

ON THE IMPACT OF LINK/NODE FAILURES AND NETWORK APPLICATIONS ON THE LOAD AND CALL PROCESSING TIMES IN ATM NETWORKS

Sándor SZÉKELY¹, Gábor SZŰCS¹, Csaba SIMON, István MOLDOVÁN and Sándor MOLNÁR

High Speed Networks Laboratory
Department of Telecommunications and Telematics
Budapest University of Technology and Economics
1117 Budapest, Magyar tudósok krt. 2, I.B.210, Hungary
Tel: +36-1-463-31-10, fax: +36-1-463-31-07
e-mail: {szekely,szucs,simon,moldovan,molnar}@ttt-atm.ttt.bme.hu

Received: December 10, 2002

Abstract

The beginning of the 1990s 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 equipment is largely introduced by users and service providers. At the end of the 1990s 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 operation costs. ATM network service providers offered mainly permanent virtual circuit connections to customers in the last years, but recently there is an increasing interest to offer switched virtual circuit (SVC) connections to end users. The SVC is based on the use of signalling protocols. Our paper focuses on the performance of call processing in ATM networks. Based on a series of measurements on four types of ATM switches we have established some generalised conclusions, which are switch invariant features of the ATM signalling flows. These results have a wide generality among ATM switches and they can be used to model and design large ATM signalling networks. Our new model is validated against the measurement results on the 7-node backbone of the TEN-155 Pan-European ATM network. The second part of the paper shows that a cascaded network is a good estimator for the signalling performance of an arbitrary network. In addition, it is demonstrated that the call density of the network is an important network parameter, which is closely related to the maximum network level call arrival rate. The impact of the link/node failures and network applications on the call establishment times are investigated in this article on a 35-node sample network. The case studies are extended from homogeneous networks to hybrid networks, where many types of switches are present.

Keywords: ATM signalling, performance measurements, modelling, network diameter, average call path, call density of the network, maximum network level call arrival rate, link and node failures.

¹The authors are currently with Siemens AG, ICN, Munich, Germany

1. Introduction

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 a network can accept and the rate at which they can be accepted. Both of these performance criteria are influenced by the User-Network-Interface (UNI) signalling performance [1]. A good survey of signalling protocols for ATM networks can be found in [24], [18] and [3].

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. [2], 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 [1], while performance benchmarking of ATM signalling software is presented in [11] and [16], showing the first results obtained on switches from four vendors in a number of hardware, software and network configurations. [16] 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 [6], very few papers addressed the congestion situation in switches due to signalling message flows. Furthermore, the evaluation in [12] is based on estimations and external inputs and expresses the necessity to have real measurements of the equipment that is going to be installed for Broadband-Video-on-Demand (B-VoD).

Based on these and many other studies we have concluded that the existing measurement results of current broadband signalling networks are far not enough to cover all aspects of the ATM signalling. Motivated by this need we addressed this research area and therefore the first part of our paper focuses on real measurements of such signalling networks.

Section 2 describes our objectives, the methodology we have used during our evaluation and establishes our definitions used in this article. Section 3 focuses on the performance of the call processing by measurements of isolated switches. The aim is to identify the main components of call processing in ATM networks and to describe quantitatively the effect of different call profiles. Section 4 deals with the construction and performance analysis of a new generic flow model based on these measurement results from Section 3. The analysis is extended to network level in section 5, different network topologies are investigated: 10-node cascaded, 4-node fully meshed, 7-node, 30-node and 35-node arbitrary networks. Hybrid networks vs. homogeneous networks are also studied. Finally, section 6 concludes the paper.

2. Objectives, Methodology and New Definitions

The objective of our paper is to analyse the performance of the ATM signalling networks and to give a new model that describes the behaviour of complex call processing. More exactly, our main goals are the following:

- to identify the main components of call processing delays by a series of measurements carried out on isolated switches manufactured by different vendors;
- to develop and analyse new models based on the above measurements that accurately describe the message latencies in an ATM signalling node and to use this model for large ATM network design.

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 [21]. Moreover, the evolution of the xDSL technologies, IN networks, IP-over-ATM and WWW applications argue the need of real performance measurements in ATM signalling networks [12], [14].

Tests to analyse the properties of a single switch are important and clearly relevant to LAN performance. It is also obvious that analysis of 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, we have turned to *simulations*². 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 the links, call throughput and signalling load of nodes. Our investigations here include an analysis of a generic message flow model by simulation. The results are validated by *measurements*.

Some *definitions* are given here in order to help understanding the subsequent statements.

Definition 1 The *call establishment time* is the amount of time needed for an ATM system to establish a switched virtual connection between network components.

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

where $t_S(\cdot)$ is the timestamp of an outgoing or incoming message at the source; $X = \text{SETUP}$; $Y = \text{CONNECT}$ are signalling messages defined in [20], [26].

This is the most fundamental signalling performance metric.

²The simulation study was carried out based on the simulation software called ACCEPT, developed by the authors at the HSN Laboratory, Dept. of Telecomm.&Telematics, BUTE, Hungary.

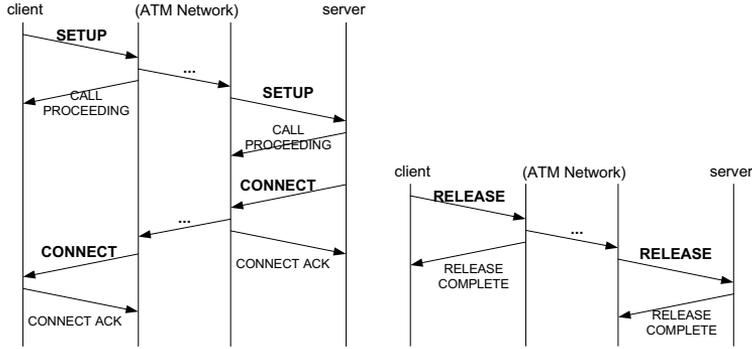


Fig. 1. Signalling message flows in ATM networks a) successful call setup; b) call release

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

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

where $X = REL$; $Y = RLC$ are signalling messages described in [20], [26].

The call release time *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, we have introduced new performance measures to properly characterise the signalling behaviour in the network side, as follows:

Definition 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.

Definition 4 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:

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

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

Definition 5 The *call throughput* (ρ_R) of a switch (network) is the number of successful calls/number of generated calls.

3. Measurements of Call Processing in ATM Switches

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, we have decided to carry out *performance measurements* for our analyses. Let us summarise a short list of the currently available ATM switches on the telecommunication market with signalling capabilities: Alcatel 7670, Alcatel (formerly Newbridge) MSX36170, Cisco BPX8620, Cisco (formerly LightStream) LS2020, Ericsson AXD301, Lucent (formerly Ascend) CBX500, Marconi (formerly Fore) ASX200/1000, Nortel Passport7400, Siemens (formerly Seabridge) XpressPass140, etc. Out of these, four ATM switches were used in our measurements: *GDC APEX DV2* [5], *Fore ASX200BX* [4], *Newbridge MSX36170* [15] and *Seabridge XP140* [27]. This is a good representative sample covering a range of products from 1995 to 2001.

We have analysed the signalling measurements obtained on the four aforementioned commercial ATM switches and then compared to some similar results obtained on five other switches by [11], [16] and [17]. In section 1 we 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*.

We have carried out a set of signalling performance measurements on *point-to-point single* and *point-to-point multi-connection* ATM calls. Our *point-to-multi-point* measurement results are not general enough yet to be included into this paper. The main interest was in 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 ATM signalling node. In this simple case we had two user-network interfaces (UNI) with STM-1 (155 Mbps) optical interfaces.

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

- we assured only successful point-to-point calls (no errors occurred 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 BSTS 75000* [9] and *GNNetest iW95000* [8]).

We have carried out both steady state and transition state signalling measurements, the detailed results were already presented in [23], therefore in section 3.1 we will only give a short summary of these results, for better understanding of the construction of the signalling model in section 4.

3.1. Signalling Performance Evaluation of Point-to-Point Connections in ATM Switches

In this section we have proven that the standard performance measures (e.g. call establishment time, release time) defined by the standardisation institutes (e.g. [1]) are not enough to properly characterise the performance of a broadband signalling network. We have shown that the effect 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. The results are grouped into 7 independent statements, as presented in *Table 1*.

Table 1. The overview of the results

Statement 1	The dominance of layer 3 message processing times
Statement 2	The differences between message processing delays
Statement 3	The dependency of T_C on routing table, bandwidth allocation and number of active connections
Statement 4	The dependency of message processing times on different call profiles
Statement 5	The impact of the call release phase on the T_C
Statement 6	The impact of the signalling overload on the T_C and T_{RN}
Statement 7	Estimation of the T_{CN} of cascaded switches

Statement 1 Based on measurements and analysis we have concluded that the dependency of call establishment time on layer 3 processing is at least one order of magnitude higher than its dependency on lower layer processing.

The layer 3 processing is a software implementation, while layer 2 and layer 1 processings are implemented in hardware. As a conclusion, in the following we have investigated only the effect of layer 3 processing on the signalling performance of ATM switches.

Statement 2 We have shown that the same destination response time (T_{DS} , resp. T_{DR}) can be either dominant or non-dominant component of the call establishment time and release latency, depending on the speed of the signalling processor of the switch. Moreover, we 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 (30...36)% of the *SETUP*³ delay, respectively (see *Fig. 2*).

³This *SETUP* represents a *SETUP* message containing the mandatory information elements only.

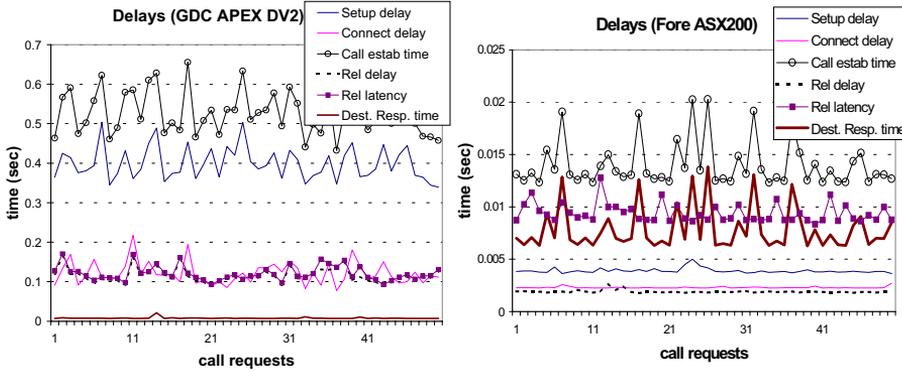


Fig. 2. Message delays, call establishment times and release latencies at 1 call/sec load case a) GDC APEX DV2 switch; case b) FORE ASX200BX switch

As an example while the destination response time on call setup ($T_{D,S}$) is less than 3% of the *SETUP* delay in the case of a GDC switch, the same response time is higher than the *SETUP* delay in the case of a FORE switch.

Statement 3 We have shown that the call establishment time depends linearly on the size of the routing table, but it is independent of the size of the allocated bandwidth. Moreover, we have proved 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 [16], e.g. the dependency on the PNNI hierarchy [19] is shown there as well. Note that, when running the tests with different bandwidth sizes, we paid attention to avoid call rejections due to overloaded links.

Statement 4 We have given a quantitative measure, how the mean delays of *SETUP*, *CONNECT* and *RELEASE* messages (and therefore the call establishment time) depend on different call profiles, when processed at an ATM signalling node:

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

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 great variety from simple voice calls 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:

$$T_C^{CCP} = (1 + s) \cdot T_C^{\text{default}}, \text{ where } 0 < s < 1. \quad (5)$$

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 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

- We 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:

$$T_{RN}^{CCP} = T_{RN}^{\text{default}} \quad \forall s, \quad 0 < s < 1. \quad (6)$$

- We have shown that the mean call establishment time does not depend on the type of call.

$$\text{e.g. } \bar{T}_C |_{VBR} \approx \bar{T}_C |_{CBR} \approx \bar{T}_C |_{ABR} \approx \bar{T}_C |_{UBR}, \quad (7)$$

where \bar{T} denotes the average time (see Table 1, the case when adding AAL1, AAL3 or AAL5 parameter IEs). Early papers in the literature expected that the call establishment time for VBR calls would be larger than for CBR calls (see [7], [25]), but our measured results contradicted to this statement.

Statement 5 We have further investigated the mean call establishment latency of simple calls (i.e. default *SETUP*) in two cases: call establishment followed by call release vs. call establishment without release phase and we have found the following relationships:

- In the 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 Fig. 3).
- In the 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.

- 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 [11] 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 in Fig. 3, our results contradict to this statement.

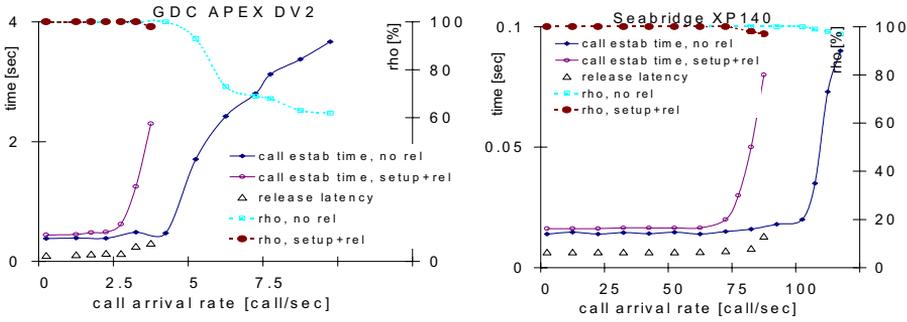


Fig. 3. Call establishment time, release latency and throughput vs. call intensity case a) GDC APEX DV2; case b) Seabridge XP140

Statement 6 In the case of a light call overload in the switch (<10% of the calls are rejected), the call establishment time increases dramatically, but the slope of the call release latency remains for this range still constant (see Fig. 3). Further increasing the call intensity leads to a dramatic increase of the release latency as well.

This may happen when there is a priority mechanism applied or when the *SETUP* and *RELEASE* messages visit partly different paths inside the signalling processor (distributed call processing). The rejected calls are caused by buffer overflow in the processor.

Statement 7 We 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:

$$r \cdot \overline{T}_{CN} |_{1\text{switch}} \geq \overline{T}_{CN} |_{r'\text{ switches}}, \quad r = 1, 2, \dots, k. \quad (8)$$

This formula can be generalised to a heterogeneous cascaded network, as follows:

$$\sum_{i=1}^P n_i \cdot \overline{T}_{CN}^{\text{type } i} \geq \overline{T}_{CN} |_{r'\text{ switches}}, \quad r = \sum_{i=1}^P n_i, \quad (9)$$

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 [13] on 3 different types of cascaded ATM switches confirm the second inequality.

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, we have shown that these features of the ATM signalling are general regardless of different architecture and processing power of ATM switches. Of course, there are features which depend on software implementations and hardware architectures (e.g. see [16]), but these are not subject of our investigations and therefore not shown here (except Statement 2).

The statements demonstrated in section 3 represent the results of a very ‘unpopular’ research, the measurements were collected during the last four years by our research team on four different generations of switches manufactured by different vendors. Moreover, the results were obtained with two different test equipment. In addition, independently of our work other research groups (see [16], [17], [13]) obtained very similar results with other types of ATM switches, thus confirming that our 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. The results were given within a 95% confidence interval, i.e.:

$$\left[m_i - 2 \cdot \frac{1}{(n-1) \cdot \sqrt{n}} \cdot \sum_{i=1}^n (x_i - m_i)^2, m_i + 2 \cdot \frac{1}{(n-1) \cdot \sqrt{n}} \cdot \sum_{i=1}^n (x_i - m_i)^2 \right]$$

where

$$m_i = \frac{1}{n} \cdot \sum_{i=1}^n x_i. \quad (10)$$

4. Construction of a Generic Call Processing Model for PNNI Networks

The construction of a call model for Private Network Node Interface (PNNI) [19] has been motivated by the fact that some service providers want to introduce dynamic call establishment in their access Digital Subscriber Line (xDSL) and backbone ATM network. 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 (with Integrated Access Devices, IADs) each having 0.66 Mbps CBR channel for its voice traffic (8×64 kbps channels) and 0.1 Mbps nrt-VBR channel for its inband management. The data channel (UBR) is 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.

Today it 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, extracting the call setup time information from a stream of ‘ping’ messages [17]. This backbone network consists of seven ATM signalling nodes, the longest path (diameter) contains four nodes. Our experience gained in a relatively small network has been validated against these results on the *TEN-155*, and extended by simulation to even larger networks. Furthermore, as in section 3 we have shown that in many cases the test equipment may be the bottleneck (see Statement 2), therefore we have developed an adequate model for the end systems as well. By multiplying the number of end systems connected to one switch we can avoid the exponential increase of the destination response time vs. call arrival rate, thus diminishing the number of error sources in our estimations.

4.1. The Architecture of the Model

We have developed a new call processing model for an ATM signalling node based on the specifications [20], [26], [19] and the measurement results presented in section 3. The above mentioned documents do not specify any call model, but describe the format of the signalling messages and the protocol how they interact. From the results presented in statements 2, 4 and 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 we have investigated two different mechanisms in the proposed model: *FIFO queueing* and *priority queueing*. Moreover, at one node we have distinguished *separate processor phases* according to the jobs to be done (see Fig.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 statement 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 following: 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 [16]).

Two other call processing models have been developed in [25] and [6]. While

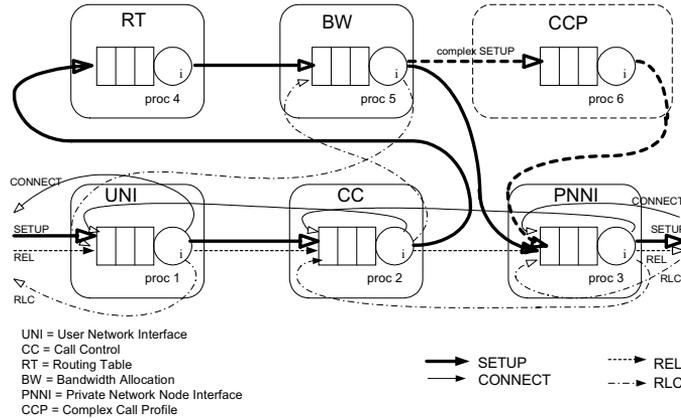


Fig. 4. Call processing model of an ATM signalling point

the model in [25] is even more complex than ours and considers different service times for CBR, VBR and UBR calls, respectively (which is not needed according to our measurements, see statement 4), none of the two other models considered the release phase. However, as described in statement 5, the T_C is increased by 15–20% when release messages are also present in the network. Furthermore, the model in [6] consists of a single FIFO queuing model, which cannot capture the differences between the setup and release latencies.

4.2. 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 Figs. 2 and 3);
- 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 ($RDLM$ ⁴) have to be determined.

We have shown that the service time of each process can be derived from the following system of equations:

⁴ $RDLM$ is the delay a lost message (other than $SETUP$) is regenerated at the switch in case of buffer overflow.

$$\begin{aligned}
\sum_{i=1}^5 \frac{1}{\mu_i} &= \min \text{ SETUP delay} & \sum_{i=1}^6 \frac{1}{\mu_i} &= \min \text{ 'complex' SETUP delay} \\
\sum_{i=1}^3 \frac{1}{\mu_i} &= \min \text{ CONNECT delay}^5 & \sum_{i=1,2,3,5} \frac{1}{\mu_i} &= \min \text{ RLC delay}^6
\end{aligned} \tag{11}$$

$1/\mu_1 = a/\mu_2 = 1/\mu_3$, where $a \in (1, 2)$ is a correction factor.

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 Eq. (11), the call throughput of the switch, the call establishment times and release latencies obtained by measurement at different call arrival rates.

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

- STEP_1.** Set the service rates according to the system of Eqs. (11). 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 $\rho_R = 1$.
- STEP_2.** Set the length of the bottleneck buffer BS_i , $i = \{j \mid q_j > q_k, \forall j, k \in \{1, \dots, 6\}\}$ so that the threshold values $|\lambda_{Th,sim} - \lambda_{Th,meas}| < 0.05 \cdot \lambda_{Th,meas}$, where $\lambda_{Th} = \min\{\lambda \mid d(T_{CN})|_{\lambda=\lambda_{Th}} >> 1\}$ and q_j is the message length of queue j .
- STEP_3.** Set the length of the remaining buffers until the throughput of the switch $|\rho_{R,sim} - \rho_{R,meas}| < 0.05 \cdot \rho_{R,meas} \forall \lambda > \lambda_{Th}$.
- STEP_4.** Set $RDLM$ to further minimise the error between the three measured and simulated curves in the overload region ($\lambda > \lambda_{Th}$).

The simulated results of all three parameters (T_{CN} , T_{RN} , ρ_R) are adjusted to the measured ones for all switches we have studied (e.g., see Fig. 5). The output parameters (BS , $RDLM$) are determined such that the errors between the simulated and measured results are less than 5%.

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 both constant rates and Poisson arrival rates, and repeated each run 5 times with different random variables. The average values were obtained from 10000 calls \times 5 runs and they were related to a 99% confidence level.

⁵or min RELEASE delay

⁶The Release Complete (RLC) message has end-to-end meaning in our model.

4.3. Validation of the Model

To validate our model in a real environment, we have selected a 7-node ATM network. The network configuration can be seen in Fig. 6 and represents the TEN-155 network offering service to the European research community [17]. The 8th node from Switzerland (CH) has not been used for signalling.

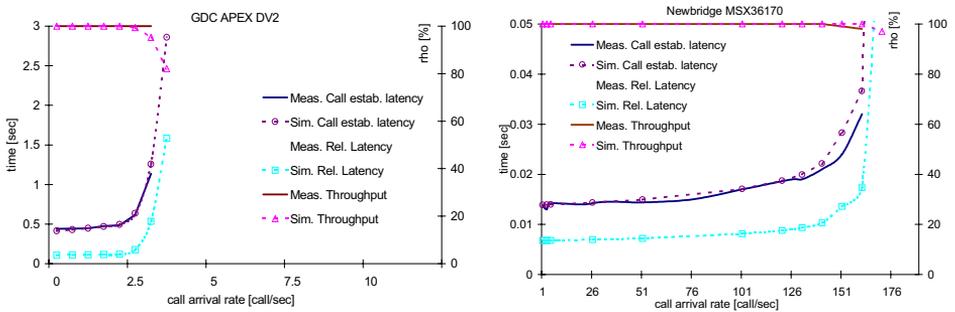


Fig. 5. Calibration of the model for 'default' *SETUP* messages: a) GDC APEX DV2 switch; b) Newbridge MSX36170 switch

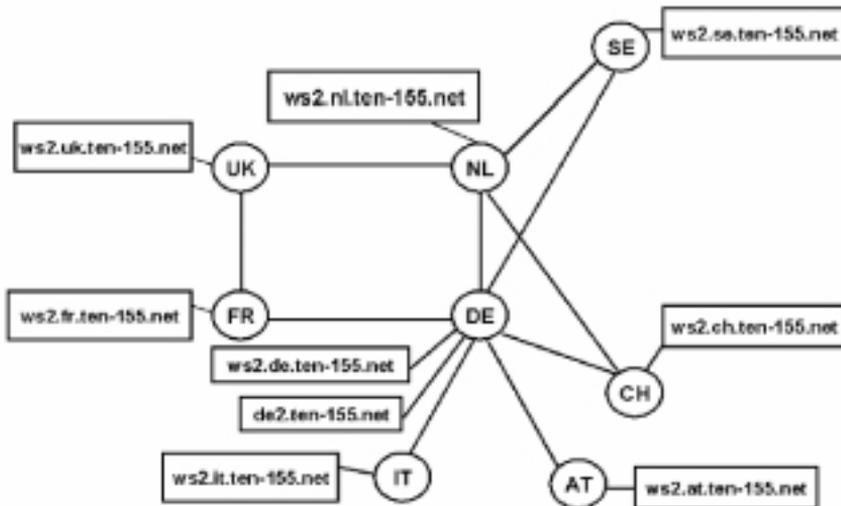


Fig. 6. The backbone nodes of the TEN-155 network

The backbone nodes are all Lucent (formerly Ascend) CBX500 (with a signalling service time of 25 msec, i.e. 40 calls/sec) except the nodes in Sweden (SE)

Table 3. Validation of our model against network-level measurements in the TEN-155 network

Simulated results → Min/Avg/Max (StDev)	DE	FR	NL	UK	AT	SE	IT
Measured results ↓ Avg	[msec]	[msec]	[msec]	[msec]	[msec]	[msec]	[msec]
DE	–	45/45.1/60.7 (0.62)	45/45.1/60.7 (0.62)	75/75.2/90 (1.2)	55/55.2/68.7 (1.05)	90/90.6/125.8 (2.47)	80/80.4/104.6 (1.57)
FR	44..45	–	65/65.2/73 (0.78)	55/55.2/69 (1.08)	75/75.3/91 (1.19)	110/110.6/132.6 (2.49)	100/100.4/127 (1.62)
NL	40..45	65..66	–	55/55.2/69 (1.08)	75/75.3/91 (1.19)	90/90.6/125.8 (2.47)	100/100.4/127 (1.62)
UK	79..80	54..66	41..52	–	105/105.4/118 (1.36)	120/120.8/143.7 (2.9)	130/130.5/152.8 (1.73)
AT	53..54	80..85	75..76	115..118	–	120/120.7/149.7 (2.7)	110/110.5/133.5 (1.72)
SE	90..98	124..125	90..92	109..112	132..133	–	145/145.9/178 (3.07)
IT	80..83	90..107	87..96	124..141	110..112	143..160	–

and Italy (IT), where a lower speed CBX550 (50 msec) is placed. The workstations at the sites are all SUN workstations with different processor capacities using SUN ATM CLIP software to generate ATM signalling. The SVC network paths consist of 2, 3 or 4 hops, however, 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-pair in [17]. Each measurement result was an average of 5–6 measurements, each consisting of a stream of ‘ping’ messages. For our simulated results we have generated 3000 calls at 1 call/sec. The results are compared in *Table 3*. The measured results are placed in the lower triangle, while our simulated results in the upper one. For the measured results we have only the average values (for both directions) and the standard deviation of 4.4, while for the simulated results we have obtained the Min/Avg/Max values and the standard deviations.

We have observed that in some cases the measured average values in the two opposite directions are very different (see e.g., IT-FR, IT-UK, IT-SE, UK-NL). This is due to the fact that the call establishment time was measured as the difference between the round-trip-time (RTT) of the first and second ‘ping’ packet when there is no SVC built up a-priori, but the ARP cache is still populated (IP to ATM mapping is kept for max. 600 sec. Otherwise, when the ARP cache is empty, the address resolution time will be added to the call establishment time). We have to admit that these types of measurements cannot be so accurate as a calibrated special test equipment used in our measurements in section 3. However, in more than 80% of the cases, the simulated results are very close to the measured ones. There are a few deviations, due to the inaccurate ‘ping-type’ measurement results. For example, between nodes UK-AT and FR-SE the measured values are 10% over the simulated average values; between nodes NL-IT, NL-UK and UK-SE the measured values are 5–10% under the simulated ones. No greater deviation was observed between the simulation and measured results.

The model defined in section 4 is quite complicated, it is not easily tractable analytically, especially when using priority mechanism or when studying a large network, therefore in the following we have used our *simulation model* to estimate the call establishment times and release latencies in real size networks.

5. Performance Analysis Based on the Proposed Model

The message latencies introduced by the signalling nodes in an ATM network are additive, as shown in statement 7. This property of the measurement implies that with growing number of nodes 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 [10]. We have defined the network diameter as the shortest path between the two furthest nodes

of the network: $\Phi = \max\{L(i, j) \mid L(i, j) < L'(i, j), \forall i, j \in \{1, \dots, N\}\}$, where $L(i, j)$ is the path length between nodes i and j .

We have studied the call establishment time 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, we have estimated these parameters for $N = 4$ -node fully meshed, $N = 30$ -node and $N = 35$ -node ATM networks in a typical xDSL topology with 20–26 DSLAMs (each having 200 subscribers). These sample networks contain all source-destination path lengths from 2 to 8 nodes, so we could compare the call establishment times and release latencies (for all path lengths) to those obtained for cascaded switches.

Statement 8 We have shown that an r -node cascaded network can be used to estimate the signalling parameters (T_{CN} , T_{RN} , ρ_R) of a large, N -node ($N > r$) network having a diameter $\Phi = r$ nodes (see Fig. 7).



Fig. 7. Ratio of the call establishment times (and release latencies) in cascaded vs. 35-node network with $\Phi = 8$ (Fore ASX200BX)

First, we have investigated a *homogeneous network* (all network nodes are identical). We have shown that the minimum values of T_{CN}^L and T_{RN}^L for each path-length ($L = 1, \dots, 8$) 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 networks, 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 λ_N^{\max} is the maximum network-level call arrival rate without any call rejection, N is the number of nodes. There is one exception to be seen in Fig. 7 for 1–2 hops, where the ratio for the maximum values is 30–50%. This is due to the fact that we have a small number of calls with 1–2 hops in the 35-node network,

approx. 2% of the maximum load (see Fig. 8), while the cascaded network of 1–2 nodes is tested under the maximum load (33, 100, calls/sec resp.).

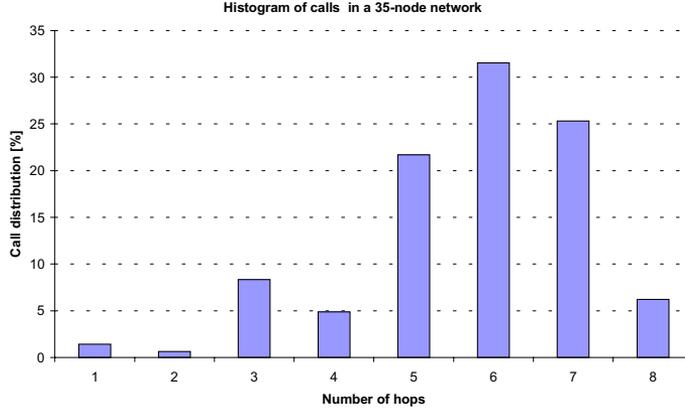


Fig. 8. Histogram of the call path length in the 35-node network

In Fig. 8 we have shown the distribution of the calls in this 35-node network. It can be seen, e.g., that 30% of the calls go through 6 network nodes, 25% travel along 7 nodes, while 20% visit 5 hops.

Statement 9 We have shown that the maximum network-level call arrival rate λ_N^{\max} depends on the network topology as follows:

$$\begin{aligned} \lambda_1^{\max} &\approx \lambda_r^{\max} \leq \lambda_N^{\max} \text{ or } \max\{\lambda \mid r \geq 1, \text{ cascade}, \rho_R = 1\} \\ &\leq \max\{\lambda \mid \Phi(N\text{-node}) = r > 1, \rho_R = 1\}. \end{aligned} \quad (12)$$

The term λ_1^{\max} is the maximum call arrival rate measured in one isolated switch without any call rejection. The average call path is defined by the following equation: $\bar{L} = \sum_{r=1}^{\Phi} p_r \cdot r$, where p_r is the probability that a call goes through r nodes. Furthermore, we have found that λ_N^{\max} can be approximated analytically by the following formula:

$$\lambda_N^{\max} \approx \lambda_1^{\max} + \left| \frac{N_{ISP} - 1}{N_{ISP} + 1} \right| \cdot \frac{\lambda_1^{\max}}{\bar{D}_N} = \lambda_1^{\max} \cdot \left[1 + \left| \frac{N_{ISP} - 1}{N_{ISP} + 1} \right| \cdot \frac{\Phi}{(1 - \frac{N_{tr}}{N}) \cdot \bar{L}} \right], \quad (13)$$

where N_{tr} is the number of transient nodes in the network, N_{ISP} is the number of internet access servers. We have defined the average call density (\bar{D}_N) of an N -node network: $\bar{D}_N = (1 - \frac{N_{tr}}{N}) \cdot \frac{\bar{L}}{\Phi}$. The call density is an important network-level parameter, because it is indicating the distribution of the end users among the network nodes. In general, $0 < \bar{D}_N \leq 1$. In our sample networks ($N = 4, 30, 35$) this parameter is $\bar{D}_4 = 0.93$, $\bar{D}_{30} = 0.59$, $\bar{D}_{35} = 0.45$, while for an isolated switch

$D_1=1$. The applicability of this formula is dependent on how easy the average call path \bar{L} can be found. We have obtained \bar{L} by simulation. Furthermore, we found the lower and upper bounds for the λ_N^{\max} ,

$$\lambda_1^{\max} \leq \lambda_N^{\max} \leq \lambda_1^{\max} \cdot \left(1 + \frac{N \cdot \Phi}{\bar{L}}\right), \quad (14)$$

but for realistic network scenarios these margins are usually tighter, e.g. $\lambda_1^{\max} \leq \lambda_N^{\max} \leq 3 \cdot \lambda_1^{\max}$ (see *Table 4*).

Table 4. Comparison of simulated and estimated values of the λ_N^{\max}

If the max. call arrival rate for 1 switch = 140 call/sec				Max. network level call arrival rate	
# of nodes	$L(i, j)$	# of transient nodes	# of ISPs (0=LANE)	Simulated [call/sec]	Estimated [call/sec] with Eq. (13)
4	1.86	0	0	280	290
7	2.71	0	0	295	347
30	5.28	10	0	385	379
35	5.72	13	0	355	451
35	4.17	12	1	100	140
35	4.67	12	2	245	231
35	4.49	11	3	335	276

Secondly, we have studied the behaviour of the call processing in a hybrid network (not all network nodes are identical). In practical situations we have often found two or three types of ATM switches in one network. In this case, the formula given in *Eq. (13)* will suffer certain changes:

$$\lambda_N^{\max} \approx \max \left[\min_{i=1..P} (\lambda_1^{\max})^{\text{type}i}, \min_{j=1..R} (\lambda_1^{\max})^{\text{type}j} \right] + \left| \frac{N_{ISP} - 1}{N_{ISP} + 1} \right| \cdot \frac{\min_{i=1..P} (\lambda_1^{\max})^{\text{type}i}}{\bar{D}_N}, \quad (15)$$

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

Next, we 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 $0 < s < 1$.

Statement 10 We have shown that, if $T_C^{CCP} = (1 + s) \cdot T_C^{\text{default}}$ for one node, then the average call establishment time of the network will be increased to $\bar{T}_C^{CCP} \leq (1 + s) \cdot \bar{T}_C^{\text{default}}$, while the average call release latency \bar{T}_{RN} remains unchanged.

The equality stands for the case when all calls in the network are of the same complexity. This statement 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.

Statement 11 We have shown that replacing one node of a network $((\mu_i^C; \mu_i^R)$, $i = 1, \dots, N$) by a switch having a signalling service rate $(\mu_0^C \leq \frac{1}{\Phi+1} \cdot \min_{i=1..N} \mu_i^C; \mu_0^R \leq \frac{1}{\Phi+1} \cdot \min_{i=1..N} \mu_i^R)$, can be used to determine the signalling load of this node. Replacing each node one-by-one will identify the bottleneck nodes in a given network topology.

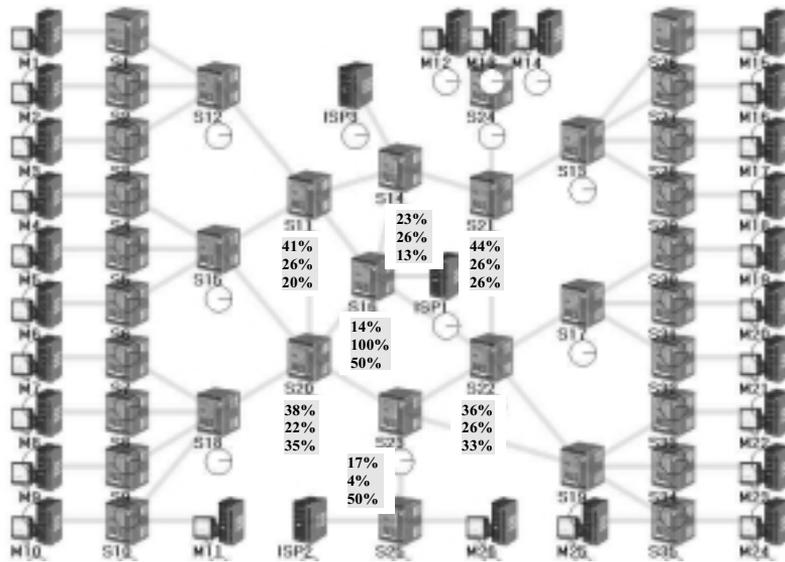


Fig. 9. Measuring the relative signalling load (p_i) in the core of the sample 35-node network a) LANE service (1st value; b) Internet access via one ISP (2nd); c) Internet access via 2 ISPs (3rd)

If it is not possible to monitor and analyse each network node individually, then the method described above is very useful. All calls with $T_C > \frac{1}{\mu_0^C}$ are processed by this test node as well. Thus, the signalling load of this network node can be defined as follows: $\lambda_i = \lambda_N \cdot p_i$, $i = 1, \dots, N$, where p_i is the relative load of node 'i'. The bottleneck node in the network has its $\lambda_i \approx \lambda_i^{\max}$ (e.g., see Fig. 9, switch S16 and S23).

We have supposed that the call arrival rate is Poissonian and the source-destination pairs are uniformly distributed in the network. Simulations have been carried out for the following network loads: 33; 100; 200; 400; 500 calls/sec (e.g. see *Fig. 10*). Each time 10000 calls have been evaluated. The zero setup times represent the lost calls in the network. An arrival rate of 500 calls/sec resulted in 20 lost calls out of 10000 (0.2%). Simulation results have been validated against measurements for 2, 3 and 4 cascaded switches (e.g., with Fore ASX200BX and Seabridge XP140).

As a next step, we have investigated the impact of link and node failures in the core of the 35-node sample ATM network (in the case of pure LANE services). In general, one node failure in the backbone resulted in 3–5 adjacent link failures.

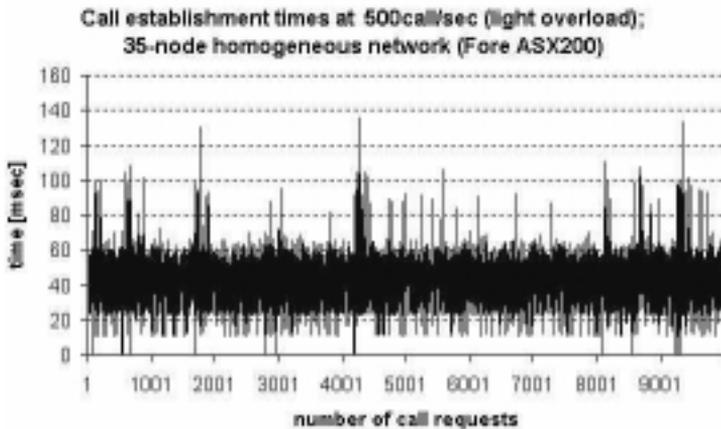


Fig. 10. Call establishment times in the 35-node homogeneous network (0.2% of calls are dropped)

Statement 12 We have shown that the failure of one link/node in the core of the 35-node sample network has the following impacts:

- it will result in a call blocking probability of 20–45%, if there is no alternative path (see *Table 5*);
- it will increase the load of the neighbour switches by 10–25%;
- it will increase the average load of the backbone by 5–10%;
- and it will change the average \bar{T}_C and \bar{T}_{RN} by maximum $\pm 5\%$.

This statement shows that the lack of redundancy is especially critical for switches carrying a load over 30% of the total network load. The load of some neighbour switches might increase up to 60% of the total load of the network. The 4-to-8 hops calls are primarily affected by a failure in the core network (LANE).

Table 5. The impact of node failure on the signalling load of the neighbour switches

Switch #	No failure, initial load/sw	Failure of sw # 11 [%]	Failure of sw # 14 [%]	Failure of sw # 16 [%]	Failure of sw # 20 [%]	Failure of sw # 21 [%]	Failure of sw # 22 [%]	Failure of sw # 23 [%]
11	40.9	–	30.94	39.17	45.47	33.05	59.74	48.59
14	23.22	22.76	–	26.94	19.02	0	44.16	26.01
16	14	32.17	37.93	–	22.42	25.23	0	25.58
20	37.95	53.49	38.41	45.03	–	49.33	61.42	33.9
21	44.71	49.38	45.01	47.89	50.24	–	48.8	46.26
22	36.17	37.11	59.64	32.1	54.9	33.39	–	45.85
23	17.7	16.25	17.3	28.29	8.95	27.41	36.21	–
lost calls [%]	0	21.27	0	0	32.86	45.02	21.49	7.51
avg load/sw [%]	30.66	35.19	38.21	36.57	33.50	28.07	41.72	37.70
avg T_C [msec]	38.73	38.42	39.62	39.04	37.27	36.54	40.47	39.77
avg T_{RN} [msec]	12.29	12.19	12.56	12.38	11.84	11.62	12.82	12.60

Furthermore, we have tested the impact of different network applications on the signalling load of switches in the backbone. Internet connectivity via one, two or three access servers has been studied vs. LAN Emulation service. The mixed scenario has been also investigated. We have found that the distribution of the network load λ_N depends on the number of ISPs (see Fig. 9).

Statement 13 We have shown that the signalling load of switches in the same network topology is strongly affected by the types of network applications (Fig. 9):

- only one internet access results in a load $\lambda_i \approx \lambda_N$ of the switch connected to this server;
- two internet servers (connected to different nodes) divide the load λ_W usually in two equal parts;
- the signalling load is more uniformly distributed among the backbone nodes in the case of LANE services ($\lambda_i \approx (0.3 - 0.4) \cdot \lambda_N$).

The investigation of PNNI hierarchical levels for the given (and some other) network topologies is in progress, but the results are not general enough yet to be included into this paper.

6. Conclusions

In this paper we have presented a performance evaluation study of ATM signalling networks. Our analysis is based on a series of comprehensive measurements using four commercial ATM switches. Moreover, our experimental results are validated against published results of four other switches. The analyses mainly concern delay components, call establishment time, release latency and call throughput. We have demonstrated the impacts of the characteristics of the call process and network components on these performance measures. It was shown that the ATM specific characteristics of the signalling procedure can significantly influence the call establishment time and the throughput. Based on our signalling procedure characterisation, we have developed a novel model for call processing of a signalling node. We have illustrated that our modelling approach can be used in the signalling performance evaluation of a real-size homogeneous/hybrid ATM network. Our model has been validated against call setup measurements in the TEN-155 network. We have compared the performance of a cascaded network vs. an arbitrary network, and have found an approximation formula for the maximum network level call arrival rate without any rejected call for many network scenarios. Moreover, we have investigated the effect of the node and link failures and the impact of different network applications (LAN Emulation, Internet Access) on the signalling load of the core switches. This helps us to find the margins of the maximum signalling load without any call rejection in the network, when a node/link fails or different applications are introduced. In the near future we plan to define 2-to-4 PNNI hierarchical levels (including the crackback mechanism) for our sample 35-node network topology (and for other topologies as well), investigating the impact of hierarchical levels on the call establishment times.

References

- [1] ATM Forum Technical Committee, Draft UNI Signalling Performance Test Suite, (ed. J. Orvis, A. Francis) ATM Forum/btd-test-uni-perf.00.09, July 1999.
- [2] BAFUTTO, M. – KUHN, P. J. – WILLMANN, G., Modelling and Performance Analysis of Common Channel Signalling Networks, Hirzel-Verlag, AEU, **47**, No. 5/6, (1993), pp. 411–419.
- [3] BLACK, U., ATM, Vol II., Signaling in Broadband Networks, Prentice Hall, New Jersey, ISBN 0-13-571837-6, 1998.
- [4] ForeRunnerTM ATM Switch Network Configuration Manual, MANU0148-03 – Rev. A – 12/19/97, Software Version 5.1.x, Fore Systems, 1997.
- [5] GDC APEX[®] ATM Switch System, Operation and Installation, Version 4.1, GDC 032R115-V410, Volume 1, General DataComm Inc., 1995.
- [6] GELENBE, E. – KOTIA, S. – KRAUSS, D., Call Establishment Overload in Large ATM Networks, *Proceedings of the ATM'97 Workshop*, Lisbon, Portugal, May 26–28, 1997, pp. 560–569.
- [7] GELENBE, E. – MANG, X. – ÖNVURAL, R., Bandwidth Allocation and Call Admission Control in High-Speed Networks, *IEEE Communications Magazine*, **35**, No. 5, May 1997, pp. 122–129.

- [8] interWatch@95000, Performance and Verification Systems, (CD ROM) DOC 96103695RH, GN Nettest, Release 3.3.0, 2000.
- [9] HP Broadband Series Test System, UNI Signalling User's Guide, Hewlett-Packard Co., 1996.
- [10] ITU-T Draft Recommendation I.35bcp, Call Processing Performance for a B-ISDN, 1997.
- [11] KAUSHAL, A. – SHUMATE, S. – HILL, R. – MURTHY, S. – NIEHAUS, D. – SIRKAY, V. – EDWARDS, B., Performance Benchmarking of ATM Signaling Software, *Proceeding of OPENSIG Workshop*, Columbia University, USA, October 1997, <http://www.ittc.ukans.edu/~niehaus/>.
- [12] KOLYVAS, G. T. – POLYKALAS, S. E. – VENIERIS, I. S., Performance Evaluation of Integrated IN/B-ISDN Signalling Platforms, *Computer Communications*, **21** (1998), pp. 606–623.
- [13] MAUROGIORGIS, M. – PAPADOUKAKIS, N. – SYKAS, E. – TSELIKIS, G., 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.
- [14] MERTZ, A. – POLLAKOWSKI, M., *xDSL & Access Networks, Grundlagen, Technik und Einsatzaspekte von HDSL, ADSL und VDSL*, Kapitel 6, Prentice Hall, ISBN 3-8272-9593-9, Germany, 2000.
- [15] MainStreetXpress 36170 General Information Book and Technical Practices, System Release 4.1, NNP 95-4936-01-00-B, Newbridge, October 1999.
- [16] NIEHAUS, D. – BATTOU, A. – MCFARLAND, A. – DECINA, B. – DARDY, H. – SIRKAY, V. – EDWARDS, B., Performance Benchmarking of Signaling in ATM Networks, *IEEE Communications Magazine*, **35** No. 8, August 1997, pp. 134–143.
- [17] NOVAK, J. – POUÉLÉ, A., Interim Report on the Results of the Quantum Test Programme, November 1999, <http://www.dante.net/quantum/qtp/QUA-99-070.pdf>.
- [18] ONVURAL, R. O. – CHERUKURI, R., *Signaling in ATM Networks*, Artech House, ISBN 0-89006-871-2, 1997.
- [19] ATM Forum Technical Committee, PNNI Draft Specification, Version 1.0, ATM Forum/94-0471R11, 1994.
- [20] 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.
- [21] SZABÓ, I. – SZÉKELY, S. – MOLDOVÁN, I., Performance Optimisation of AAL2 Signalling for Supporting Soft Handoffs in UMTS Terrestrial Radio Access Networks, *5th IEEE Symposium on Computers and Communications, ISCC'00*, Juan les Pins, France, 4-6 July 2000, pp. 46–52.
- [22] SZÉKELY, S. – MOLDOVÁN, I. – SIMON, CS., 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.
- [23] SZÉKELY, S. – SZÜCS, G. – SIMON, CS., 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.
- [24] STILLER, B., A Survey of Signaling Systems and Protocols for ATM Networks, *ACM Computer Communications Review*, **25** No. 2, April 1995, pp. 21–33.
- [25] WU, C. S. – JIAU, J. C. – CHEN, K. J. – CHOY, M, Minimizing Call Setup Delay in ATM Networks via Optimal Processing Capacity Allocation, *IEEE Communications Letters*, **2** No. 4, April 1998, pp. 110–113.
- [26] ATM Forum Technical Committee, ATM User-Network Interface (UNI) Signalling Specification, Version 4.0, ATM Forum/95-1434R8, April 1996.
- [27] XpressPass 140/140HD/142/144 General Information Book and Technical Practices Manual, System Release 4.3, CD-ROM (P/N: CD-1400043), Seabridge, June 2001.