission of lost messages, unexpected messages in a certain state, biased results due to the layer 2 flow control mechanism, etc.)

The importance of these results is reflected in the fact that they provide detailed analyses of each component of the call establishment time and release latency. Moreover, I have shown that these features of the ATM signalling are very general regardless of different architecture and processing power of ATM switches.

Construction of a generic call processing model for ATM networks

Thesis 2 *I have developed two call models to describe the processing of signalling messages in ATM nodes (see Thesis 2.1 for simulation studies and Thesis 2.2 for analytical studies).*

The construction of a call model for UNI and Private Network Node Interface [PNNI] has been motivated by the fact that the service providers want to introduce dynamic call establishment in their access Digital Subscriber Line (xDSL) and backbone ATM network. However, today is almost impossible to have access to large ATM networks with signalling capabilities. Some very basic results have been obtained on the *TEN-155* Pan-European ATM network [Nov99]. This network consists of seven ATM signalling nodes, the longest path (diameter) contains only four nodes. Therefore, our experiences gained in a small network have been extended by simulation to real size networks. Moreover, in Thesis 1 we have shown that in many cases the test equipment may be the bottleneck, therefore we have developed an adequate model for the end systems as well.

Thesis 2.1 *I have developed a new model (shown in Figure 4) and presented an algorithm to set its parameters.* **[J-4], [C-9], [C-10], [D-5]**

The architecture of the model

2.1.1. I have developed a new call processing model for an ATM signalling node based on the specifications [Q2931], [UNI40], [PNNI] and the measurement results presented in Thesis 1, respectively.

The [Q2931], [UNI40] and [PNNI] standards do not specify any call model, but describe the format, content of the signalling messages and the protocol how they interact. From the results presented in Theses 1.1.2, 1.1.4 and 1.1.6, it is clear that the average call release latency is always shorter than the average call establishment latency, furthermore, the ratio between these parameters varies as the call arrival rate is increased. Therefore I have investigated two different mechanisms in the proposed model: *FIFO queueing* and *priority queueing.* Moreover, at one node I have distinguished *separate processor phases* according to the jobs to be done (see Figure 4). E.g., there are 5 processes visited by a *default SETUP* message and one more by a *complex SETUP* with additional capabilities (according to Thesis 1.1.4). Instead, *CONNECT* and *RELEASE* messages visit only 3 processes.

The separated processes are as follows: *UNI/PNNI* to decode/encode the incoming/outgoing messages, *CC* to create and update the objects related to one call, *RT* for path selection, *BW* for bandwidth allocation/de-allocation on the outgoing link, and finally, *CCP* for execution of a complex call profile, buffer allocation and Quality of Service issues. The ATM specific part of this model is given by the followings: different bandwidth requirements can be served in the *BW* block, different internal paths are visited according to the message types, and different service rates are obtained due to specific call profiles, routing, QoS service guaranties, buffer allocation for CBR and VBR traffic, etc. The service rate of *RT* is dependent on the size of the routing table and on the level of PNNI peer-group hierarchy. A higher level gets a lower service rate (according to the results of [Nie97]).

Two other call processing models have been developed in [Wu97] and [Gel97]. While the model in [Wu97] is even more complex than ours and considers different service times for CBR, VBR and UBR calls respectively (which is irrelevant according to our measurements, see Thesis 1.1.4), none of the two other models considered the release phase. However, as described in Thesis 1.1.5, the T_C

is increased by 15-20% when release messages are also present in the network. Furthermore, the model in [Gel97] consists of a single FIFO queueing model, which cannot capture the differences between the setup and release latencies (see also Thesis 2.2.1).

Parameter settings of the model

The problem can be formulated as follows:

- The known (measured) parameters are the message delays, the call throughput of the switch, the call establishment times and the release latencies vs. call arrival rate (e.g. see Figure 2);
- It has to be determined the service rate (μ_i) and the buffer size (BS_i) of each process of the signalling processor and the retransmission delay of lost messages *(RDLM)*[3].

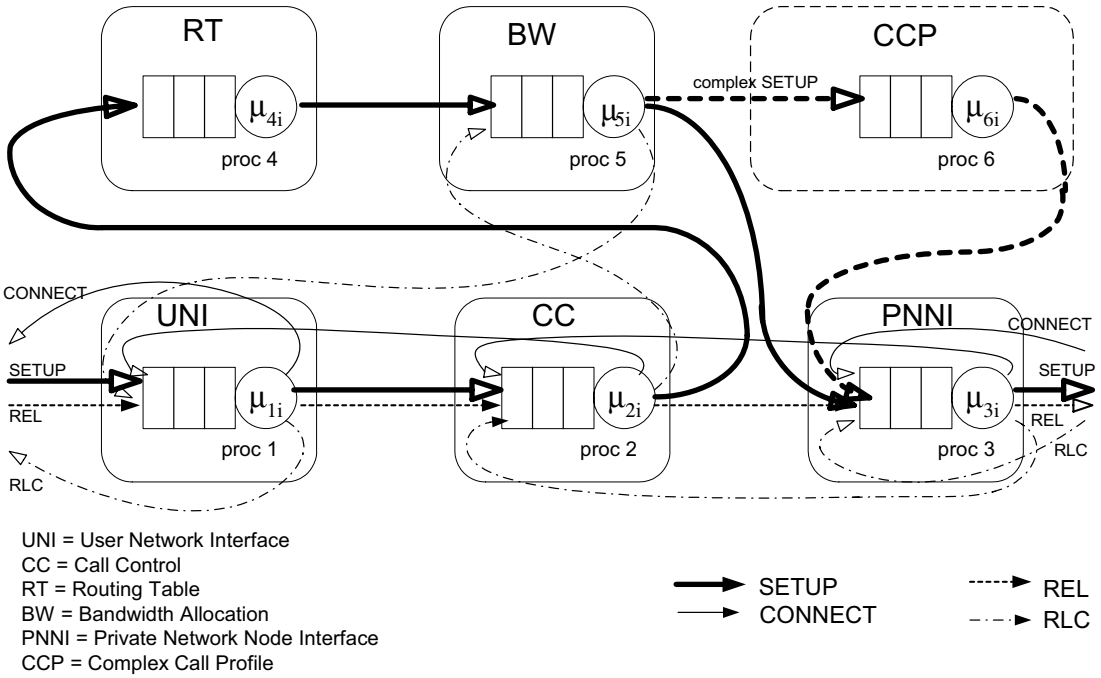

Figure 4 Call processing model of an ATM signalling point

I have shown that the *service time* of each process can be derived from the system of equation (14). The *buffer size (BS_i)* of each process and the *retransmission delay of lost messages (RDLM)* are derived by simulation. The input parameters needed for these settings are the service rates, obtained by solving the system of equation (14), the call throughput of the switch, the call establishment times and release latencies obtained by measurement at different call arrival rates.

2.1.2. I have developed an algorithm to obtain the μ_i, $BS = \sum_{i=1}^{6} BS_i$ and *RDLM*, which consists of four steps:

I. Set the service rates according to the system of equations (14).

$$\sum_{i=1}^{5} \frac{1}{\mu_i} = \min \ \textit{SETUP delay} \qquad \sum_{i=1}^{6} \frac{1}{\mu_i} = \min \text{ '}\textit{complex}\text{' SETUP delay}$$

$$\sum_{i=1}^{3} \frac{1}{\mu_i} = \min \ \textit{CONNECT delay}^{4} \qquad \sum_{i=1,2,3,5} \frac{1}{\mu_i} = \frac{1}{2} \cdot \min(\textit{RELEASE time} + \textit{RELEASE delay}) \qquad (14)$$

[3] *RDLM* is the delay a lost message (other than *SETUP*) is regenerated at the switch in case of buffer overflow.
[4] or min *RELEASE delay*

$1/\mu_1 = a/\mu_2 = 1/\mu_3$, *where* $a \in (1,2]$ *is a correction factor. Modify* $a \in (1,2]$ *until:* $|T_{CN}^{meas} - T_{CN}^{sim}| < 0.05 \cdot T_{CN}^{meas}$ *and* $|T_{RN}^{meas} - T_{RN}^{sim}| < 0.05 \cdot T_{RN}^{meas}$, *while* $\gamma_R = 1$.

II. *Set the length of the bottleneck buffer* BS_{j^*}, $j^* = \min\{j \in \{1,...,6\} \mid \forall k \in \{1,...,6\} : q_j \geq q_k\}$ *so that the threshold values* $\left|\lambda_{Th}^{sim} - \lambda_{Th}^{meas}\right| < 0.05 \cdot \lambda_{Th}^{meas}$, where $\lambda_{Th} = \min_{i \in I}\{\lambda_i \mid \frac{T_{CN,i+1} - T_{CN,i}}{\lambda_{i+1} - \lambda_i} \geq 1\}$, *I is a finite set of measured/simulated data and* q_j *is the message length of queue j.*

III. *Set the length of the remaining buffers until the throughput* $|\gamma_R^{meas} - \gamma_R^{sim}| < 0.05 \cdot \gamma_R^{meas}$, $\forall \lambda > \lambda_{Th}$.

IV. *Set RDLM to further minimise the errors between the three measured and simulated curves in the overload region* $(\lambda > \lambda_{Th})$.

Validation of Thesis 2.1

First of all, I have shown that the simulated results of all three parameters (T_{CN}, T_{RN}, γ_R) can be adjusted to the measured ones for all four switches we have studied (e.g., see Figure 5). The *BS* and *RDLM* parameters are determined such that the errors between the simulated and measured results are less than 5%.

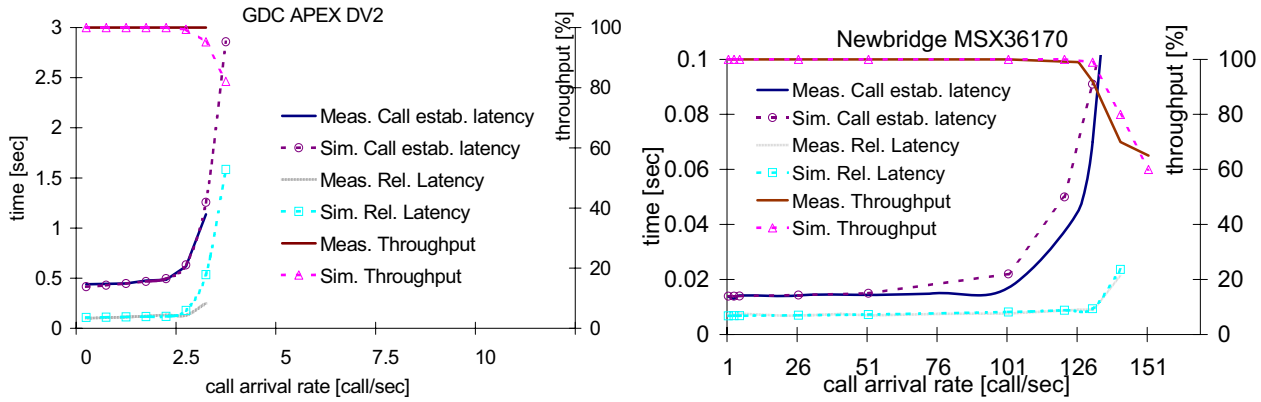

Figure 5 Validation of the model for 'default' SETUP messages: a) GDC APEX DV2 switch; b) Newbridge MSX36170 switch

To reduce the effect of the destination response time, we have replaced the (1 source; 1 switch; 1 destination) configuration by that of (10 source; 1 switch; 10 destination). In our simulation studies we have generated constant and Poisson arrival rates, respectively. The average values were obtained from 10000 calls. To validate our model in a real environment, we have selected a 7-node ATM network, representing the *TEN-155* network offering service to the European research community [Nov99]. The backbone nodes are all *LUCENT* (formerly *ASCEND*) *CBX500* or *CBX550*. The SVC network paths consist of 2, 3 or 4 hops, but we have observed 12 different call establishment times due to different switches and workstation speeds.

The measured call establishment results are presented for each node-pairs in [Nov99]. Each measurement result was an average of 5-6 measurements, each consisting of a stream of 'ping' messages. Our simulated results consisted of 3000 calls at 1 call/sec rate. The comparative results are shown in [D-5]. In more than 80% of the cases, the simulated results are very close to the measured ones, however there are a few deviations, due to the inaccurate "ping-type" measurement results.

Thesis 2.2 *I have constructed a simplified signalling flow model for analytical studies and additionally I have given a new heuristical formula for computing the T_C and T_{RN}.* **[C-6], [C-9], [D-5]**

In Thesis 2.2.1 I have shown that the flow model given in Thesis 2.1 can be simplified in a way that it allows analytical solutions. Due to the fact that the T_C of a cascaded network will always slightly overestimate that of an arbitrary network configuration (as shown later in Thesis 3.1), I have offered analytical solutions for the cascaded network only. The selected network model is a special case of the BCMP model (off-the-shelf technique, see [BCMP75]), where the node model consist of one equivalent central server and one queue only. The messages represent the classes, and there is a mandatory class change at the end users: *SETUP* (`class 1`) changes to *CONNECT* (`class 2`), *CONNECT* to *RELEASE* (`class 3`), *RELEASE* to *RELEASE COMPLETE* (`class 4`). The only external call pattern for `class 1` (*SETUP*) calls at the first node is assumed to be state independent Poisson arrival with a mean rate of λ_N [calls/sec].

2.2.1. I have shown that the call establishment time of a cascaded network with *r*-nodes can be approximated analytically by using an M/M/1/K queueing model for call processing in each node and an infinite service model for modelling the call duration.

In this queueing network the load distribution can be examined by solving the traffic equations:

$$e_{js} = (p_{0,js} - q_{0,js}) + \sum_{i,u} e_{ir} \cdot (p_{iu,js} - q_{iu,js}), \quad i,j = 1,...,r+2; \quad u,s = 1,2,3,4 \tag{15}$$

where: $p_{0,11} \neq 0, \quad p_{0,js} = 0 \ \forall j,s > 1$. The routing probability in each node satisfies the condition: $\sum_{j,s}(p_{iq,js} + q_{iq,js}) = 1$. The matrix form of this system contains an $[s_{\max} \cdot (r+2)$ x $s_{\max} \cdot (r+2)]$ routing matrix, which is a sparse matrix.

Suppose that: $p_{iu,js} > 10 \cdot q_{iu,js}$, $p_{iu,js} + q_{iu,js} = 1$. Then, in the equation (15) all the terms containing a product of at least two *'q'*-s, i.e., $(q_{iu,js} \cdot q_{kv,lw})$ can be neglected. Furthermore, note that: $q_{\max} = \max_{i,j,u,s}\{q_{0,11}, q_{iu,js}\}$.

Then the most complex term (e_{14}) out of all e_{js} can be approximated under the following form:

$$e_{14} \geq \prod_{\substack{i=0 \\ s=1,3}}^{r} p_{is,(i+1)s} \cdot \prod_{\substack{i=0 \\ s=2,4}}^{r} p_{(i+1)s,is} - q_{\max} \cdot \left(\sum_{\substack{i=0 \\ s=1,3}}^{r} p_{is,(i+1)s} + \sum_{\substack{i=0 \\ s=2,4}}^{r} p_{(i+1)s,is}\right)$$

In this expression we need to find the conditions for $q_{\max}$ so that the second term can be neglected:

$$q_{\max} \cdot \left(\sum_{\substack{i=0 \\ s=1,3}}^{r} p_{is,(i+1)s} + \sum_{\substack{i=0 \\ s=2,4}}^{r} p_{(i+1)s,is}\right) \leq 10^{-2} \cdot \prod_{\substack{i=0 \\ s=1,3}}^{r} p_{is,(i+1)s} \cdot \prod_{\substack{i=0 \\ s=2,4}}^{r} p_{(i+1)s,is}$$

If we find an upper bound for the denominator, and a lower bound for the numerator, then we arrive to even more severe conditions:

$$q_{\max} \leq \frac{10^{-2} \cdot (p_{\min})^{4 \cdot (r+1)}}{4 \cdot (r+1)}$$

where, $p_{\min} = \min\{p_{iu,js} > 0, \ \forall i,j = \overline{0,r}; \ \forall u,s = \overline{1,4}\}$.

Case study: $\begin{matrix} p_{\min} = 0.99 \\ r = 8 \end{matrix}$ $\qquad q_{\max} \le \frac{10^{-2} \cdot (0.99)^{36}}{36} = 2 \cdot 10^{-4}.$

This case is a realistic scenario for signalling networks under study for a light overload $(p_{\min} = 0.99)$. If we admit this condition, then we can neglect the term containing $q_{\max}$, thus arriving to the same formula for equation (16) as with M/M/1 queues.

After some certain steps, we obtain 'r+1' nodes with geometrical distribution and one with Poisson distribution (for the call duration server).

$$\rho_i = \begin{cases} \sum_{s=1}^{4} \frac{\lambda_N \cdot e_{is}}{\mu_{is}}, & i = 1,...,r+1 \\ \frac{\lambda_i}{\mu_i}, & i = r+2 \end{cases}; \qquad (16)$$

Equation (17) instead will suffer the following changes compared to the case with M/M/1 queues:

$$p_i(k) = P(n_i = k) = \begin{cases} \frac{(1-\rho_i) \cdot \rho_i^k}{1-\rho_i^{K+1}}, & i = 1,...,r+1; \ \ k \le K \\ 0, & i = 1,...,r+1; \ \ k > K \\ \frac{e^{-\rho_i}}{k!} \cdot \rho_i^k, & i = r+2 \end{cases} \qquad (17)$$

where n_i is the total number of jobs of all classes in node *i*. The stability condition is $\rho_i < 1$. As shown in Thesis 1.1.5, a finite call duration does not have any influence on the T_C, therefore we can neglect the effect of the node (i=r+2). The product form solution is obtained only if we have $\mu_{is} = \mu_i \ \ \forall s = 1,2,3,4$ (see eq.(16)), which is not true. Let's assume an empirical relationship between the service rates of different classes at one node (according to Thesis 1.1.2): $\frac{1}{\mu_{i1}} = 3 \cdot \frac{1}{\mu_{i2}} = 3 \cdot \frac{1}{\mu_{i3}} = 12 \cdot \frac{1}{\mu_{i4}}, \ \ i = 1,...,r,$ and denote $\frac{1}{\mu_{i1}} = \alpha_i$, which is the minimum *SETUP delay* (see Thesis 2.1.2). Moreover, because $\frac{1}{\mu_i^C} = \frac{1}{\mu_{i1}} + \frac{1}{\mu_{i2}} = \frac{4}{3} \cdot \alpha_i$, we can express ρ_i in terms of μ_i^C. The expected number of jobs of all classes appearing in queue *i* can be derived from the definition of the mean value of random variables, then applying the Little's formula we have:

$$E(T_i) = \frac{1}{\lambda_i} \cdot \sum_{k=0}^{K} k \cdot p_i(k) = \frac{1}{\lambda_i} \cdot \left[\frac{\rho_i}{1-\rho_i} - \frac{(K+1) \cdot \rho_i^{K+1}}{1-\rho_i^{K+1}} \right], \quad i = 1,...,r+1 \qquad (18)$$

The stability condition is $\rho_i < 1$. From this formula we can obtain the call establishment time and release latency for the cascaded network of *r*-nodes (let us consider the *ideal case,* when all call attempts are successful and suppose that all nodes are identical):

$$\begin{aligned} E(T_C)\big|_{'r'\text{switches}} &= \sum_{s=1,2} \sum_{i=1}^{r+1} E(T_i) = 2 \cdot r \cdot E(T_r) + E(T_{r+1}), \\ E(T_{RN})\big|_{r'\text{switches}} &= \frac{1}{2} \cdot \sum_{s=3,4} \sum_{i=1}^{r+1} E(T_i) = r \cdot E(T_r) + \frac{1}{2} \cdot E(T_{r+1}), \end{aligned} \qquad (19)$$

where T_r is the time a messages spends in one switch, $E(T_i)$=$E(T_j)$=$E(T_r)$ $\forall i,j$=1,…,r, while T_{r+1} denotes the response time of a terminal equipment (i=r+1). The maximum buffer size '*K'* has to be determined such that the formula will best approximate the measured results. An *M/M/1 infinite buffer* solution has been presented in [C-6], which is a particular case of the equation (18), i.e. when $K \to \infty$ then the second term converges to zero. It gives a good approximation of the call establishment time if the load $\lambda_N / \mu^C < 0.65$.

If there are many different types of switches in the network, or more complex calls are generated and more than one PNNI levels are defined, then the equation (19) is difficult to be applied. For this case I have constructed another approximation formula that can be used for the network design (if the T_C and T_{RN} of one node is a-priori known):

2.2.2. I have shown that the inequality in equation (10) regarding the call establishment latency of calls that go through '*r*' cascaded switches can be approximated by the following formula:

$$E(T_{CN})\big|_{'r'} \approx \begin{cases} (1+s)\cdot\dfrac{m+1}{2}\cdot[1-0.1\cdot\log(r)]\cdot\exp\left(\dfrac{\lambda_N}{r\cdot\mu_0^C}\right)\cdot\sum_{i=1}^{P}(n_i\cdot\overline{T}_{CN}^{\text{type}\,i}),\ \ r=\sum_{i=1}^{P}n_i,\ \mu_0^C=\min_{i=1..P}(\mu_i^C),\ \lambda_N<0.9\cdot\mu_0^C \\ (1+s)\cdot\dfrac{m+1}{2}\cdot[1-0.1\cdot\log(r)]\cdot\dfrac{C\cdot\left(\dfrac{\lambda_N}{\mu_0^C}\right)^2}{\sqrt{\exp\dfrac{\lambda_N}{\mu_0^C}}}\cdot\sum_{i=1}^{P}(n_i\cdot\overline{T}_{CN}^{\text{type}\,i}), \qquad \lambda_N\geq 0.9\cdot\mu_0^C \end{cases} \tag{20}$$

Similarly, for the estimation of the T_{RN}, I have found the following formula:

$$E(T_{RN})\big|_{'r'} \approx \begin{cases} [1-0.1\cdot\log(r)]\cdot\exp\left(\dfrac{\lambda_N}{r\cdot\mu_0^R}\right)\cdot\sum_{i=1}^{P}(n_i\cdot\overline{T}_{RN}^{\text{type}\,i}),\ \ r=\sum_{i=1}^{P}n_i,\ \mu_0^R=\min_{i=1..P}(\mu_i^R),\ \lambda_N<0.9\cdot\mu_0^R \\ [1-0.1\cdot\log(r)]\cdot\dfrac{C\cdot\left(\dfrac{\lambda_N}{\mu_0^R}\right)^2}{\sqrt{\exp\dfrac{\lambda_N}{\mu_0^R}}}\cdot\sum_{i=1}^{P}(n_i\cdot\overline{T}_{RN}^{\text{type}\,i}), \qquad \lambda_N\geq 0.9\cdot\mu_0^R \end{cases} \tag{21}$$

The term containing *log(r)* expresses the message overlapping, *m* represents the PNNI peer-group hierarchical level, $s=0$ for the mandatory *IE*s (default *SETUP* message), $0<s<1$, P is the number of different types of ATM switches in the network and n_i is the number of switches of one type. The (μ_i^C, μ_i^R) are the average values of the equivalent service rates of the switches for the call establishment and release, where $\mu_i^C < \mu_i^R$, and λ_N is the call arrival rate. Finally, the term $C=\exp\left(\dfrac{0.9}{r}\right)\cdot\dfrac{\sqrt{\exp(0.9)}}{(0.9)^2}$ is the normalising factor. It can be observed that the T_{RN} is independent of the call complexity and of the PNNI hierarchy, respectively.

Validation of Thesis 2.2

Both formulas of Thesis 2.2 have been validated against measured and simulated results of $r \leq 4$ cascaded switches. The model in Thesis 2.2.1 can be used as a rule of thumb, for the first estimations in network planning. The disadvantage of this method is that it gives the average value estimation only. It has lost its ATM specific characteristics due to the single FIFO queueing model. Moreover, the model cannot handle priority, because then the product form solution will disappear. The sparse matrix is already of size [16x16] for 2 cascaded switches, with one calling and one called party, respectively. Each additional node adds 4 rows and 4 columns to the matrix, while each generator adds 8 of them. The *M/M/1/K* finite buffer model gives an overestimation of the call establishment time of cascaded switches for $\lambda_N/\mu^C > 0.5$, moreover this model is also inaccurate when estimating the call release latency (it gives $T_C \approx 2T_{RN}$). In Thesis 2.2.2, I have found that the proposed approximation formula, i.e. equation (20) has acceptable accuracy in practice. Except the impact of the PNNI hierarchy, which has been only checked for $m=1$ and $m=2$, the other parameters have been validated against different types of cascaded switches (by measurement and simulation).

Performance analysis of signalling in large ATM Networks

Thesis 3 *I have provided a performance analysis of signalling traffic in real-size ATM networks.* **[C-4], [C-10], [J-4], [D-6]**

The model defined in Thesis 2.1 is quite complicated, it is not easily tractable analytically, especially when using priority mechanism or when studying a large network, therefore in the followings I have used this *simulation model* presented in Thesis 2.1 to estimate the call establishment times, release latencies and signalling load/switch in real size networks. The message latencies introduced by the signalling nodes in an ATM network are additive, as shown in Thesis 1.1.7. This property of the measurement implies that as more nodes are traversed on a signalling message path, the signalling message latency increases. The latency measured across a small number of nodes can be used *to predict* the performance of a larger network of similar nodes. Thus one can obtain the upper bound (diameter, Φ) of the network in order not to exceed the maximum latencies set by the ITU-T recommendations [ITU97].

Def 3.1 I have defined the *network diameter* as the shortest path between the two furthest nodes of the network:

$$\Phi = \max\{L(i,j) \mid L(i,j) < L'(i,j), \forall i,j \in \{1,...,N\}\}, \tag{22}$$

where *L(i,j)* is the path length between nodes *i* and *j*.

Def 3.2 I have defined the *average user density* ($\overline{D}_N$) of an *N*-node network:

$$\overline{D}_N = (1 - \frac{N_{tr}}{N}) \cdot \frac{\overline{L}}{\Phi}. \tag{23}$$

where N_{tr} is the number of transit nodes in the network (i.e., no end-user directly connected to it). The user density is an important network-level parameter, because it is indicating the distribution of the end users among the network nodes. In general, $0 < \overline{D}_N \le 1$. The bigger the value, the denser the network is. In our sample networks (N = 4, 30, 35) this parameter is $\overline{D}_4 = 0.93$, $\overline{D}_{30} = 0.59$, $\overline{D}_{35} = 0.45$, while for an isolated switch $D_1 = 1$. The applicability of this formula is dependent on how easy the average call path $\overline{L}$ can be found. I have obtained $\overline{L}$ by simulation.

Def 3.3 I have defined the *signalling load balance* of a given switch in the network:

$$\theta_i = \frac{\lambda_i - \frac{1}{N} \cdot \sum_{i=1}^{N} \lambda_i}{\max_{1 \le i \le N}(\lambda_i)}; \quad -1 \le \theta_i \le 1; \quad 1 \le i \le N \tag{24}$$

where λ_i is the signalling load of a given switch.

I have studied the call establishment times and release latencies for 2...10 cascaded switches versus signalling load. A short chain of nodes has been validated by measurements (2…4 switches). Furthermore, I have estimated these parameters for *N*=4-node *fully meshed*, *N*=7-node (having a 4-node backbone *ring*), *N*=30-node and *N*=35-node ATM networks in a *typical xDSL topology* with cca. 25 DSLAMs (Figure 6). This later topology contains all known basic topologies (e.g., access multiplexers, redundant links, backbone ring, etc.). These sample networks contain all source-destination path lengths from 2 to 8 nodes, so I could compare the call establishment times T_{CN}^{L} and release latencies T_{RN}^{L} (for all path lengths, *L(i,j)*=1,…,8) to those obtained for cascaded switches.

Thesis 3.1 I have shown that an *r*-node cascaded network can be used to estimate the signalling parameters (T_{CN}, T_{RN}, γ_R) of a large, *N*-node ($N > r$) network having a diameter $\Phi = r$ nodes.

First, I have investigated a <u>*homogeneous network*</u> (all network nodes are identical). I have shown that the minimum values of T_{CN}^{L} and T_{RN}^{L} for each path-length (L) are the same in both network topologies. The average and the maximum values of T_{CN}^{L} and T_{RN}^{L} for the cascaded topology slightly overestimate those of the network's, i.e. with 1-5% the average values and with 15-20% the maximum values respectively, for a range of network load $0 < \lambda_N < \lambda_N^{\max}$, where $\lambda_N^{\max}$ is the maximum network-level call arrival rate without any call rejection, N is the number of nodes.

It is quite obvious that the maximum network-level call arrival rate $\lambda_N^{\max}$ depends on the network topology as follows:

$$\lambda_1^{\max} \approx \lambda_r^{\max} \le \lambda_N^{\max} \quad \text{or} \quad \max\{\lambda \,|\, r \ge 1, \text{cascade}, \gamma_R = 1\} \le \max\{\lambda \,|\, \Phi(\text{N - node}) = r > 1, \gamma_R = 1\}$$

The term $\lambda_1^{\max}$ is the maximum call arrival rate measured in one isolated switch without any call rejection (γ_R=1). The average call path is defined by the following equation: $\overline{L} = \sum_{r=1}^{\Phi} p_r \cdot r$, where p_r is the probability that a call goes through r nodes.

Thesis 3.2 I have found that $\lambda_N^{\max}$ can be approximated analytically by the following formula:

$$\lambda_N^{\max} \approx \left[1 - \frac{\theta_N^{\max} \cdot (1 + N_{tr})}{N} + \left| \frac{N_{ISP} - 1}{N_{ISP} + 1} \right| \cdot \frac{\Phi}{(1 - \frac{N_{tr}}{N}) \cdot \overline{L}} \right] \cdot \frac{\lambda_1^{\max}}{1 + \frac{\overline{s}}{2 - p_s}}, \tag{25}$$

where, N_{ISP} is the number of internet access servers (ISP = Internet Service Provider),

$\overline{s} = \frac{1}{m} \cdot \sum_{i=1}^{m} s_i$ is the average signalling overhead for the "complex" call profile,

m is the number of different types of "complex" call profiles,

p_s is the probability that a call has a "complex" call profile.

Secondly, I have studied the behaviour of the call processing in a <u>*hybrid network*</u> (not all network nodes are identical). In practice, we have often found two or three types of ATM switches in one network. In this case, the formula given in equation (25) will suffer some certain changes:

$$\lambda_N^{\max} \approx \left[\max[\min_{i=1..P}(\lambda_1^{\max})^{\text{type } i}, \min_{j=1..R}(\lambda_1^{\max})^{\text{type } j}] \cdot \left(1 - \frac{\theta_N^{\max}}{N/(1 + N_{tr})} \right) + \left| \frac{N_{ISP} - 1}{N_{ISP} + 1} \right| \cdot \frac{\min_{i=1..P}(\lambda_1^{\max})^{\text{type } i}}{\overline{D}_N} \right] \cdot \frac{1}{1 + \frac{\overline{s}}{2 - p_s}} \tag{26}$$

where $(\lambda_1^{\max})^{\text{type } j}$ is the maximum call arrival rate of the switch connected to an ISP server (j=1..R).

Next, I have studied the evolution of the T_C and T_{RN} in the N-node network when instead of default *SETUP* messages (containing only the mandatory information elements) more complex calls are

injected into the network. The complex call profile (*CCP*) can be described by a parameter '*s'*, where $0 < s < 1$ (see Thesis 1.1.4).

Thesis 3.3 I have shown that, if $T_C^{CCP} = (1+s) \cdot T_C^{default}$ for one node (see Thesis 1.1.4)**, then the average call establishment time of the network will be increased to**

$$\overline{T}_C^{CCP} = (1 + \frac{\overline{s}}{2 - p_S}) \cdot \overline{T}_C^{default}, \quad (27)$$

while the average call release latency $\overline{T}_{RN}$ remains unchanged:

$$\overline{T}_{RN}^{CCP} = \overline{T}_{RN}^{default}. \quad (28)$$

p_S is the probability that the generated calls are "more" complex (according to Table 2). When all calls in the network are of the same complexity then the $p_S = 1$, when none of them is complex then $p_S = s = 0$. Thesis 3.3 is especially important in the network design to avoid overload situations and underlines again the differences in the behaviour between T_C and T_{RN}.

Let us note the equivalent service rates for the call establishment of a node $\mu^C = \mu_{eq}^{setup}$, and for the call release $\mu^R = \mu_{eq}^{release}$, respectively.

Thesis 3.4 I have shown that replacing one node of a network ($(\mu_i^C; \mu_i^R)$, *i=1,...,N*) by a switch having a signalling service rate $(\mu_0^C \le \frac{1}{\Phi + 1} \cdot \min_{i=1..N} \mu^C; \mu_0^R \le \frac{1}{\Phi + 1} \cdot \min_{i=1..N} \mu^R)$, can be used to determine the signalling load of this node. Replacing each node one-by-one will identify the bottle-neck nodes in a given network topology.

If it is not possible to monitor and analyse each network node individually, then the method described above helps us to obtain the signalling loads. All we need is to replace one node with the above given parameters and run the simulation at very low call arrival rate to obtain the call establishment times and release latencies. Then analysing the results, we should see that those calls with $T_C > \frac{1}{\mu_0^C}$ are processed by the test node as well. The number of these calls over the total number of generated calls gives the load of the investigated node. Thus, the signalling load of this node can be defined as follows: $\lambda_i = \lambda_N \cdot p_i$, i=1,...,*N,* where p_i is the probability that the call will be processed by node i. The bottle-neck node in the network has its load equal to:

$$\lambda_i^{\max} = \max\{\lambda_i | \lambda_i \ge \lambda_j, \forall i, j = 1,..., N\}.$$

Validation of Thesis 3

We have considered that the call arrival rate is poissonian and the source-destination pairs are either uniformly distributed in the network (for LANE) or there are centrally located servers (ISP1,2,3) for Internet connectivity (see Figure 6). Simulations have been carried out for the following network-level call arrival rates: 1, 5, 33, 100, 200, 400, 500 calls/sec. Each time 10000 calls have been evaluated. The simulation results have been validated for 2, 3 and 4 cascaded switches against measured data (e.g., *FORE ASX200BX* and *SEABRIDGE XP140*, but see also [Mau01] for mixed types). Furthermore, I have implemented by simulation the *TEN-155* network scenario as described in [Nov99], and validated my results against those measured data of T_C. In addition, I have compared my results to the simulated network results in [Gel97] and I have shown that my model provides better results than that one.

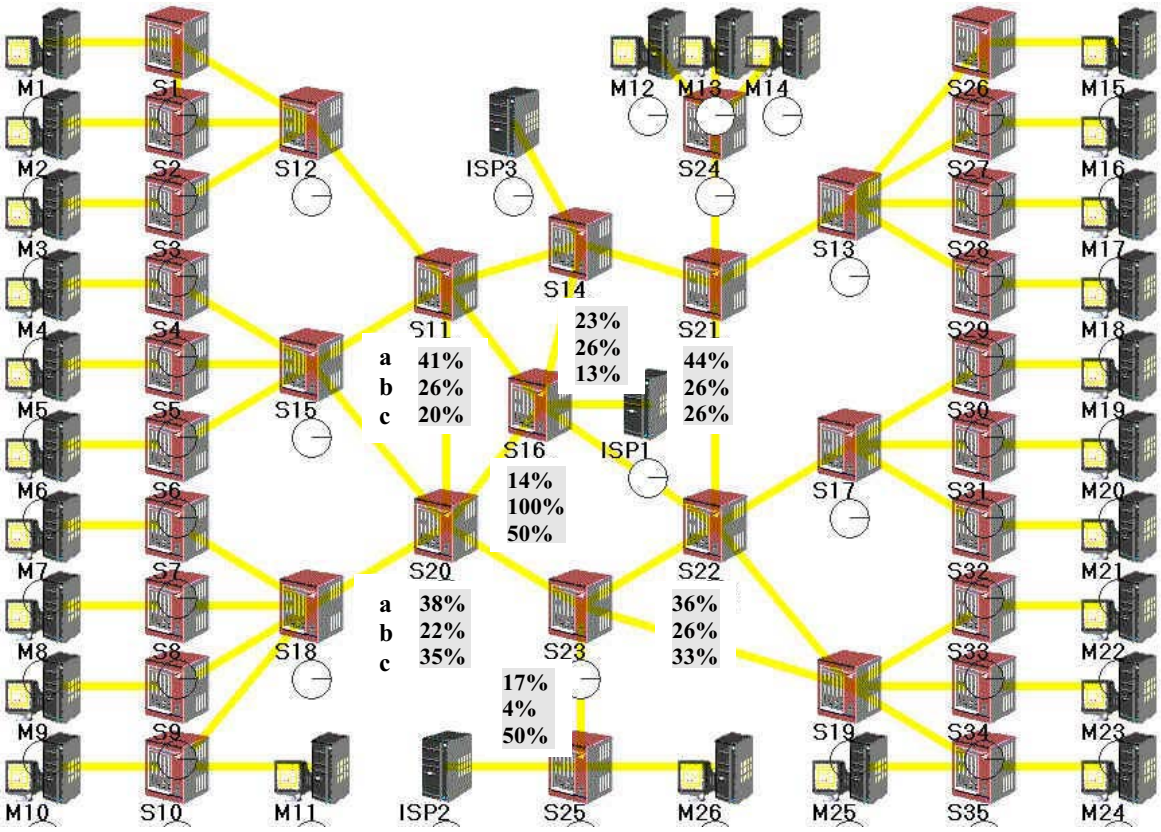

Figure 6 Measuring the relative signalling load (p_i) in the core of the sample 35-node network
a) LANE service (1st value); b) Internet access via one ISP (2nd); c) Internet access via 2 ISPs (3rd)

The importance of the Thesis 3.1 is that independent of the network topology, a cascaded network always (slightly) over-estimates the mean and maximum of the call establishment times and release latencies in the network. The formula in Thesis 3.2 is a conservative one, and it has been validated by simulating many different network scenarios and applications. The formula can be used in IP networks as well, when PPP connectivity is used for connections to ISP servers. In Thesis 3.3, I have pointed out again the different behaviour of the T_C and T_{RN} in an N-node network for complex calls. A case study for the impact of link/node failures and network applications on the signalling load of a 35-node network has been investigated and presented in [J-4] and [D-6].

Optimisation of the call processing architecture of the AAL2 switch in UMTS networks

Thesis 4 *I have constructed a new model for AAL2 signalling nodes and I have given a performance analysis of this model for FIFO and priority queueing.*

The ATM adaptation layer type 2 (AAL2) defined in [AALI97] has been selected as the transmission technology in the landline part of the radio access network of UMTS systems [Ene99]. To suit AAL2 to a network where support for soft handoff is essential, an additional switching level on top of ATM has been developed. AAL2 Signalling is a new protocol which is capable of handling on-demand, switched AAL2 connections [AALQ99]. It makes possible carrying small data packets on top of an ATM infrastructure with low delay while using the bandwidth efficiently [Bal97]. In this thesis, I am dealing with certain design considerations of an AAL2 signalling point and propose techniques to further decrease the AAL2 connection establishment time for soft handoff legs [Won97]. It should be mentioned that at the time of writing and publishing my results (see [C-7]) there was no AAL2 switch available on the market (1999-2000).

Thesis 4.1 *I have constructed a new model and a new algorithm to obtain an optimised AAL2 signalling node architecture.* **[C-7], [D-7]**

The architecture of the model

4.1.1. I have shown that the message flow model described in Thesis 2.1 can be applied to model an AAL2 signalling node with certain re-design according to the protocol specification in [AALQ99].

A re-design means, that there is no need for complex call profile process (CCP) in the model, the messages have shorter formats and different notations (see Figure 3). Other features are similar to the description in Thesis 2.1.

Parameter settings of the model

The problem can be formulated as follows: given that the overall capacity of the signalling processor is constant, we have to (re)allocate the service times of each process in a way *to minimise* either the AAL2 connection establishment time T_C or the overall handling time T_H (see its definition in Thesis 1.1.1). A single AAL2 switch and two connection endpoints are used in the optimisation process, the Soft Hand-Off device (SHO) can only initiate and the Base Station will terminate the connections. The connection arrival process is poissonian, and the distribution of the connection holding time is exponential.

4.1.2. I have developed a simulation-based iterative algorithm *QUE*, optimising the allocation of processor resources in order to minimise the call establishment times and release latencies of soft handoffs in UMTS networks.

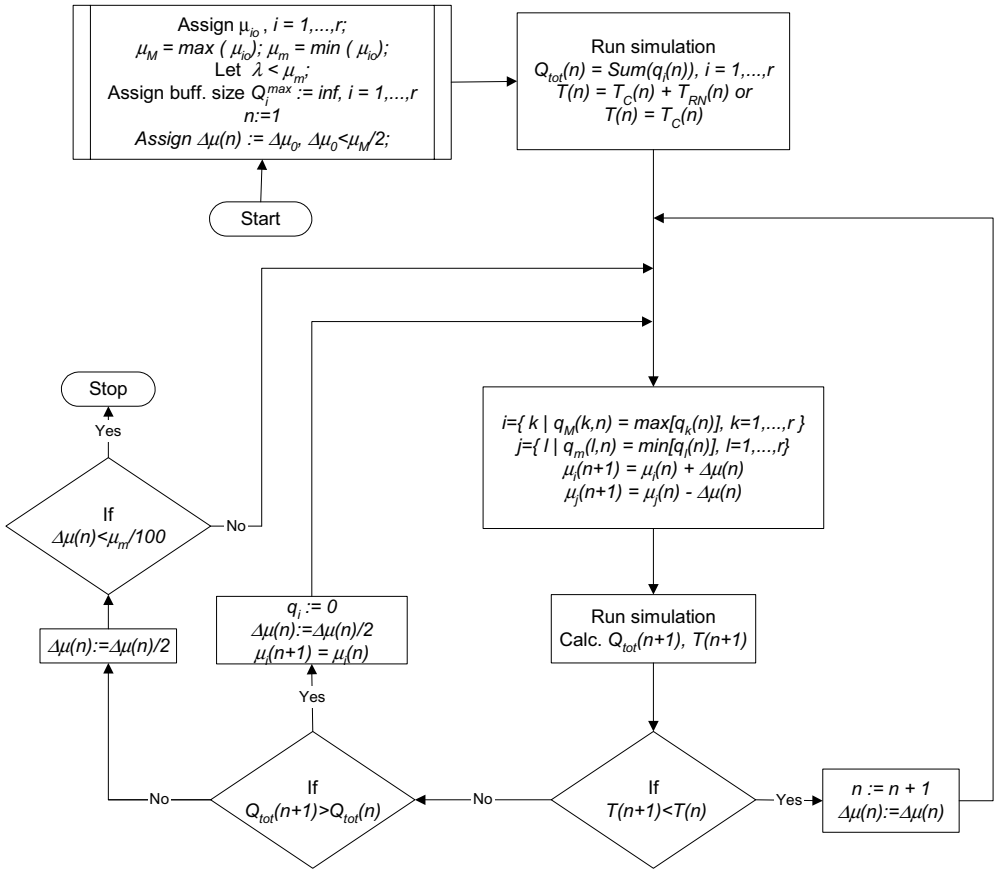

Figure 7 Block diagram of the optimisation algorithm QUE (FIFO queueing)

The algorithm (see Figure 7) starts with assigning an initial service rate to each process, sets the connection arrival rate and the step size $\Delta\mu$. The buffer sizes are set to infinity, to avoid connection rejection due to buffer overflow. The second phase is to run the simulation, and calculate $Q_{tot}(n)$ and $T(n)$, where n represents the number of correct steps. $Q_{tot}(n)$ is the sum of the average queue length measured in the buffers of individual processes in step n. $T(n)$ is equal to the connection establishment time, or to the overall connection handling time in step n, depending on the optimisation criterion. In the next phase, the processor capacity is reallocated in a way that the service rate of the busiest process is increased by $\Delta\mu$, and the service rate of the process with the shortest average queue length is decreased by the same value. Thus the overall processor capacity is kept constant. We repeat the simulation, and calculate $Q_{tot}(n+1)$ and $T(n+1)$. If the last step improved the performance ($T(n+1)<T(n)$), the step is considered correct, n is incremented, and the capacity allocation is updated. If $T(n+1)>T(n)$, then the step size ($\Delta\mu$) is halved subsequently, until the readjusted capacity allocation results in $T(n+1)<T(n)$, or we arrive to the predefined minimum value of $\Delta\mu$, when the algorithm is stopped. The comparison of $Q_{tot}(n+1)$ to $Q_{tot}(n)$ allows the algorithm to find the global optimum point instead of a local one.

To further decrease the call establishment time, *I have introduced prioritised handling of signalling messages* in the AAL2 switch. Particularly, I have assigned higher priority to the messages involved in the connection establishment phase. The optimisation algorithm QUE will not always find the global optimum if priority handling is used, therefore another criterion should be applied.

4.1.3. I have shown that replacing $Q_{tot}(n)=\sum_{i=1}^{r} q_i(n)$ by $Q_{tot}(n)=\sum_{i=1}^{r}\frac{q_i(n)}{\mu_i(n)}$ will enhance the performance of the algorithm QUE and will be applicable to both FIFO queueing and priority handling of the signalling messages (algorithm *QUE/MU*).

While the algorithm QUE minimises the sum of the queue lengths of processes, the algorithm QUE/MU minimises the average time the jobs spend in one node. Therefore in case of FIFO queueing we obtained very closed results. The definition of a global minimum depends on the objectives: do we want to obtain a minimum for the call establishment time or for the overall handling time? Figure 8a compares the results of the above two algorithms obtained step-by-step when no priority for messages is applied. The call establishment time is shorter in the case of algorithm QUE/MU, but the call release time is longer at the same time, so that the overall handling time is very close in both cases. I have chosen the starting point of the algorithms being the *equilibrium state* (reference point), when the initial distribution of the resources is uniform among the five processes, the total capacity was C (e.g., C=166 instructions/time unit, see Figure 8a). The result delivered by the algorithm is independent of the processor capacity.

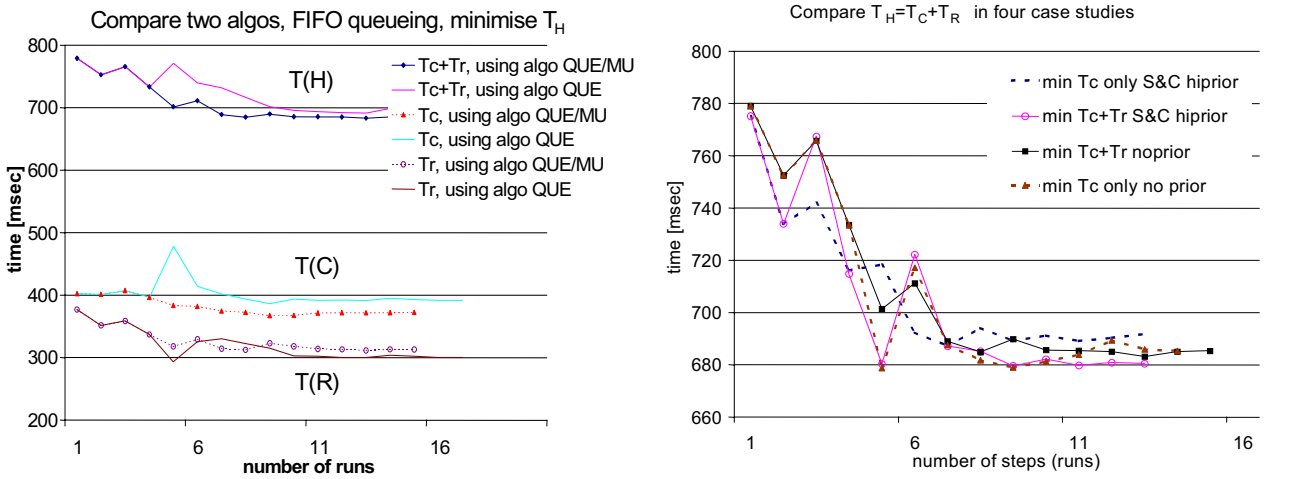

Figure 8 a) Effect of two algorithms on T_C, T_H and T_{RN}; b) Case studies for minimising T_H

Thesis 4.2 *I have given a performance evaluation of the optimisation algorithm and of the priority handling scheme.* **[C-7], [D-7]**

I have investigated the distribution of the processor resources in the following four cases (see Table 4), using the optimisation algorithm QUE/MU. I have not investigated the case when priority for messages in the release phase is given, because there was no argument in practice for such scenario.

Table 4 Definition of four case studies

Priority handling	Minimise	
	T_C of an AAL2 connection	T_H of an AAL2 connection
no priority (FIFO)	Case 1	Case 2
priority for setup phase	Case 3	Case 4
priority for release phase	n.a.	n.a.

According to Thesis 2, the latencies in the network are proportional to the delays in one switch, therefore the following investigations refer to one isolated switch. The reference point (*equilibrium state*) is not the worst case situation, but it was obvious to select for the starting point a state with uniformly distributed resources among the processes (i=1,...,5).

4.2.1. I have found that prioritised handling of the signalling messages *ERQ* (≡ *SETUP*) **and *ECF*** (≡ *CONNECT*) **results in a noticeable 8-12% decrease of the call establishment time T_C** (in any state). However, **at the same time, an increase in the call release latency T_{RN} of 9-11% can be also observed.**

The decrease of T_C by introducing priorities is more than what could be achieved by applying the algorithm in any case. As a consequence of the Thesis 4.2.1, the overall handling time T_H does not change when introducing priorities compared to the values of the *FIFO* case (ΔT_H ≈1-3%, $\forall n$=1,...,16, see Figure 8b). Based on these results, **I have concluded that the bandwidth utilisation on the outgoing link cannot be increased by introducing priority mechanism for the setup phase, as the overall handling time T_H remains constant.**

Having a detailed look at our case study, I have found that applying the algorithm QUE/MU to all four cases, the call establishment time T_C is reduced by 10% if no priority was applied (case 1&2), 6% in the case 3 and only 3% in the case 4. Furthermore, in the case 3 we obtained the maximum gain in reducing the call establishment time T_C, but the minimum gain in reducing the overall handling time T_H.

4.2.2. I have shown that applying the algorithm QUE/MU to the cases 1 to 4, leads to different optimum points, but in all these points the overall handling time T_H is at least 15% less than in the *equilibrium state* (reference point)**.**

The call release time T_R is decreased by 25–32% in all cases compared to the equilibrium state (but e.g., the gain in the case 3 is with 7% less than in the case 4). The call release time T_R is significantly decreased, even though the algorithm is aiming to reduce the time needed for the connection establishment. This is due to the fact that the setup and release messages partly travel along the same path within the processor.

Validation of Thesis 4

We have tested both algorithms starting from different initial states (extreme states as well), in all cases the three parameters (T_C , T_R and T_H) converged to the same optimum values, only the convergence speed varied (from 10 to 20 steps). The algorithm converges usually in 13-15 steps, in

worst case (starting from *EXTREME_x* values) this could go up to 20 steps. Also when trying for different capacities *C*, the ratios in the final distribution among the processes remained the same. The algorithm QUE should not be applied for the case when certain messages get higher priority, because it will not necessarily deliver the global optimum. Instead, the algorithm QUE/MU converges always to the same values, regardless of *FIFO* queueing or priority handling. I haven't found any similar paper in the literature to compare these algorithms to. [Wu97] presented a model for the setup phase only and minimised it using the lagrangean method, however it did not presented any numerical results of this optimisation method.

The importance of Thesis 4.2.1 is to show that the application of a priority mechanism for setup messages reduces the call establishment time and thus advantageous for handling soft handoff procedures, but cannot save bandwidth on the link, because of the increased release latency. In the same time the Thesis 4.2.2 points out that applying the algorithm QUE/MU the processor states can be optimised so that it will reduce the overall handling time significantly, thus reducing the bandwidth usage on the link. In addition, network level simulations have been carried out with the optimised processor architecture (of case 3 and 4) for the UTRAN network (for both flat and tree topologies, see [C-7]). My overall recommendation is to use both mechanisms (introducing priority handling and applying the algorithm QUE/MU) in a UTRAN network design.

Application of the Blocked Call Queueing (*BCQ*) mechanism to wide-band calls in ATM networks

Thesis 5 I have introduced a new mechanism (BCQ) in access nodes of ATM networks and I have given a performance analysis of this queueing mechanism. **[C-2], [C-3], [C-4], [D-8]**

The Blocked Call Queueing mechanism holds some blocked calls by storing their signalling information in a separate buffer at the access ATM switch. These calls will be later connected when network resources become available. The call blocking probability of wide-band calls is much higher than that of narrow-band calls' [Ber96]. The introduction of queueing the wide-band calls rather than reject is going to reduce the call blocking probability of wide-band calls [Bla96], at the expense of a small increase in setup time and call blocking probability of narrow-band calls. My contribution is to investigate the possibility of supporting this BCQ mechanism by using signalling protocols (i.e., proposal for extension of existing protocols). The signalling overload associated with wide-band *BCQ* and questions related to grade of service are also in focus of my investigations. I was interested only in a part of the repeated call attempts, when rejection of wide-band calls is caused by the network (neglecting those refused by the called party). According to my proposal, in such a case there is no need for any action by the calling party.

Thesis 5.1 *I have shown that it is possible to apply the Blocked Call Queueing (BCQ) mechanism by extension of the current signalling protocols [Q2931], [UNI40]. I have introduced a new state (U*), a new message (CALL QUEUED) and a new timer (T3xx) to implement the BCQ mechanism (see Figure 9).* **[C-2], [C-4], [D-8]**

The extension details are presented in [C-2] and [D-8]. The implementation of wide-band *BCQ* mechanism is quite simple, no complexity arises when extending the flow charts based on [Q2931], [UNI40]. Five original states are affected by introducing a new state for BCQ. All the new elements are represented in Figure 9. I have extended our simulation model with this *BCQ* mechanism for the access nodes and activated for the blocked wide-band calls. While in Thesis 1, 2, 3 and 4 I have

supposed that all call attempts are successful, in this thesis I have analysed the bandwidth requirements of calls, and focused especially on the blocked wide-band calls.

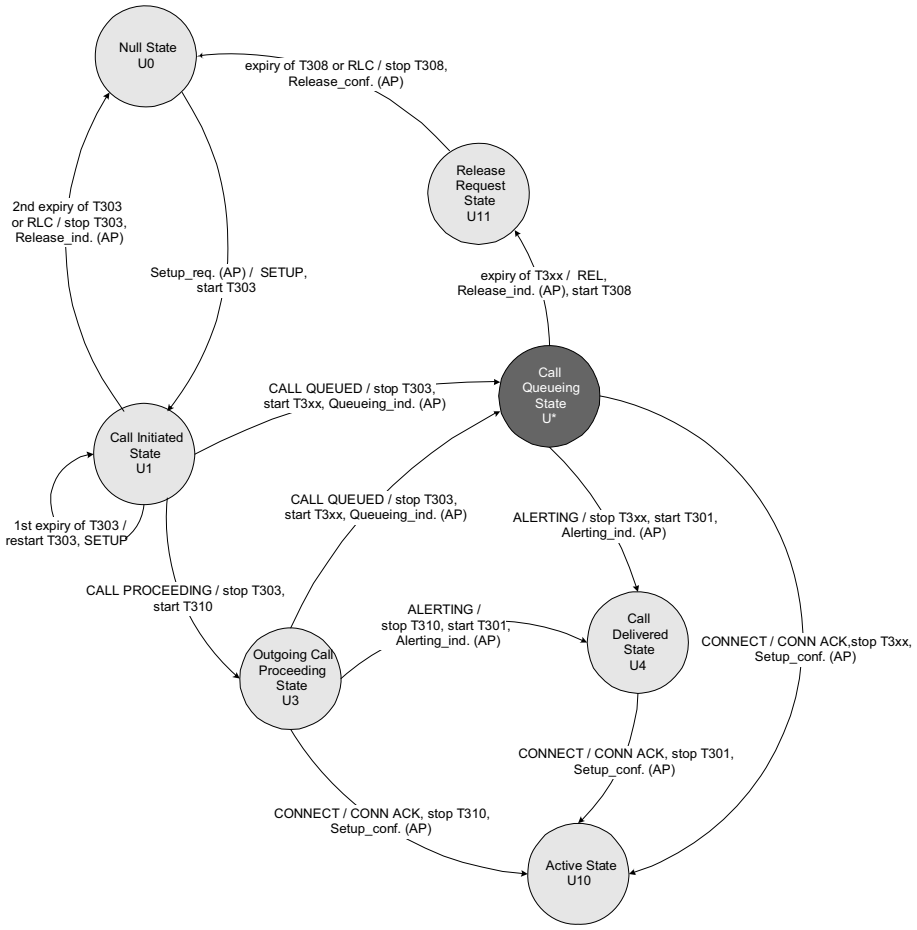

Figure 9 FSM graph (user side of the UNI) for the BCQ mechanism (simplified graph)

Thesis 5.2 *I have given a performance evaluation of the Blocked Call Queueing (BCQ) mechanism.* **[C-2], [C-3], [C-4], [D-8]**

In this thesis I have tried to find an analytical solution to the call establishment time of the blocked and retransmitted wide-band calls. The study of the call release time is out of interest, as this parameter is identical to the non-blocked calls presented in the previous theses.

5.2.1. I have given an analytical solution for lower and upper bounds of call establishment time of the blocked and retransmitted wide-band calls (T_C^{BCQ}):

$$ET_C^0(i,j) \le ET_C^{BCQ}(i,j) \le ET_C^0(i,j) + T_{3xx} \tag{29}$$

where: $ET_C^0(i,j)$ is the expected call establishment time of narrow-band calls along the main path between nodes *i* and *j*. The new timer (T_{3xx}) for the blocked calls has to be shorter than the time defined for the grade of service parameter, $T_{3xx}<T_{GoS}$, but long enough to allow a couple of retrials. Suppose that *E(N)* is the expected number of retrials and d_n is the expected time spent in the access queue at the n^{th} retrial.

$$N = \max\{n : t_1 + t_2 + ... + t_n < T_{3xx}\},$$

where: $t_1 = d_1 + T_C^1(i,j); t_2 = d_2 + T_C^2(i,j); ...; t_n = d_n + T_C^n(i,j)$, and $T_C^k(i,j)$ is the average call establishment time of a narrow-band call between nodes *i* and *j* along the alternative path $\pi_k(i,j)$, $k=1,...,n$.

According to Lorden's Theorem [Lord70], we can find lower and upper bounds for *E(N)*:

$$\frac{T_{3xx}}{E\left[T_C^0(i,j)+d_0\right]}-1\le E(N)\le\frac{T_{3xx}}{E\left[T_C^0(i,j)+d_0\right]}+\frac{E\left[T_C^0(i,j)+d_0\right]^2}{\left[E\left[T_C^0(i,j)+d_0\right]\right]^2}$$

where $T_C^0(i,j)$ is the average call establishment time of a narrow-band call between nodes *i* and *j* along the main path $\pi_0(i,j)$ and d_0 is the expected time spent in the access queue at any retrial (using BCQ).

5.2.2. Furthermore, conditional on *N=n*, **I have found an approximation formula for the expected call establishment time when using wide-band *BCQ* mechanism:**

$$ET_C^{BCQ}(i,j)\approx(1-p_0^{n+1})\cdot T_C^0(i,j)+\frac{p_0}{1-p_0}\cdot(1-p_0^n)\cdot d_0\,, \tag{30}$$

where p_0 is the call blocking probability of wide-band calls along the main path $\pi_0(i,j)$ between nodes *i* and *j*, $T_C^0(i,j)$ is the expected call establishment time of narrow-band calls between nodes *i* and *j* along the main path $\pi_0(i,j)$, and d_0 is the waiting time between two trials at the access node after a wide-band call being rejected. Each of these terms can be calculated as follows:

$$T_C^0(i,j)=2\cdot\sum_{k\in\pi_0(i,j)}E_{kZ}(m)\cdot E_k(T)=\frac{2}{\mu^C}\cdot\sum_{k\in\pi_0(i,j)}E_{kZ}(m) \tag{31}$$

while, similarly (supposing that after blocking, the rejected call enters again the end of the queue in the access node after a waiting time w_t equal to the expected mean holding time of call type *'t': $w_t=E(MHT_t)$*):

$$d_0=w_t+\frac{1}{\mu^C}\cdot\rho_A\cdot(1-p_t) \tag{32}$$

where: ρ_A is the utilisation of one access node, μ^C is the equivalent service time of one node in a homogeneous network and p_t is the average link blocking probability of a wide-band call of type '*t*'. If p_{kZ} is defined as the call blocking probability on the link *k* directed from node *Z*, then:

$$p_0=1-\prod_{k\in\pi_0(i,j)}(1-p_{kZ})=1-(1-p_t)^{\bar{L}(i,j)}.$$

To find the probabilities p_{kZ} for all links is a non-trivial problem. Here we admit that the backbone and external network have the same call blocking probability 'p_t' for a wide-band call of type *'t'*.

Finally, after some steps we achieve:

$$ET_C^{BCQ}(i,j)\approx(1-p_0^{n+1})\cdot\frac{2}{\mu^C}\cdot(1-p_t)\cdot\sum_{k\in\pi_0(i,j)}\rho_{kZ}+\frac{p_0}{1-p_0}\cdot(1-p_0^n)\cdot\left[w_t+\frac{1}{\mu^C}\cdot\rho_A\cdot(1-p_t)\right]. \tag{33}$$

Numerical examples are given in [D-8] for different values of the parameters p_t, w_t, n. Throughout my investigations, I have shown that the wide-band *BCQ* mechanism does not introduce any significant signalling overload for moderate call load conditions. Finally, I have investigated the range of usability of the *BCQ* mechanism by simulations.

5.2.3. I have concluded that the wide-band *BCQ* mechanism is most beneficial when $0.5\le\rho_{kZ}\le0.67$ and the presence of wide-band calls in the network is higher than 10 percent. When $\rho_{kZ}<0.5$ the *BCQ* is not needed, while for $\rho_{kZ}>0.67$ the *BCQ* is not effective (T_C^{WB} >>, p_0>>).

Validation of Thesis 5.2

To validate my analytical results, I have conducted simulation studies to obtain the call establishment time of narrow-band and wide-band calls, call blocking probabilities (p_0) and queue lengths for two different scenarios of call mixture in a number of network topologies using the wide-band *BCQ* mechanism. These case studies are described in details in [C-4] and [D-8]. I have shown that using wide-band *BCQ*, the blocking probability p_0^{WB} is significantly reduced (approx. 50%) for wide-band calls (2-10Mbps) at the expense of a small increase (max. 20%) for narrow-band calls (0.1Mbps) in a large network configuration. Thus an increase in network revenue is obtained, which is advantageous for both users and network operators. The dependency on the utilisation of the access nodes, call distribution and mean holding time is also shown in [C-4]. Furthermore, I have shown that the average call establishment time (T_C^{WB}) is not significantly longer (<10%) in case of *BCQ*, compared to the case without using the queueing mechanism for $p_0 \leq 0.05$, independent of the lengths of w_t. The applicability of the formula in the equation (33) is limited by the fact that it is difficult to obtain the probabilities p_{kZ} on the links.

5. Applications of the new results

The measurement results presented in Thesis 1 can be widely used by network operators to design their ATM signalling network. With the rapid expansion of the *x*DSL applications today (e.g., in Europe, USA and South Korea) between residential customers and business users (SMEs), there is more and more obvious to introduce signalling in the existing ATM networks. The today's applications are either PPPoA or PVCs. One of the driving factors is that full provisioning with PVCs over a single STM-1 path (155Mbps) from the DSL Access Multiplexer (DSLAM) to the ATM switch is not possible for more than 200 customer lines each having 0.66Mbps CBR channel for its voice traffic (8x64kbps channels). The bandwidth needs of the data channels (UBR) are not even considered here. In such a case the most economical solution is the introduction of ATM signalling, because there is no need for hardware change.

The method and the "population-diagram" described in Thesis 1.2.1 has been widely used by the author and his colleagues at the 'System Test Department' at Siemens AG in their every-day work by testing Voice over DSL, Voice over IP (MGCP, H.323, SIP) and virtual trunking applications in the new generation VoIP network. The 'population-diagram' is in a patent pending state at the German Patent Office (see [T-2]).

The model described in Thesis 2 has been adapted in Thesis 4 to be used for network design in the radio access network of UMTS systems. These investigations are important, as I have shown that our optimised AAL2 switch model achieves the fastest possible connection setup, which is the most important requirement when it comes to supporting soft handoff. This model can be a basis for AAL2 switch implementations.

Another application of ATM signalling is created by the flexible bandwidth allocation on demand for efficient Video Contribution Services between TV-Studios and Broadcast Centers. The applicability of Thesis 5 is quite obvious here for blocked wide-band calls.

6. References

[AALI97] ITU-T Recommendation, *"B-ISDN ATM Adaptation Layer Type 2 Specification"*, I.363.2, Geneva, Switzerland, 1997

[AALQ99] Draft ITU–T Recommendation, *"AAL type 2 Signalling Protocol (Capability Set 1)"*, Q.2630.1, March 1999

[ATMF00] ATM Forum Technical Committee, *"UNI Signalling Performance Test Suite"*, (ed. J. Orvis, A Francis), *af-test-0158.000*, October 2000

[Baf93] M. Bafutto, P.J. Kuhn, G. Willmann, *"Modelling and Performance Analysis of Common Channel Signalling Networks"*, Hirzel-Verlag, AEU, Vol.47, No.5/6, 1993, pp. 411-419

[Ber96] S.A. Berezner, A.E. Krzesinski, *"Call queueing in circuit switched networks"*, *Telecommunication Systems* No.6, 1996, pp. 147-160

[Bal97] J. H. Baldwin, B. H. Bharucha, B. T. Doshi, S. Dravida, *"AAL-2 – A New ATM Adaptation Layer for Small Packet Encapsulation and Multiplexing"*, Bell Labs Tech Journal, April 1997, pp. 111-131

[BCMP75] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios, *"Open, Closed and Mixed Networks of Queues with Different Classes of Customers"*, *Journal of ACM*, 22 (2), 1975, pp. 248-260

[Bla96] S. Blaabjerg, G. Fodor, A.T. Andersen, *"Reducing Wide Band Blocking by Allowing Wide Band Calls to Queue"*, *COST 244 Technical Report*, TD(1996/14)

[Ene99] G. Eneroth, G. Fodor, G. Leijonhufvud, A. Rácz, I. Szabo, *"Applying ATM/AAL2 as a Switching Technology in Third-Generation Mobile Access Networks"*, IEEE Communications Magazine, June 1999, pp. 112-122

[Far01] S. Farraposo, E. Monteiro, *"Evaluating PNNI Performance", Proceedings* of 4th IEEE International Conference on ATM and High Speed Intelligent Internet Symposium, ICATM, Seoul, South Korea, April 22-25, 2001, pp. 295-299

[Gel97] E. Gelenbe, S. Kotia, D. Krauss, *"Call Establishment Overload in Large ATM Networks"*, *Proceedings* of the ATM'97 Workshop, Lisbon, Portugal, May 26-28, 1997, pp. 560-569

[Kaus97] A. Kaushal, S. Shumate, R. Hill, S. Murthy, D. Niehaus, V. Sirkay, B. Edwards, *"Performance Benchmarking of ATM Signaling Software"*, *Proceeding* of OPENSIG Workshop, Columbia University, USA, October 1997, http://www.ittc.ukans.edu/~niehaus/

[Lord70] G. Lorden, *"On excess over the boundary",* Annals of Mathematical Statistics, Vol. 41, 1970, pp. 520-527

[Mau01] M. Maurogiorgis, N. Papadoukakis, E. Sykas, G. Tselikis, *"ATM Signalling Overview and Performance Measurements in a Local Area ATM Network", IEEE Symposium on Computers and Communications, ISCC'01,* Hammamet, Tunisia, July 3-5, 2001, pp.635-640

[Mer00] A. Mertz, M. Pollakowski, "*xDSL & Access Networks", Grundlagen, Technik und Einsatzaspekte von HDSL, ADSL und VDSL, Kapitel 6, Prentice Hall,* ISBN 3-8272-9593-9, Germany, 2000

[Nie97] D. Niehaus, A. Battou, A. McFarland, B. Decina, H. Dardy, V. Sirkay, B. Edwards, *"Performance Benchmarking of Signaling in ATM Networks," IEEE Communications Magazine*, Vol. 35 No. 8, August 1997, pp. 134-143

[Nov99] J. Novak, A. Pouélé "Interim Report on the Results of the Quantum Test Programme", November 1999, http://www.dante.net/quantum/qtp/QUA-99-070.pdf

[Pil99] R.Pillai, K.Su, J.Biswas, C.Tham "*Call Performance Studies on ATM Forum UNI Signalling Implementations*", Computer Communications, Elsevier Science, Vol. 22, Issue 5, April 1999, pp. 463-469

[PNNI] ATM Forum Technical Committee, *"PNNI Draft Specification",* Ver.1.0, *94-0471R11*, 1994

[Q2931] ITU-T Recommendation Q.2931, *"B-ISDN. Dig. Subscriber Sign. Syst. No.2 (DSS2). UNI Layer 3 Specification for Basic Call/Connection Control", COM 11-R 78-E*, October 1994

[UNI40] ATM Forum Technical Committee, *"ATM User-Network Interface (UNI) Signalling Specification",* Version 4.0, *ATM Forum/95-1434R8*, April 1996

[Won97] D. Wong, T. J. Lim, "Soft Handoffs in CDMA Mobile Systems", IEEE Personal Communications, December, 1997

[Wu97] C. S. Wu, J. C. Jiau, K.J. Chen, M. Choy, *"Minimizing Call Setup Delay in ATM Networks via Optimal Processing Capacity Allocation"*, *IEEE Comm. Letters*, Vol.2, No.4, April 1998, pp. 110-113

7. Own Publications

[J] ***Journal papers***

[J-5] S. Székely, G. Szűcs, S. Maly: Performance Benchmarking of Data Transmission over Symmetrical DSL Networks, Communications Magazine, Special Issue on Broadband Access, re-submitted on June 30, 2004 for a 2nd review, at the first review being rejected, number of pages: 13, **<L>** 0.3x1.5=**0.5p**

[J-4] S. Székely, G. Szűcs, Cs. Simon, I. Moldován, S. Molnár: On the Impact of Link/Node Failures and Network Applications on the Load and Call Processing Times in ATM Networks, Periodica Polytechnica Electrical Engineering, Journal of the B.U.T.E ., Vol.46, No.1-2, 2002, pp.99-122, **<LH1>** 0.5x4=**2p**

[J-3] O. Pop, S. Székely, M. Nafornita: Performance Evaluation of Forward Error Correction Schemes in ATM Networks, Buletin Stiintific of T.U. Timisoara, 42 (56), 1997, (in Romanian), pp. 89-98, 0.3x3=**1p**

[J-2] I. Cselényi, P. Haraszti, S. Székely: The First Hungarian ATM Pilot Network, Selected Papers from the Hungarian Telecommunication Periodical, 1997, pp. 6-10, **<LH1>** 0.3x2=**0.6p**

[J-1] M. Otesteanu, S. Székely: User-Network Interfacing Problems in ATM Networks, Buletin Stiintific of T.U. Timisoara, Special Edition, 40 (54), 1995, (in English), pp. 153-160, 0.5x3=**1.5p**

[C] ***Conference and Workshop papers***

[C-10] S. Székely: Signalling Performance Evaluation of Large ATM Networks Based on Performance Measurements of Isolated Switches, 7th IEEE Symposium on Computers and Communications, ISCC'02, Taormina, Italy, 27-30 July 2002, pp.670-675, *avail. online in IEEE DL, www.ieee.org,* **<LRE>** 1x4=**4p**

[C-9] S. Székely, G. Szűcs, Cs. Simon: Modelling of Call Processing in ATM Switches Based on Performance Measurements, Proc. of the IEEE International Conference on Telecommunications, ICT'01, Bucharest, Romania, 4-7 July 2001, pp. 327-335, **<LE>** 0.3x4=**1.3p**

[C-8] Cs. Simon, S. Székely, K. Németh: Point-to-Multipoint ATM Signalling Performance Measurements, Proc. of the 7th IFIP Workshop on Performance Modelling and Evaluation of ATM/IP Networks, PMEAN'00 Ilkley, U.K., 16-19 July 2000, pp. W17/1-10, **<R>** 0.3x3=**1p**

[C-7] I. Szabó, S. Székely, I. Moldován: Performance Optimisation of AAL2 Signalling for Supporting Soft Handoffs in UMTS Terrestrial Radio Access Networks, Proc. of the 5th IEEE Symposium on Computers and Communications, ISCC'00, Juan-les-Pins, France, 4-6 July 2000, pp.46-52, *available online in the IEEE Digital Library, www.ieee.org* **<LREH1>** 0.3x4=**1.3p**

[C-6] S. Székely, G. Szűcs: Performance Measures of Call Establishment in ATM Networks, Proc. of the ATMTU'99 Int. Symposium, Kosice, Slovakia, 17-19 February 1999, pp.116-121, **<LE>** 0.5x3=**1.5p**

[C-5] S. Székely, Cs. Simon, G. Szűcs: Performance Testing on Switched Virtual Connections in ATM Networks, Proc. of the 6th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, PMEAN'98 Ilkley, U.K., 20-22 July 1998, pp. 102/1-10, **<RH1>** 0.3x3=**1p**

[C-4] S. Székely, I. Moldován, Cs. Simon: Overload Generated by Signalling Message Flows in ATM Networks, IFIP TC6 WG6.3 Conference "Performance of Information and Communications Systems", PICS'98, Chapman&Hall Publisher (eds. A. Nilsson, U. Körner), Lund, Sweden, 25-28 May 1998, pp. 51-64, *available at DBLP Uni-Trier* **, <LRH1>** 0.3x4=**1.3p**

[C-3] S. Székely: On Bandwidth Allocation Policies in ATM Network using Call Queueing, Proc. of the 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, PMEAN'97 Ilkley, U.K., 21-23 July 1997, pp. 46/1-10, **<R>** 1x3=**3p**

[C-2] S. Székely, G. Fodor, S. Blaabjerg: Call Queueing: The Design and Performance Analysis of a New ATM Signalling Functionality, Proc. of the IEEE Workshop, ConTEL B&MW, Zagreb, Croatia, 11-12 November 1996, pp.99-113, **<LEH1>** 0.3x3=**1p**

[C-1] A. Faragó, S. Blaabjerg, W. Holender, B. Stavenow, T. Henk, L. Ast, S. Székely: Enhancing ATM Network Performance by Optimizing the Virtual Network Configuration, International Conference on Data Communications and their Performance, Chapman&Hall Publisher (eds. R. Onvural, H. Perros), Istanbul, Turkey, 23-26 October 1995, pp. 401-414, *available at DBLP Uni-Trier*, **<LREH1>** 0.14x4=**0.5p**

[T] ***Technical papers***

[T-2] S. Székely: Schaltungsanordnung und Verfahren zur Fehleranalyse (*Eine neue Methode für System*

Tester zur Fehlersuche: Repräsentierung von Signalisierungsnachrichten in einem Plot-diagramm), Phase 1: Erfindungsmeldung – 2003 E18241 DE, Phase 2: 2004 P00559 DE, Germany, patent pending 1x2=**2p**

[T-1] S. Székely: VINCE: A Tool for Protocol Verification and for Interfacing Between Different ATM Signalling Protocols, Technical Report, T.U. Karlsruhe, Institute of Telematics, Germany, pp.1-37, March, 1996, (http://www.telematik.informatik.uni-karlsruhe.de/forschung/atm-info/vince/), *avail. at: Citeseer.com,* **<RES1>**

[H] ***Hungarian journal papers***

[H-3] G. Szűcs, S. Székely: Hívásszintű teljesítmény-mérések ATM hálózatokban, Magyar Távközlés, 99/5, május 1999, pp. 3-7, 0.5x1=**0.5p**

[H-2] S. Székely, A. Arató: Videókonferencia alkalmazások ATM hálózaton, Magyar Távközlés, 96/6, június 1996, pp. 6-10, **<H1>** 0.5x1=**0.5p**

[H-1] S. Székely, K. Szarkowicz: Bepillantás az ATM jelzésekbe, Magyar Távközlés, 95/5, május 1995, pp. 27-30, 0.5x1=**0.5p**

[R] ***Academic works which cited my publications***

[R-6] G. Fehér: Resource Control in IP Networks, Ph.D. dissertation, Budapest University of Technology and Economics, April 2004, *available on request*, cited [J-4]

[R-5] S. Waller: New AAL2 signalling protocol to support the UMTS Terrestrial Radio Access Network, M.Sc. thesis, University of Cape Town, 2003, http://crg.ee.uct.ac.za/progress.html, cited [C-7]

[R-4] A. Schwarz: Modellierung und Bewertung von Verfahren zur Last- und Leistungsregelung in Steuereinheiten von B-ISDN/ATM Vermittlungssystemen, Ph.D. dissertation, University of Stuttgart, LDB #33612, 2000, *available on request*, cited [C-4] and [C-5]

[R-3] I. Cselényi, R. Szabó, I. Szabó, A.L.-Henner, N. Björkman: Experimental Platform for Telecommunication Resource Management, Computer Communications, Vol. 21, pp.1624-1640, 1998, http://qosip.ttt.bme.hu/papers/local/cselenyi98compcomm.pdf, cited [C-2]

[R-2] B. Kaan: A Virtual Path Routing Algorithm for ATM Networks based on the Equivalent Bandwidth Concept, M.Sc. thesis, Bogaziti University, 1998, http://mercan.cmpe.boun.edu.tr/~kaanbur/thesis.pdf, cited [C-1]

[R-1] G. Seres, F. Baumann, G. Gordos, T.Henk: ATM alkalmazása a lokális hálózatban. A BME-TTT tanszék helyi ATM hálózata, Networkshop'97, Keszthely, Hungary, May 27-29, 1997, http://www.iif.hu/rendezvenyek/networkshop/97/tartalom/NWS/1/7/index.htm, cited [J-2] and [H-2]

[D-n] ***Chapters of my dissertation*** n = 1 to 10

Min. required points: 12p **Total points: 25p**

8. Acknowledgements

My special thanks go:

- to T. Henk, head of the High Speed Networks Laboratory, for his continuous support since 1995;
- to my supervisors, S. Molnár and B. Frajka for their helpful suggestions and critical comments;
- to H. Perros, A. Nilsson, J. Kühn, J. Eberspächer, G. Krüger, G. Carle, J. Sztrik and to the anonymous reviewers for their precious comments and reviews of some of my publications;
- to my colleague, I. Moldován for implementing the simulator ACCEPT (used in Theses 2, 3, 4, 5);
- to my colleagues, Cs. Simon, G. Szűcs, K. Németh, F. Fábián and their students for the invaluable assistance while preparing the measurements, running the tests and processing the results (Thesis 1);
- to my colleagues, Cs. Antal, I. Cselényi, K. Szarkowicz, I. Szabó and G. Fodor for the long discussions about the design solutions in the UMTS networks and queueing in broadband networks, respectively (Theses 4, 5);
- to the Hungarian Ministry of Education for a four years scholarship;
- last but not least to G. Gordos, former head of the department of Telecommunications and Telematics at BUTE, and to M. Boda, former head of the R&D department at Ericsson Hungary for their continuous support during my Ph.D. studies.