5

Overload generated by signalling message flows in ATM networks

S. Székely, I. Moldován, Cs. Simon
High Speed Networks Laboratory, Department of Telecommunications and Telematics, Technical University of Budapest
Sztoczek u. 2, H-1111 Budapest, Hungary, Fax: +36 1 4633107
{szekely, moldovan, simon}@ttt-atm.ttt.bme.hu

Abstract

Although the importance of the signalling performance of ATM networks has been recognized as a potential bottleneck (Gelenbe, 1997a), very few papers address the congestion situation in switches due to signalling message flow.
In this paper the overload generated by signalling message flow in ATM networks is investigated by measurements and simulation. The results obtained by measurements*** on a university campus network highlight the strong influence of call attempts arriving in a burst on setup time, and form a basis for the simulation model. We simulate the flow of call establishment messages to estimate the queue lengths of signalling messages at access and intermediate nodes, the call blocking probability of different traffic classes and the average round trip time delay (RTT) of connection establishment in two cases with or without wide-band Blocked Call Queueing (BCQ).
In B-ISDN the call blocking probability (CBP) of wide-band (WB) calls is much higher than that of narrow-band (NB) calls'. The introduction of queueing the WB calls rather than reject is going to reduce the CBP of WB calls, at the expense of a small increase in setup time and an increase in CBP of NB calls. The signalling overload associated with WB BCQ and questions related to grade of service are also investigated. The main contribution of the second part of this paper is to show, that the WB BCQ mechanism does not cause congestion of signalling message flow at the network level, when it is applied for moderate overload conditions.

Keywords
signalling protocol, round trip time delay, blocked call queueing, call retrial

---

*** This work was supported by Ericsson Traffic Lab. and HSN Lab., Hungary

Performance of Information and Communication Systems U. Körner & A. Nilsson (Eds.)

1. INTRODUCTION

Very few academic and industrial papers have appeared in the field of ATM signalling performance evaluation until now. (Gelenbe, 1997a) and (Gelenbe, 1997b) argue that congestion can occur at the nodes due to request messages themselves, involving path selection, routing and call establishment. In addition, it offers a simplified analytical model to obtain the call blocking probabilities.

Signalling performance measurements on ATM switches are in focus in the ATM Forum. Automated test cases for performance testing addressed the following aspects of UNI signalling layer: limits test, burst measurements, latency test and endurance test (AF-SIG, 1997). The proposed test suite is a full automated testing solution. The test suite will be responsible to test some performance aspect of UNI signalling (first UNI 3.0, UNI 3.1, Q.2931, then UNI4.0 and PNNI).

In this paper the overload generated by signalling message flow in ATM networks is investigated by measurements and simulation. Our aim is to give an estimate of the RTT delay, and CBP and to determine the signalling CPU processing queue length at each node (access and backbone) as function of different signalling traffic load and network topology. For simulation we use some similar assumptions as (Gelenbe, 1997a). In addition to that, we have implemented the BCQ mechanism for wide-band calls and investigated the RTT delay, the CBP, not only the queue lengths, while using different topologies and signalling CPU speed.

Section 2 presents signalling performance measurement results on some commercial ATM switches. The results obtained from measurements are incorporated into the simulation model. Section 3 motivates the introduction of queueing of blocked wide-band calls at the access nodes of the broadband network, rather than simply accept or reject them. The simulation model is given in Section 4, and comparative results are shown to highlight the difference between the two mechanisms, with or without wide-band BCQ in terms of RTT delay, CBP and queue lengths. It is shown that for moderate overload conditions, call queueing of wide-band blocked calls achieves a substantial reduction in the loss probability at the expense of a small call establishment delay. Section 5 draws some conclusions. Finally, the Appendix presents an extension of the UNI signalling protocol in order to support this new B-ISDN supplementary service at the user-network interface, called Blocked Call Queueing.

2. SIGNALLING PERFORMANCE MEASUREMENT TEST RESULTS

The aim of this section is to show by measurements the overload generated by signalling messages on commercial ATM switches. The testing configuration is very simple, one isolated ATM switch connected to a tester (as calling user) and to an other terminal equipment (called party). To emphasise the signalling overload in the ATM switches, we have chosen a switch with a slow CPU service rate. The limits, latency and burst testing results are shown in Table 2.1 and Figure 2.1 a), b). Table 2.1 presents the RTT delay for the first and last successful connections establishment within a burst and the average RTT delay is also determined, for a burst containing subsequently 10, 20, ..., 100 SETUP messages. We have measured the total processing time, and monitored the number of successful and

rejected calls. Because of T303 timer expiry, some SETUP messages have been retransmitted for a burst containing more than 30 SETUP messages. Except the last two columns, the results are obtained for the standard values specified in (AF-UNI, 1996). T303 and T310 are two timers at the UNI interface and are shown in detail in Figure A.1. in the Appendix. It is important to see that for the shortest path possible (only two links) the RTT delay is of the order of seconds when having bursty arrivals.

Table 2.1. Burst and limits measurement test results. Note: n.v. = not valid (because of the testing equipment limits we could not obtain valid results here)

# of initiated Setup in a burst:	*RTT delay [ms] (between Setup and Connect)*			*# of resent Setup msg*	*Total proc. time [ms]*	*# of reject ed msg.*	*T303 increased from 4 sec to 100*	
	First conn	*Last conn.*	*Average RTT*				*Total proc. t.*	*Reject msg.*
10	405	2 414	1 705	0	3 238	0	3 271	0
20	596	4 819	2 998	0	4 994	0	5 003	0
30	425	4 269	6 418	5	6 823	5	6 850	5
40	822	10 684	6 364	13	11 341	11	8 600	10
50	886	9 840	4 505	4	15 330	9	11 500	11
70	609	14 912	7 950	6	23 250	33	16 390	13
80	n.v.	n.v.	n.v.	n.v.	50 000	50	30 000	30
100	n.v.	n.v.	n.v.	n.v.	65 000	70	30 000	42

The line 50 shows some anomalies. Then line 80 and 100 face a huge number of refused connection, but we could not capture the detailed time stamps, because of the tester's buffer capacity limits. The total processing time is longer than T303 and T310 expiry time, that causes some of the retransmissions and rejected calls. For a burst of 100 initiated calls, the number of successful calls increase to 70 when we increase the values of timers T303 and T310 high above the standard values (not figured in Table 2.1). Even though we still have 30% rejected calls. This happens because of buffer overflow in the processing queue of the signalling CPU.

Figure 2.1 a) shows that the RTT delay of a call establishment increases linearly with the number of open connections (the presented values are obtained for user--one switch--user configuration). Figure 2.1 b) shows that the influence of a bursty signalling arrival on the RTT delay is much stronger than that of the number of open connections. E.g., the 10^{th} SETUP message within a burst suffers five times longer delay than the first one. We can easily calculate the average service time of the tested switch's CPU, being $1/\mu = 100$ ms.

Some other commercial ATM switches showed lower signalling CPU service time ($1/\mu = 20$, 33 and 50 ms). These later three results are used in the simulation model in Section 4. The measurement results shown here highly motivated the investigation of signalling performance by simulation on the network level.

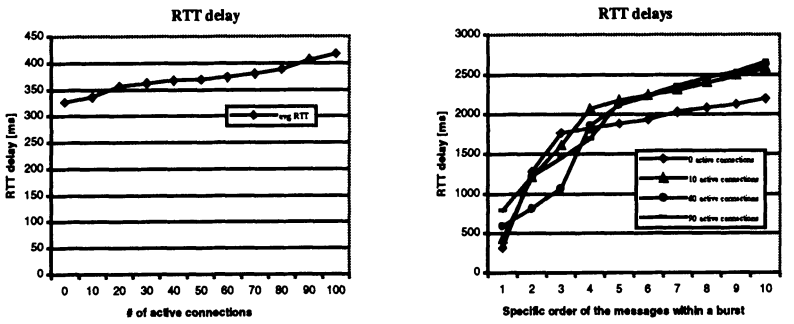

Figure 2.1. (a) The average RTT delay of a new call establishment when we increase the number of active connections from 0 to 100; (b) The average RTT delays for each of the call establishment messages within a burst of 10, having different number of open connections

3. QUEUEING OF BLOCKED WB CALLS AT THE ACCESS NODE

Since ATM is the switching and multiplexing technology of Broadband Integrated Services Digital Networks (B-ISDN's), it is essential for its signalling system that it supports the co-existence of narrow- and wide band services. B-ISDN's have been traditionally modelled on the call scale by the theory of multi-rate loss networks (Ross, 1995). A call request requiring a certain amount of bandwidth between a given originating - destination pair is blocked and disappears from the system if sufficient resources are not available at the time the call request arrives to the network, see e.g. (Ritter, 1994) and (Chung, 1993). Adopting this rule to a multi-rate environment with a large difference in bandwidth requirements between traffic classes implies that calls requiring a large amount of bandwidth will experience a much higher blocking probability than calls requiring only a small amount of bandwidth. By applying either trunk reservation or class limitation it is possible to level out the blocking probabilities. However, in most cases, the disadvantage put on the narrow-band traffic is much bigger than the advantage obtained for the wide-band traffic, because network utilisation is inherently low at multi-rate pure loss networks, when request sizes may differ with orders of magnitude, as it has been shown in e.g. (Sykas, 1991).

The effect of repeated call attempts appears in telephony too, numerous traffic measurements and theoretical investigations were made, see (Elldin, 1967), (Le Gall, 1973), (Gosztony, 1975) and the bibliography attached to them. Measurements showed that several time periods related to the repeated call phenomenon are not exponentially distributed. On the 6th and 7th International Teletraffic Congresses several other papers were presented on this field. Finally, a quite recent paper opened the subject of call queueing in circuit switched network (Berezner, 1996). In this paper we are interested only on a part of the repeated calls, when rejection is caused by congestion in the network. In such a case, according to our proposal, there is no need of any action by the calling party.

Since with an advanced signalling protocol it may be possible to allow calls requiring a large amount of bandwidth to wait in a queue until resources become available, there is a hope to significantly reduce the blocking probabilities of these calls. These types of systems constitute an important generalisation of the pure loss systems and have already been studied in the literature as "mixed delay and loss" or "mixed queueing and loss" systems. In particular, numerical examples indicate that per-class blocking probabilities of wide band services decrease at the expense of a short delay during call set up and a slight increase of narrow band service class blocking. It is concluded that letting wide band calls to queue can decrease wide band blocking and increase network revenue and thus advantageous for both users and network operators (Szekely, 1996), (Fodor, 1999). In fact, these papers do not take into account the effect of signalling message flows and secondly, they look at a rather small network only (4 node fully connected network).
The queueing of wide-band calls has been found as being efficient at the access node of the network, but it requires enhancement of the Call Admission Control (CAC) function to queue the unsuccessful wide-band calls. The Blocked Call Queueing mechanism holds some blocked calls by storing their signalling information in a buffer at the access ATM switch. These calls will be later connected when network resources become available.

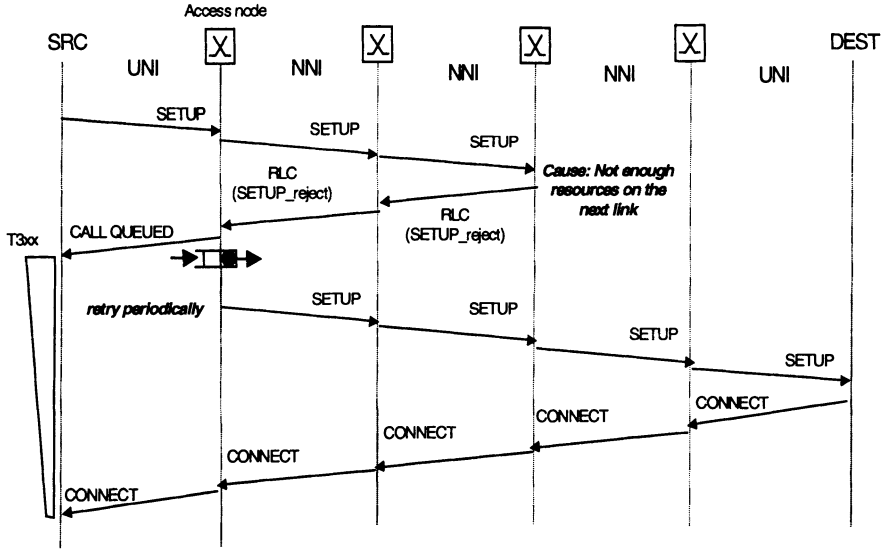

Figure 3.2 Successful call establishment with Blocked Call Queueing (simplified)

In case that a wide-band call attempt arrives to the access node of the network, the CAC function tests whether the network resources are available and proceeds either with resource allocation or the unsuccessful call joins the waiting queue.
It is not trivial to determine which mechanism is better: to retry after ***"d"*** delay (according to the mean holding time of call type *"t"*, MHT_t.) or get feedback from the network first, then proceed again. We are mainly interested in the signalling overload it generates on intermediate and access nodes and particularly its effect on

the RTT delay. We assume here out-of-band signalling, and that signalling channel (VCI=5) is never congested (but it may create overload on the signalling CPU's and suffer long delays).

As a solution for Blocked Call Queueing we have been using the existing UNI and public NNI protocols. A simplified example is given in Figure 3.2, when a call is blocked at the first attempt somewhere in the network because of unavailable resources on a link, rejected, then queued at the access node, waiting for a certain time and finally it succeeds at the second trial. The CALL QUEUED message and the T_{3xx} timer are not part of the standard UNI specification (ITU-T, 1994), but an extension of that, which is described in the Appendix.

To indicate the best algorithm for re-sending the SETUP message after the blocked call has been queued, is subject of further investigations. Hence, we assume that these messages are retransmitted within a *'d'* delay as many times as necessary until success, i.e. a positive acknowledgement (CONNECT message) arrives back to the source.

We assume here a preventive control at each node, by regulating the admission of new calls into the network according to their required equivalent bandwidth. The difference between the total and used capacity on the outgoing link has to be compared to the capacity of the new call $C(\alpha)$. The new call is rejected if:

$$\exists X_k \in \pi_n(i,j): \quad C_{available}(X_k) = C_{total}(X_k) - C_{used}(X_k) < C(\alpha), \quad \forall n = 1,2,...,N$$

where: X_k = link k, $\pi_n(i,j) = n^{\text{th}}$ path between source i and destination j

$C(\alpha)$ = capacity of call α, $\quad C_{used}(X_k) = \sum_{\alpha \in X_k} C(\alpha)$

$C_{total}(X_k)$ = total capacity of link k, $\quad N$ = number of retrials

According to our assumptions, if a narrow-band call is rejected, that is lost. In case of a rejected wide-band call, that is queued at the access node. At a possible next trial a new path is searched again for that call. The round trip time delay of this wide-band call (*RTT*) is bounded by:

$$T_{est}^{main}(i,j) \le RTT(i,j) \le \sum_{n=1}^{N} (T_{est}^{n}(i,j) + d_{alt,n}) \le T_{3xx} + T_{est}^{main}(i,j)$$

where: $d_{alt,n}$ = time spent in the access queue at the n^{th} retrial,

$T_{est}(i,j)$= call establishment time on one specific path,

T_{3xx}= time-out value of the BCQ timer

The capacity check is inherently a sequential process. If we assume that all nodes have the same service rate (μ) and we count each message transmission, reception and processing as a delay unit, it introduces a time delay which is linear with the length of the path ($L(\pi_n(i,j)$= number of nodes in the path "*n*"). In most practical network, this linear time cost is not critical since the diameter of the network is kept bounded by the topology design because of end-to-end considerations. However, in principle if the end-to-end paths are excessively long, this linear time cost may be a delay bottleneck in the process of the call setup, e.g.:

1) 20 nodes in the path , when $1/\mu$=50ms for nodes, no queue ➔ *RTT*=2 sec;

2) 20 nodes, $1/\mu$=50ms, but avg. queue length=20 ➔ *RTT*=40 sec > T_{310}=30 sec!

The expected value of the RTT for a wide-band call is given by,

$$RTT(i,j) = E(T_{est}(i,j)) = T_{est}^{main}(i,j)\cdot(1-p_{main}(i,j)) + d_{alt,1} +$$

$$+ T_{est}^{alt,1}(i,j)\cdot(1-p_{alt,1}(i,j)) + d_{alt,2} + \ldots$$

where: $p_{main}(i,j)$ = probability of unsuccessful call establishment on the main path between nodes i and j

$p_{alt,r}(i,j)$ = probability of unsuccessful call establishment on the alternative path "r" between nodes i and j, if the main path is blocked.

For simplification, let us consider:

$$T_{est}^{main}(i,j) = T_{est}^{alt,r}(i,j) = T(i,j) \text{ and } p_{main}(i,j) = p_{alt,r}(i,j) = p_0,\ r = 1,2,\ldots,N$$

Then it results : $RTT(i,j) = T(i,j)\cdot N\cdot(1-p_0) + \sum_{r=1}^{N} d_{alt,r}$

The $T(i,j)$ depends on the processing speed of signalling CPU on each switch, the signalling traffic load in the network, the number of switches in the path, the channel bit rate and the message format (e.g. the DTL information element in PNNI protocol is a source of overhead). While using 155Mbps optical interfaces we can neglect the dependency on the channel bit rate. Further on, if p_{kZ} is defined as the call blocking probability on the link k directed from node Z,

$$p_0 = p_{main}(i,j) = 1 - \prod_{X_k \in main\_path(i,j)} (1 - p_{kZ})$$

To find the probabilities p_{kZ} for all nodes is a problem. If all links in the backbone have the same CBP, p_{BB} as well as all links from the external network have p_{ext}, and $E(L(\pi_n(i,j)))$ is the average length of the path, moreover if $p_{BB} = p_{ext} = p$, then:

$$p_0 = 1 - (1-p)^{E(L(\pi_n(i,j)))}$$

The average number of calls on the link k directed from node Z is:

$E_{kZ}(m) = \rho_{kZ}(1-p_{kZ})$, then

$$T(i,j) = 2\cdot \sum_{X_k \in main\_path(i,j)} E_{kZ}(m)\cdot E_k(T) = \frac{2}{\mu}\cdot \sum_{X_k \in main\_path(i,j)} E_{kZ}(m), \text{ while}$$

$$d_{alt,r} = d = \frac{1}{\mu}\cdot\rho\cdot(1-p_0), \quad r = 1,2,\ldots,N$$

where $\rho = \lambda/\mu$ is the utilisation of one access node. The above formulas are even more complicated if we consider different CBP for different types of calls.

Instead of considering the average RTT delay it may be more appropriate to consider the maximum delay, which is the delay experienced by an arrival to the queue that finds (m-1) queued calls in the node. The grade of service (GoS) requirement is stated in terms of both the loss probability and the call establishment delay. Queueing will therefore be beneficial if the maximum setup delay satisfies the GoS requirements. The distribution of the maximum delay at a node is a gamma distribution with parameters ($1/\mu$, m-1). The expected value of the maximum delay (in a node) is $(m-1)/\mu$ and the variance is $(m-1)/\mu^2$. The variance is thus very small for a range of parameters under consideration and the maximum delay is very close to its expectation. Call queueing will therefore be

beneficial if $RTT < d_{GoS}$, the maximum delay for an acceptable service. So the timer T_{3xx} (see Appendix) has to be defined as being equal to d_{GoS}.

4. THE SIMULATION MODEL

We simulate the call set-up phase of the ATM connection and the flow of call establishment messages in order to estimate the queue lengths of signalling messages at the ATM switches, the RTT delay and the CBP of different types of calls. In addition we simulate the WB BCQ mechanism (see Section 3) and compare the two methods (with and without WB BCQ).

4.1. Simulation overview

Our evaluation is concerned with the load due to message processing by the nodes in the network. Therefore, we will focus on the number of messages which need to be processed for establishing and terminating a call, and the manner in which these messages are exchanged and routed is particularly relevant. The tested topology is a ring topology, with a core and an access network, where we can modify the number of nodes in the network. Despite of this, the topology used in (Gelenbe, 1997a) consisted of 100 nodes in a 10*10 mesh topology, where all link capacities were 45Mbps, and the call establishment message was processed in 30ms at each node. The processing of a call of type *"t"* from the source node *"i"* to destination node *"j"* with bandwidth requirement $C_t(\alpha)$ is similar in both cases, but we focus on more parameters and in addition we have implemented the WB BCQ mechanism. Moreover, the measurement results presented in Section 2 are incorporated into our simulation model.

Paths may have different lengths, and $L(\pi_n(i,j))$ is the length of the n^{th} path between i and j. Because of the given ring topology (see Figure 4.1), the average length of the path is $E(L(\pi_n(i,j)))< 4$ in the total network, while it is $E(L(\pi_n(i,j)))< 2$ in the backbone network. A call establishment means a subsequent flow of SETUP and CONNECT messages travelling up and down the path. Hence, a new call attempt generates $2 * E(L(\pi_n(i,j)))$ messages in the backbone network. For the sake of simplicity we use a very generic signalling protocol, generating only 4 types of basic messages: SETUP, CONNECT, RELEASE and RELEASE COMPLETE (RLC).

If the call is successfully established, then the bandwidth is reserved for the holding time of the call. Upon termination, the bandwidth is released at each link by a message which travels up the path, from source to destination, and the network state table is then updated.

We have *"n"* nodes fully interconnected in the backbone of the ATM network, and *"n*m"* switches on the external network, connected via double homing. We can specify the capacity of each link separately, and if one link does not exist, we can simply assume that the specified link's capacity is zero. In our case, we considered 4, 5, 6 or 8 switches in the backbone network fully interconnected by 310 Mbps links, and each switch-pair has 4 external nodes connected by 155 Mbps links to the backbone.

There are three types of traffic classes and two scenarios (see Table 4.1). The call arrivals are generated using Poisson distribution, and the source-destination pairs are selected by uniform distribution.
The 'wait and retry' mechanism can be simply implemented in different ways, e.g. a separate queue for blocked calls, different priorities for different types of calls, etc. (Fodor, 1997). Our node model is very simple, it has only one queue and one processor per node. We assume static, alternate routing and there are no restrictions about the length of the processing call queues.

4.2. Simulation results

The following figures highlight some of the simulation results. As a first result we obtained is as follows: the shorter the mean holding time of calls, the larger the bandwidth that is supported by the network. Secondly, the higher the signalling processor speed, the lower the RTT delay of connection setup and higher the bandwidth limit. The specific RTT delays obtained here confirmed the measurement results of Section 2. The minimum RTT delay is given by 2 * the average length of the path * the average service time of the nodes.
Figure 4.1 (a) shows the average queue length of backbone nodes for 3 types of service rate, when increasing the number of call attempts in the network. The bandwidth requirements of connections are given very small to avoid link congestion. The higher the service rate, the smaller the buffer occupancy for the same load. In Figure 4.1 (b) the average backbone queue lengths are presented, when we have a given service rate ($1/\mu$=30ms) and different number of backbone nodes. As we increase the network load above 25 calls/sec, in the configuration with 4 backbone nodes the service time of individual nodes is longer than the interarrival time of incoming calls, then their queues grows rapidly.

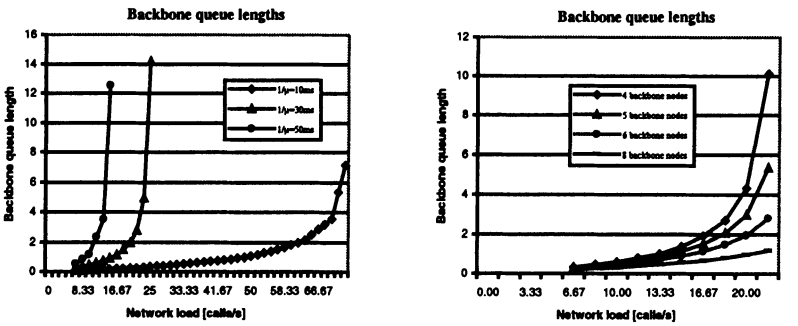

Figure 4.1 (a) Effect of the service time and (b) number of backbone nodes on the queue lenght

The backbone can not carry out the generated load, and congestion of signalling messages occurs. To avoid congestion for the same load, we can increase the number of backbone nodes, thus distributing the load to more switches and reducing the queue length. One solution for this problem is to increase the number

of the backbone switches. If we increase the number of backbone switches from 4 to 5, 6 or 8 (fully interconnected), the load will be distributed and the probability of the congestion decreases. The differences become important starting from medium signalling load (20 calls/sec), e.g. while the average queue length at the backbones is 10 having 4 nodes, it decreases to 1 for 8 nodes in the backbone. So a bottleneck having less speedy switches can be diluted by increasing their number in the backbone network. In the rest of the configurations we have mixed traffic. The network load is increased from low (0.1calls/sec) to high (33calls/sec). All nodes have the same service time ($1/\mu$=30ms). The network topology is the same, using 6 nodes in the backbone. We can re-scale the x-axis (network load) to relative scales.

$$\rho = \rho_{node} = \frac{\lambda_{node}}{\mu_{node}} = \frac{1}{\mu_{node}} \cdot \frac{\lambda_{network}}{6} = \frac{30}{6} \cdot 10^{-3}\,\text{sec} \cdot [0.00 \ldots 33.33] \frac{call}{\text{sec}} \cdot 4 \frac{mesg}{call} = [0.00 \ldots 0.67]$$

Three traffic classes are specified: narrow-band class (1Mbps), medium-band class (10Mbps) and wide-band class. Two scenarios are chosen (see Table 4.1).

Table 4.1. Traffic classes and their distribution

Call Type [Mbps]	*Mean Holding Time [sec]*	*Scenario 1*		*Scenario 2*	
		Call type distribution [%]	*Link occupancy [%]*	*Call type distribution [%]*	*Link occupancy [%]*
1	100	89	33	70	8
10	10	9	33	20	23
60	2	2	33	10	69

In scenario 1 the average link capacity is uniformly distributed between the three types of calls. As a result we have 89% narrow-band calls and only 2% wide-band calls. In scenario 2 the number of wide-band calls is increased, so it constitutes 10% of the total call attempt. The mean holding time for the WB calls is relatively short (2 sec). The maximum waiting time for a blocked wide-band call is set to *T3xx*=30 sec. Only WB call are queued, the other two classes are rejected immediately when network resources are not available for that specific call. The queue length of the access nodes has an average of less than 1 for both scenarios (see Figure 4.2 (a), (b)).

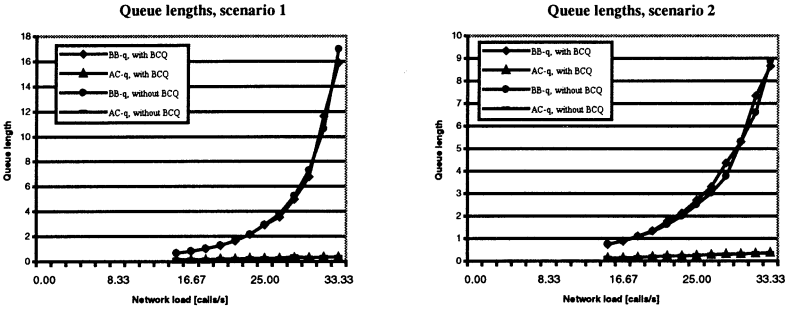

Figure 4.2 Access and backbone queue lengths for mixed traffic and two scenarios

The simulation results are obtained for 95% confidence intervals. The average queue length of the backbone nodes increases to order of 10^{th} for ρ=0.67 (high load). Neither in scenario 1, nor in scenario 2 the WB BCQ mechanism does not have any impact on the queue length of access and backbone nodes. However the queue length is less than 20 for even a high signalling load, one can address schemes to trigger recovery actions (e.g. re-sending REL messages if the correspondent RLC message was not received, because of buffer overflow).
Figures 4.3 (a)-(c) show that for the given scenarios the CBP of narrow-band calls (1Mbps) is not deteriorated by applying the WB BCQ mechanism. The CBP of wide-band calls (60Mbps) drops from 0.8 to 0.5 %, respectively 5.5 to 1.9 % at the expense of a small increase in CBP of medium-band calls, that increases from 0.62 to 0.8 %, respectively from 1.7 to 3%.

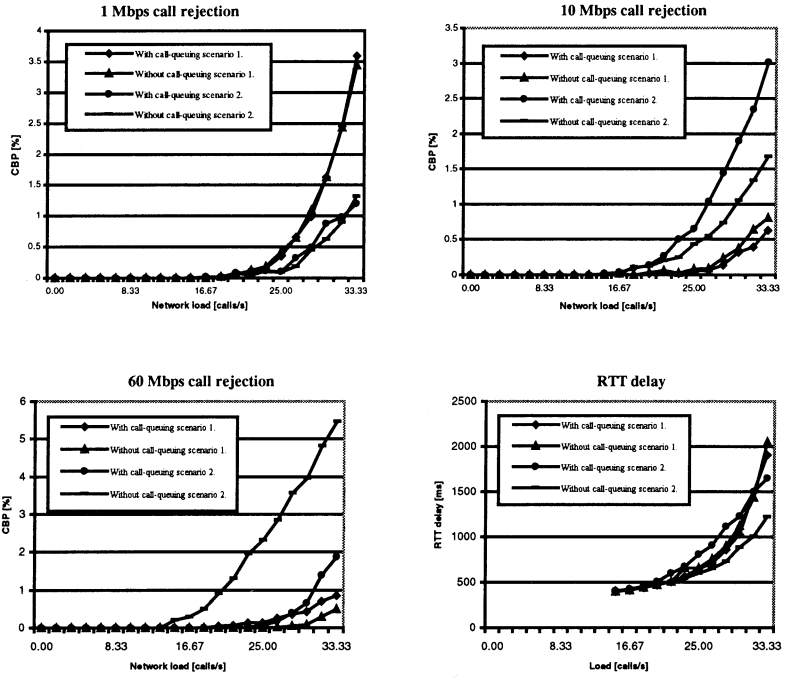

Figure 4.3 (a)-(c) Call blocking probability of narrow-band, medium-band and wide-band calls (d) The average RTT delay for all class of calls

When only 2% of the total calls requires wide-band capacity (scenario 1) the average RTT delay does not change significantly when using WB BCQ. When the WB calls form 10% of the total offered calls, the average RTT delay slightly increases. At a high traffic load (ρ=0.67) the RTT delay is still acceptable, however

the difference between the two curves (with and without BCQ) is equal to the value of RTT for a low traffic (400 ms).
For a low traffic load, the RTT delay is 400ms for the given topology (the avg. length of the path = 4 nodes, and using switches with $1/\mu$=30ms). That means the specific RTT delay/node = approx. 100ms. While at high load conditions RTT/node increases to 400ms with, respectively 300ms without WB BCQ. Hence, real size networks (less than 20 nodes in a path) will have approx. 6-8 sec RTT delay. That satisfies the grade of service requirements. We conclude that WB BCQ is most beneficial when ρ=[0.5 ... 0.67]. When $\rho < 0.5$, the BCQ mechanism is not needed, while for overload conditions ($\rho > 0.67$) this mechanism is not effective.

5. CONCLUSIONS

In this paper the overload generated by signalling message flow in ATM networks was investigated by measurements and simulation. The results obtained by measurements for point-to-point connections showed the strong influence of call attempts arriving in a burst on setup time. By simulation we estimated the performance parameters on the network level, namely the queue lengths of signalling messages, the call blocking probability of different traffic classes and the average round trip time delay of connection establishment in both cases with or without wide-band Blocked Call Queueing. We have given an analytical method to compute the RTT delay. The signalling overload associated with WB BCQ and questions related to grade of service were also investigated. Finally we have shown, that the WB BCQ mechanism does not cause congestion of signalling message flow at the network level, when it is applied for moderate overload conditions. The implementation of WB BCQ was very simple, no complexity problems appeared (see Appendix). We plan to investigate in the future the implications of ABR connections setup and point-to-multipoint calls setup on the network level.

6. REFERENCES

ATM Forum Technical Committee (1994) PNNI Draft Specification, Version 1.0, *ATM Forum/94-0471R11*

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

ATM Forum Technical Committee, Testing SWG (1997) UNI Signalling Performance Test Suite, *ATM Forum/97-0468*

Berezner, S.A. and Krzesinski, A.E. (1996), Call queueing in circuit switched networks, *Telecommunication Systems,* **6**, 147-160

Chung, C-P. and Ross, K.W. (1993) Reduced Load Approximations for Multi-Rate Loss Networks, *ACM/IEEE Trans. on Networking*, 1222-1231

Elldin, A. (1967) Approach to the theoretical description of repeated call attempts, *Ericsson Techn.*, **23**, 345-407

Fodor, G., Blaabjerg, S. and Andersen, A.T. (1999) Modeling and Simulation of Mixed Queueing and Loss Systems, *Kluwer Acad. Publisher*, Personal Wireless Communications, to appear

Gelenbe, E., Kotia, S. and Krauss, D. (1997a) Call Establishment Overload in Large ATM Networks, in *Proc. ATM'97 Workshop*, Lisbon, Portugal, 560-569

Gelenbe, E., Mang, X. and Önvural, R. (1997b) Bandwidth Allocation and Call Admission Control in High-Speed Networks, *IEEE Communications Magazine*, Vol.35, No.**5**, 122-129

Gosztony, G. and Ágostházi, M. (1975) Characteristics of repeated telephone calls (in Hungarian), *Híradástechnika*, **26**, 109-119

ITU-T Recommendation Q.2931 (1994) B-ISDN. DSSS No.2 (DSS2). UNI Layer 3 Specification for Basic Call/Connection Control", *COM 11-R 78-E*

Le Gall, P. (1973) Sur l'utilisation et l'observation du taux d'efficacité du trafic téléphonique, 7th ITC, Stockholm, *Prebook*, 443/1-8.

Ritter, M. and Tran-Gia, P. (1994) Multi-Rate Models for Dimensioning and Performance Evaluation of ATM Networks, *COST 242 Interim Report*

Ross, K.W. (1995) Multiservice Loss Models for Broadband Telecommunication Networks, *Springer Verlag*, ISBN 3-540-19918-7

Sykas, E.D., Vlakos, K.M., Venieris, I.S. and Protonotarios, E.N. (1991) Simulative Analysis of Optimal Resource Allocation and Routing in IBCN's, *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. **3**

Szekely, S., Fodor, G. and Blaabjerg, S. (1996) Call Queueing: The design and performance analysis of a new ATM signalling functionality, in *Proc. B&MW'96 Workshop*, Zagreb, Croatia, 99-113

Szekely, S. (1997) On Bandwidth Allocation Policies in ATM Network using Call Queueing, in *Proc. 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, U.K., 46/1-10

Additional reading

Onvural, R.O. and Cherukuri, R. (1997) Signaling in ATM Networks, *Artech House*, ISBN 0-89006-871-2

APPENDIX. SIGNALLING CAPABILITIES NEEDED TO SUPPORT WIDEBAND BLOCKED CALL QUEUEING AT THE UNI

This section extends the ITU-T Recommendation Q.2931 for point-to-point signalling protocol to support blocked call queueing capability. As shown below, the first message sent by the user to the network for call establishment (SETUP message) needs an extension to support blocked call queueing service. This and some other extensions of the current signalling protocol Q.2931 necessary to support blocked call queueing are as follows: a new message (CALL QUEUED), a new timer (T3xx), a new information element (BCQ IE) and a new state (U*).

Blocked call queueing service can be requested by the calling user's application process. In this case the signalling layer sends a SETUP message across the UNI, which contains the desired Blocked Call Queueing information element (BCQ IE). The network access node may ignore the BCQ IE if that information element is of no interest or that service is not implemented. When the service is implemented but there are not enough resources available in the network, the calling party is notified by a CALL QUEUED message that its call has been set in a waiting queue at the first node of the network. This CALL QUEUED message is sent to the caller only

in the case when he has asked previously for it by BCQ IE. This call is either served within a given time limit or removed from the queue by the calling party when Blocked Call Queueing timer T3xx expires. The Finite State Machine graph at the user side of the UNI in Figure A.1 gives a detailed description about all the possible scenarios. The timer T3xx is used on the user side. The network side FSM graph is very similar and is not shown here because of lack of space.

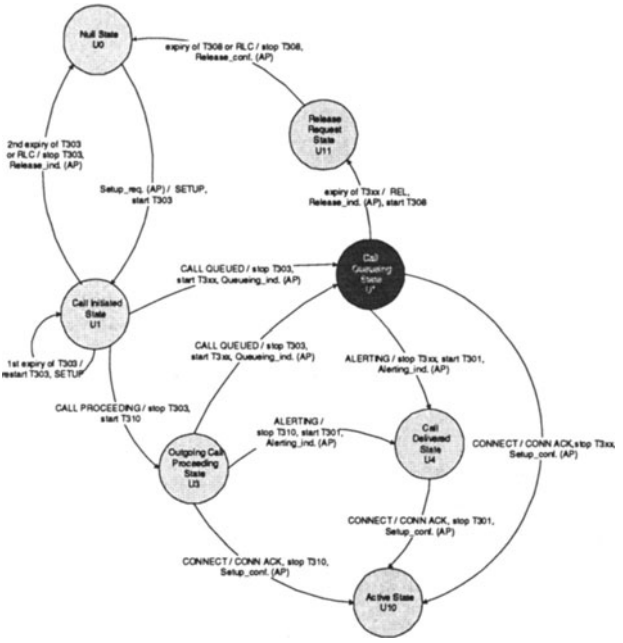

Figure A.1. Partial FSM graph of Q.2931 with the new BCQ state (user side)

7. BIOGRAPHY

Sándor Székely received the M.Sc. degree in communications engineering from the Technical University of Timisoara, Faculty of Electrical Engineering, Timisoara, Romania, in 1995. In 1994 he joined the High Speed Networks Laboratory at the Department of Telecommunications and Telematics, Technical University of Budapest, Hungary, where he is currently working towards the Ph.D degree. His research interests are related to optimisation of signalling protocols in ATM networks, and performance analysis of call establishment by measurements, simulation and analytical study. He is a student member of IEEE since 1997.

István Moldován received the M.Sc. degree in computer engineering from the Technical University of Tirgu Mures, Faculty of Automation, Tirgu Mures, Romania, in 1996. Now he is a Ph.D student at the TU of Budapest.

Csaba Simon received the M.Sc. degree in computer sciences from the Technical University of Timisoara, Faculty of Computer Science, Timisoara, Romania, in 1997. Now he is a Ph.D student at the TU of Budapest.