Budapest University of Technology and Economics

Lightweight Control Techniques for Supporting Voice Communication over Packet Switched Networks

István Szabó
High Speed Networks Laboratory
Department of Telecommunications and Telematics

Summary of the Ph. D. Dissertation

Advisor:
Dr. Sándor Molnár
High Speed Networks Laboratory
Department of Telecommunications and Telematics

Budapest, Hungary
2003

1 Introduction

There are three important trends in the development of telecommunication systems over the past couple of years. The first important development is the evolution of public landline mobile network infrastructure, which makes possible to offer sophisticated real–time multimedia and high speed data transfer services to mobile customers in addition to the widely used mobile telephony. This new networking infrastructure is referred as 3G or UMTS [1, 2]. Tremendous research effort has been committed to extend IP networks with the capability of offering differentiated services and providing QoS guarantees for a subset of traffic flows. Finally, these two networks are converging, IP transport network will be deployed in the fixed part of the mobile infrastructure and mobile terminals are expected to include an IP protocol stack and be capable of accessing web and real–time multimedia content over the Internet. My research was focused on *control techniques*, which can be used to support carrying real–time (mainly narrowband voice) traffic over packet switched networks, and it takes also into account that the above mentioned evolution and convergence results in a fast changing and quite unstable networking environment, where *incremental deployment* of new services and solutions is a must.

In cellular wireless systems the most precious resource is the radio spectrum. In order to effectively support packet switched services over the radio interface, a new technology, W–CDMA [3] has been developed for UMTS, which enables guaranteeing QoS over the air interface, while allocating radio resource with fine granularity. After completing W–CDMA, research has been started to build a fixed transport network infrastructure (UMTS Terrestrial Radio Access Network) around it. At that time ATM has been the most mature network technology, capable of matching W–CDMA in terms of QoS guarantees and flexibility of resource allocation. In addition, many incumbent telecommunication operators —especially in Japan, where the need for the next generation of cellular networks was most articulated— had huge capacity ATM networks, which they wanted to utilize. In order to efficiently support compressed voice communication, and to satisfy other radio access network specific requirements ATM needed a major enhancement. A new adaptation layer (AAL2 [4]) has been developed, and a corresponding new switching level has been introduced. My research was focused on *a new signalling protocol* which makes possible the introduction of this new switching level.

In the meantime, IP took off as a promising candidate to be the base technology for a QoS enabled, multiservice network, and more and more enterprises and new telecommunication service providers wanted to carry not just data traffic but also telephony sessions over IP networks. In order to provide toll quality voice communication, the traditional best effort service paradigm is not sufficient. Some method is needed to offer prioritised treatment for real-time traffic, and block new call arrivals if the capacity limit of the network is reached. Many approaches have been proposed in the past couple of years for service differentiation and resource allocation in IP networks [5, 6, 7, 8, 9]. These methods represent different trade–off between efficient utilisation of bandwidth and implementation complexity. A significant problem hindering the widespread deployment of IP QoS mechanisms is the large installed base of equipment supporting only the best effort service. My research was focused on finding *a lightweight call admission control method*, which is capable of providing soft, but reasonable

QoS guarantees for IP telephony, and the offered QoS guarantees are not much effected by intermediate islands of best effort routers.

Recent measurement results show that the Internet is dominated by TCP traffic [10], but the fraction of real–time multimedia content is steadily increasing [11]. Floyd and Fall [12] considers the potentially negative impacts of an increasing deployment of non–congestion–controlled traffic over the Internet. It concludes that all applications (including real–time multimedia services) deployed over the Internet should use some sort of end–to–end congestion control. Instead of deploying end–to-end congestion control for voice applications, one can regulate the amount of real–time traffic in the network by blocking new real–time call arrivals whenever the amount of network resources is found to be insufficient. My research was focused on *investigating the interaction of call level admission control and end–to–end TCP congestion control* when trying to support narrowband voice traffic over a best effort IP network.

2 Research Objectives

The general aim of my research was to find and solve open research problems with regards to efficiently carrying (compressed) voice traffic over packet switched networks.

I dealt with 3rd generation mobile radio access networks (UTRAN). The goal of my research was to propose a new signalling protocol (AAL2 Signalling), which is capable of supporting a new switching level (AAL2 switching) on top of the well–known ATM switching. Earlier, signalling solutions were tied to a specific bearer network. Further goal was to come up with a new protocol architecture, which allows AAL2 Signalling to be independent of the underlying transport network, and makes it possible to operate the protocol over IP, ATM or traditional SS7 signalling network. Finally, I had to analyse protocol implementation options to see their effect on the speed of connection establishment, which is a crucial factor when it comes to supporting soft handover in UTRAN.

One of the most important design principle of Internet was the so called "end–to–end argument"[13]. Applying this principle means that the network nodes should be bothered only with functionality, which is required by all applications, and the functionality, which is required only by a subset of end–systems shall be implemented by applications. QoS support and call admission control is definitely a functionality, which is not needed for all end–users, so applying the end–to–end argument literally routers in the network should not be burdened with call admission control. The goal of my research was to propose a call admission control solution which does not require any involvement from the routers, and to evaluate whether it is possible to perform call admission control based on passive end–to–end measurements and to find out what sort of QoS guarantees such a solution can provide.

Finally, with the increase of real–time content over the IP infrastructure, it is more and more important to deploy control algorithms for real–time traffic to prevent expelling legacy TCP flows. The goal of my research was to study the interaction of end–to–end measurement based call admission control and TCP congestion control in a best–effort IP network.

3 Methodology

Part of my research work was to conceptualise two new control protocols. As a proof of concept of the proposed protocols, both of them were implemented. The AAL2 Signalling Protocol were implemented even in commercial telecommunication systems. On the other hand, I developed the prototype implementation of the proposed call admission control method as an extension to ns–2 simulator [14].

In the performance evaluation and characterisation work, I used elementary queuing theory if it was possible but I mainly used simulation because of the following reasons:

- In most of the cases, the complexity of the investigated scenarios and systems does not make possible a proper analytic description.
- ns-2 is basically a de–facto standard platform for simulation based evaluation of IP networks, which very much decreases the time needed for the development of simulation models for new methods.
- The widespread use of the same simulator gives good confidence in the obtained numerical results.

4 New Results

Thesis 1 : A Signalling Protocol for Supporting Switched AAL type 2 Connections in UMTS Terrestrial Radio Access Networks

ATM AAL2 [4] has been selected as the transmission technology in the radio access network of UMTS systems, mainly because it is ideal for carrying compressed voice traffic [15]. To suit AAL2 to a network where support for soft handoff is essential, an additional switching level on top of ATM has been created. After thorough analysis of existing ATM signalling protocols and different realisation alternatives, it was found that none of the existing signalling protocols is suitable for controlling this new level. Most importantly, all ATM signalling solutions were combined call and connection control protocols carrying also complex traffic management information. In this thesis I propose a new protocol (AAL2 Signalling) and a protocol architecture for handling on–demand, switched AAL2 connections.

Thesis 1.1 : AAL2 Signalling Protocol [C1, J2, J3]

I proposed a new signalling protocol for controlling switched AAL2 connections.

The protocol has the following features:

- It is a separate, new connection control protocol and not an extension of any existing ATM call and connection control signalling solution.
- It is able to control AAL2 connections on multiple underlying ATM VCCs that can either be permanent or switched ones.

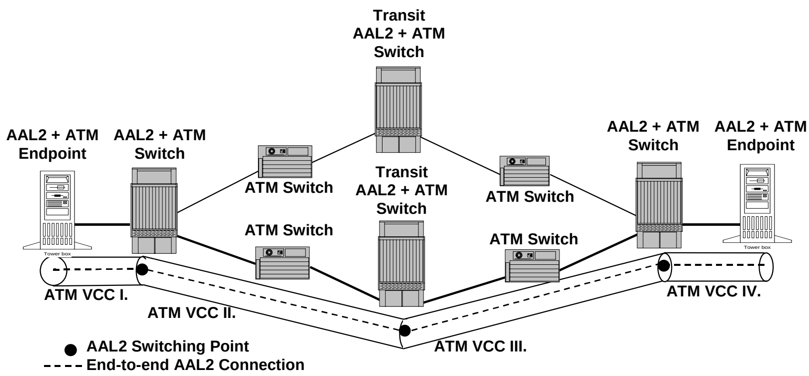

Figure 1: ATM virtual connections and an end–to–end AAL2 connection in an AAL2 overlay network

- It provides mechanisms for the establishment and release of point–to–point AAL2 connections over an AAL2 network comprised of AAL2 endpoints and AAL2 switching points. (Figure 1.)
- A forward and backward compatibility mechanism is included in the protocol that makes it possible to extend the protocol in the future without causing problems for switches operating according to an earlier version of the specification.
- It supports transparent transfer of served user generated information between originating and terminating AAL2 connection endpoints which enables the co–ordination of radio channels and corresponding fixed network connections in a UMTS radio access network.
- It supports hop–by–hop routing.
- The protocol provides basic maintenance capabilities to cope with congestion in the signalling network and temporary outage of signalling links and peer switching entities.

Probably the best validation of the proposed protocol is that it has been implemented by many companies (Ericsson [16], Trillium [17], Spirent Communications [18]), and it is now part of commercial product offerings. Further details about the validation and formal specification of the protocol can be found in Section 2.2.1 of the dissertation. The proposed protocol operates at a new, additional switching level on top of ATM switching, therefore its performance and functionality can not really be fairly compared to existing ATM signalling solutions. The benefits of this additional switching level can be shown at an overall system level. For a system level analysis of AAL2 switching see [J2, 19].

Thesis 1.2 : Bearer Independent Signalling Protocol Architecture [P1, C1]

I introduced the concept of *Bearer Independent Signalling Protocol*, which makes it possible for a signalling protocol to operate on top of any underlying bearer network, for example SS7(MTP), IP or ATM.

AAL2 can be used in many different application scenarios, and AAL2 signalling should be designed to enable easy deployment of the new protocol in all those networks. The question that needs to be answered is which the ideal signalling bearer protocol stack is for carrying AAL2 signalling messages.

If we consider only the needs that can be deducted from the capabilities supported by the AAL2 signalling protocol, the reasonable choice is to follow ATM Forum PNNI's approach and connect the adjacent AAL2 switches with AAL5 based Signalling AAL ATM VCC links (ITU-T Q.2110 & Q.2130).

If the UMTS Terrestrial Radio Access Network case is considered, the broadband SS7 network (that is indispensable for carrying the mobile network related signalling) can be reused also for carrying the AAL2 signalling. Reusing the SS7 network simplifies the application of AAL2, since it eliminates the need for maintaining a separate signalling network for the AAL2.

In UMTS core networks and in landline trunking (voice over ATM) application, the legacy system used for carrying signalling messages is the narrowband SS7 network, therefore being able to carrying the AAL2 signalling on narrowband SS7 is also beneficial.

Considering the ongoing effort in IETF with the aim of transporting SS7 signalling on IP, design of a future proof AAL2 signalling should mean that IP is also supported as a signalling bearer.

The above listed arguments leads us to the conclusion that the notion of "bearer independent signalling protocol" should be introduced and the AAL2 signalling protocol should be designed in a way that it can theoretically be carried on any available signalling bearer protocol stack. The resulting protocol architecture for AAL2 Signalling is depicted in Figure 2.

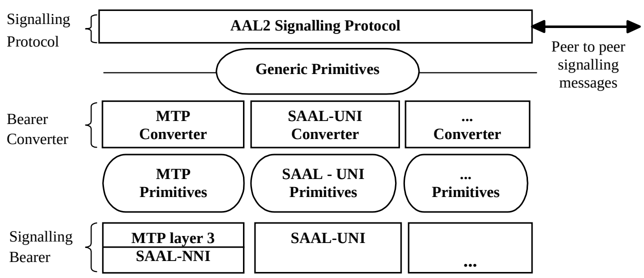

Figure 2: AAL2 Signalling Protocol Architecture

Describing the concept at a high level, the Signalling Bearer Converter provides a set of "generic primitives" that is used by the signalling protocol when exchanging signalling messages with the peer signalling entities and receiving information about the conditions in the signalling network. The Signalling Bearer Converter is responsible for translating these "generic primitives" into primitives offered by a concrete signalling bearer protocol and vice versa. It hides all differences between signalling bearer services from the signalling protocol. It can monitor the availability of remote signalling entities and the signalling links in the network. The bearer converter can implement segmentation and reassemble or a sequence numbering and retransmission service depending on the requirements imposed by the underlying layer that is intended to be used as signalling bearer below the signalling protocol.

The flexibility provided by the new architecture comes for some additional implementation complexity. To prove the viability of the concept of Bearer Independent Signalling Protocol, I provided two bearer converter designs. Both of them operate for AAL2 Signalling. The first one makes it possible to run the signalling protocol on top of SS7 networks, while the second one relies directly on the Signalling ATM Adaptation Layer as it is standardised for the UNI interface. Probably the best validation of the proposed bearer converter designs is that both of them have been implemented (Trillium [17], Spirent Communications [18]), and they are now part of commercial product offerings. As an example, the operation of the SS7 bearer converter can be followed in Figure 3.

A real challenge for those, who design, build and operate telecommunication networks today is to handle the convergence from different networking solutions towards a multiservice IP network. In this transition phase a large installed based of ATM equipment operates which carries more and more PSTN traffic. This PSTN traffic is controlled using the old SS#7 signalling system. On the other hand, continuous improvement of IP technology makes it more and more capable for carrying signalling and real–time data. The main benefit of the proposed protocol architecture is that it provides a tool for efficiently coping this transition by enabling the signalling protocol to be independent of any particular networking technology, therefore avoiding the need for protocol redesign when migrating a telecommunication system to a new networking technology. The proposed principle is followed by many new protocols for example the Bearer Independent Call Control Protocol (ITU-T Q.1901) and the Stream Control Transmission Protocol (IETF RFC 2719).

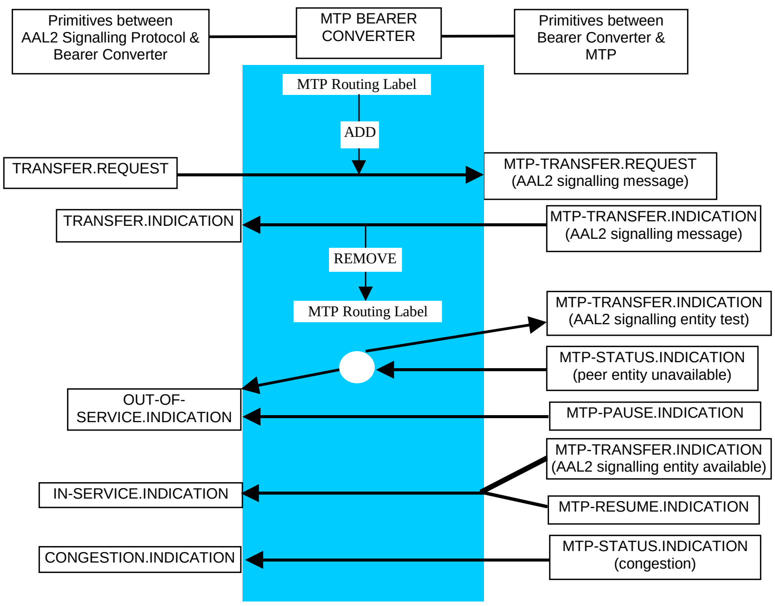

Figure 3: Primitive Mapping in the MTP Bearer Converter

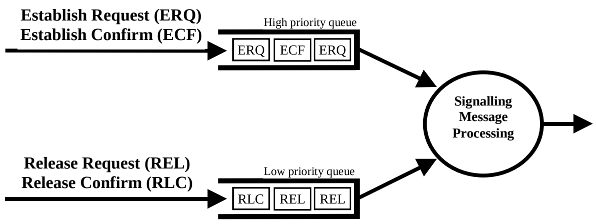

Figure 4: Priority Queuing of AAL2 Signalling Messages

Thesis 1.3 : Optimising the Performance of AAL2 Signalling [C2]

I proposed to implement AAL2 Signalling using priority queuing of the protocol messages. I evaluated the solution using elementary queuing theory and simulation.

As mentioned earlier, AAL2 has been selected for transporting soft handover legs between base stations and diversity combining devices located in radio network controllers. When it comes to supporting soft handover, the most important requirement is fast connection establishment.

A way to increase the speed of AAL2 connection establishment is to introduce priority handling of signalling messages, particularly, to assign higher priority to the messages involved in the connection establishment. This priority mechanism is depicted in Figure 4, where it is shown that two queues are implemented in all the processing elements of the AAL2 switch. The first queue collects the *ESTABLISH REQUEST (ERQ)* and *ESTABLISH CONFIRM (ECF)* messages. The messages in this queue get absolute priority over messages in the second queue, which means that the server starts fetching messages from the second queue only if the first one is empty. The second queue collects the messages that are used for connection release. Such a mechanism is widely used for user plane traffic, but it is not that typical in the control plane.

Single Node Case

In case of a single AAL2 switch, the effect of prioritisation can be determined using elementary queuing theory [20]. An AAL2 switch can be modelled as an M/G/1 queuing system. The decrease in connection establishment time resulting from the priority queuing can be expressed by dividing the average waiting time of the connection establishment messages in the priority case with the average waiting time of the messages in the non–priority case. This quotient can be written as:

$$\frac{1-\rho}{2-\rho} \tag{1}$$

where ρ is the utilisation factor. Figure 5 shows the percentage gain as defined above when the utilisation of the signalling processor is gradually increased.

AAL2 Switching Network Case

The UMTS Terrestrial Radio Access Network (UTRAN) has some salient architectural constraints: the AAL2 switches are located in radio network controllers

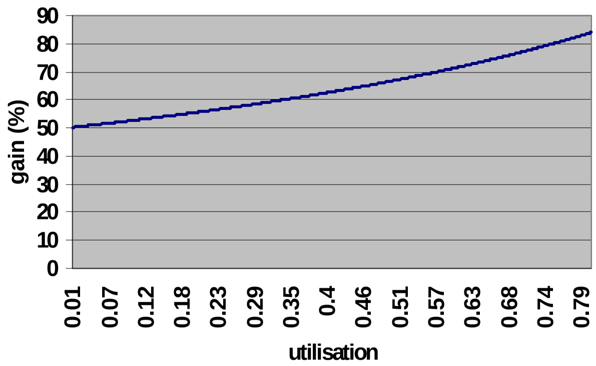

Figure 5: The Effect of Prioritisation on the Connection Setup Delay as the Processor Utilisation Increases

(RNC) and base station (BS) nodes. A BS is connected to its controlling RNC, and there is no direct link between base stations. Furthermore, AAL2 connections are always initiated by the diversity combining unit (SHO) unit and terminated in the BS. In theory, it is possible to establish AAL2 connections from an SHO located in a particular RNC, to any BS in the network when tracking a fast moving mobile. However, SHO–relocation is allowed, which limits the AAL2 connections to span a maximum of two RNCs. I constructed two AAL2 switching network models, which resemble typical UTRAN topologies. The first, "Tree", contains two RNCs controlling 10 base stations which are connected in a tree topology. In the second case, "Flat", the difference is that all the 10 base stations are directly connected to the controlling RNCs.

Unfortunately, these switching network cases can not be handled analytically. I have carried out simulation experiments for both network configurations, with two different node models. In the first case, the network is constructed from AAL2 nodes which apply FIFO queuing of the signalling messages. In the second case, the switching nodes give higher priority for the ERQ and ECF messages. The simulation results proved that the priority mechanism is capable of providing faster connection establishment. The gain ranges between 4–11% depending on the network configuration and the signalling processor utilisation. The better the connection setup performance the worse the connection release performance. However, as explained in the introduction, the most important design concern in a radio access network is the efficient use of the scarce radio resource, which is best supported by providing the fastest possible soft handoff leg establishment.

Thesis 2 : A Core–stateless End–to–end Measurement Based Call Admission Control Method for Supporting IP Telephony

One of the most actively studied research area of the past couple of years is the design of connection admission control methods for IP networks with different complexity, and the evaluation of the trade–off between the complexity of the method and the offered performance guarantees. Routers implementing the Expedited Forwarding per–hop behavior [21] offer a scalable way of providing a low delay and low jitter service for packets belonging to a particular traffic class, but only if the load in the class is regulated by some means. A lightweight, end–to–end measurement based call admission control method is proposed in this thesis, which is tailored to the needs of IP telephony gateways. It belongs to the recently proposed family of end–to–end measurement based call admission control methods [8, 22] but instead of probing the network on a per–call basis, it relies on measuring the quality experienced by ongoing live calls.

Thesis 2.1 : Call Admission Control for IP Telephony Based on Passive End–to–end Measurements [C3, P2]

I proposed a core–stateless, end–to–end measurement based call admission control method.

In the network depicted in Figure 6, telephony calls are travelling through the core network between pairs of IP Telephony Gateways. There are hundreds of simultaneous calls between any two IP Telephony Gateways. The base of the new call admission control method is that the gateways collect aggregate statistics (such as packet loss ratio) about all the ongoing calls, and regularly share this information with the peer gateway by sending a feedback message. When a new call arrives at the gateway, it compares the latest available statistics with the target value set for the telephony calls, and accepts the call if the performance indicators representing the quality of the transmission path towards the target gateway are above the preset threshold. The target values can be set in a way that the packet loss and delay suffered by the individual voice calls is below the limit, which can be compensated by the voice codec or the human brain[1].

The operation of the gateways can be followed in Figure 7. The "receiving gateway" is delivered an IP packet encapsulating a voice frame. Before sending the voice frame to the destination access network, the gateway determines which gateway have sent this packet (based on the source IP address in the packet header), and which call the frame belongs to (based on the call identifier included in the packet header). It checks the frame sequence number to calculate the number of packets that have been lost, and updates the packet loss counter corresponding to the sending gateway. It also checks the timestamp to calculate the transmission delay, and to update the delay counter. The delay counter is simply increased by the delay of the currently received packet calculated from the timestamp. Another variable, counting the number of packets received in the current measurement interval is also increased by one. At the end of each measurement period the receiving gateway sends a control packet to each sending gateway to inform them about the loss and delay statistics of their calls. The control packet carries the packet loss ratio and the average delay which is calculated by dividing the delay counter with the total number of packets received during the whole measurement interval.

Upon receiving a control packet from the remote gateway, the state variables "lossrate" and "actdelay" which store the loss and delay suffered by the voice packets of the gateway are updated. The received delay value is just stored while

[1]A state of the art voice codec can completely conceal 1% frame loss, and it is capable of providing acceptable speech quality up to 10% loss. The delay tolerance across the core network is around 150ms.

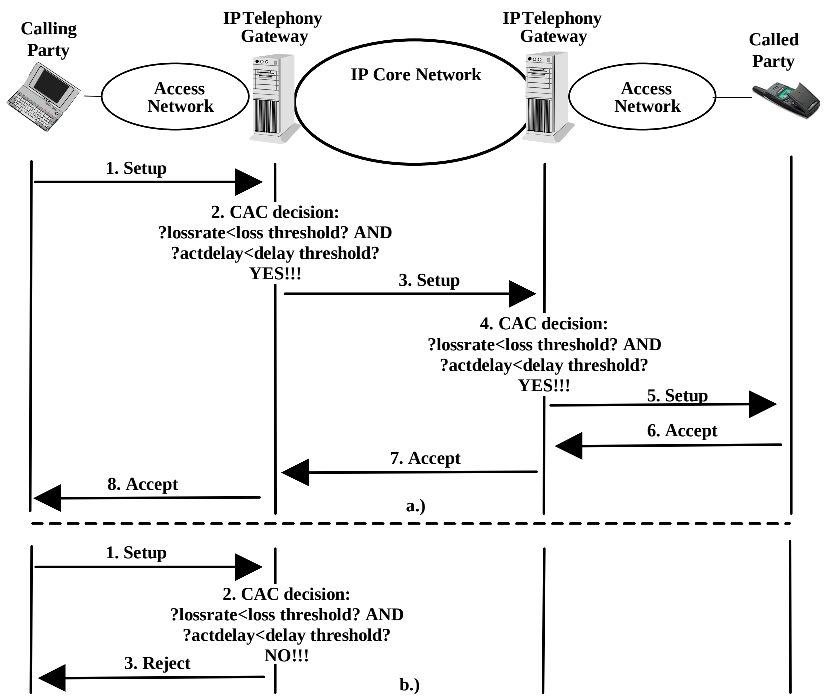

Figure 6: Sequence of actions for successful (a.) and unsuccessful (b.) connection establishment

the "lossrate" is updated by calculating an exponentially weighted moving average (EWMA) of the received measurement result, and the previous "lossrate" value.

Figure 6 depicts the sequence of actions taking place when an end user initiates a new voice call. Note that the message names appearing in the figure don't imply the use of any particular call control protocol. The end–to–end measurement based admission control method can work together with all existing call control signalling solutions. A connection setup message travels through the access network of the calling party, and arrives at the "sending" IP telephony gateway. The gateway derives the address of the "receiving" gateway from the address of the called party. It compares the actual value of loss and delay measurements towards the "receiving" gateway with the preset thresholds ("losstarget" & "delaytarget"). If the actual values are below the thresholds, the new call is accepted, and the connection establishment message is forwarded to the remote gateway (Step 3 of Fig 6.a). If any of the current values is above the respective threshold, the call is rejected (Step 3 of Fig. 6.b). When making the call admission control decision (Step 4), the "receiving" gateway uses the measurement reports submitted by the "sending" gateway.

As far as the start–up phase is concerned, when no information about the core of the network is available, a reasonable assumption is that a core network can handle thousands of simultaneous calls, so the first few calls can for sure be admitted without compromising the performance.

If the networks supports the Expedited Forwarding PHB [21], the low delay

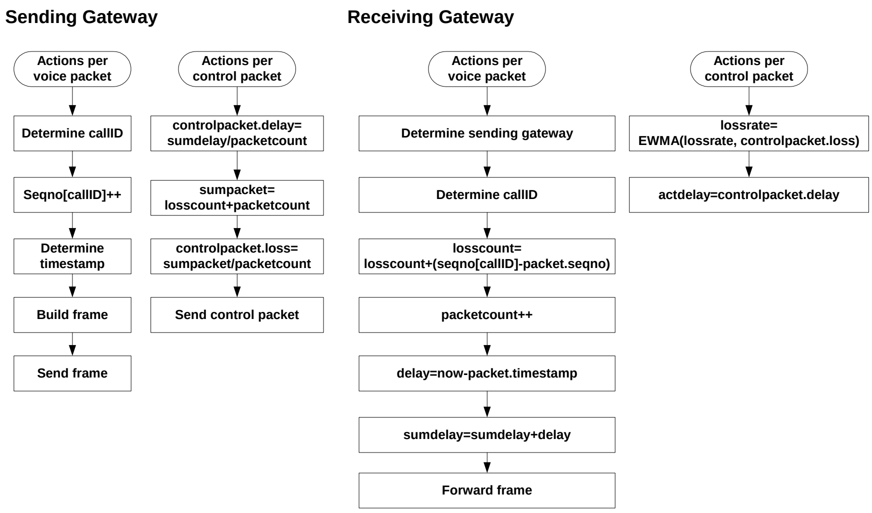

Figure 7: Flowchart of the Call Admission Control Algorithm

and jitter for voice traffic can be guaranteed by marking voice packets to receive preferential treatment. In this case, the required low delay can be provided by configuring the routers to assign small buffers for the traffic class, where the voice calls are carried which means that the CAC method can work utilising only loss measurements.

The method is further enhanced by allowing continuous adaptation of the admission control threshold. This feature is beneficial when the offered traffic in the network varies in a wide range. By detecting incipient congestion, the adaptive method aims for keeping the network utilisation close to a desired constant level in order to prevent serious congestion. Two methods were tested to achieve this goal. The main advantage of the methods is simplicity, the first one works for the case when the traffic follows a daily profile, the second one is more general:

1. The first method comes from the plain old telephony. If the network is used mainly for voice traffic, or there is a virtual private network type of arrangement for interconnecting the voice gateways, one can exploit the fact that it is very easy to characterise the daily profile of the phone traffic. Periods of modest network load are intercepted with periods of heavy traffic during the so–called busy hours. The gateways can be configured based on the daily traffic profile. A lower CAC threshold can be configured for the busy hours.

2. A more general solution is to continuously update the CAC threshold. Upon receipt of a new loss report the sending gateway compares the value to a second threshold. If the latest loss report is above the second threshold, the gateway selects a smaller CAC threshold. It uses this stricter CAC threshold until the average loss reported by the remote gateway decreases again to the desired value.

To protect the system against the potential negative effect of lost update packets, a timer is introduced to supervise the receipt of update packets. This timer is restarted upon receipt of a new update packet. If the timer expires, the admission control threshold is lowered, more precisely it is divided by a factor called *backoff*. This procedure ensures that the admission control is made stricter. If an update packet is finally received, the call admission control threshold is reset to the initial value.

I have validated the proposed solution using event driven simulation. The simulation results demonstrate that the method works. Further details about the performance evaluation can be found in Thesis 3.

Thesis 2.2 : A Distributed Measurement Architecture for Performing Call Admission Control Based on End–to-end Measurements

I proposed a measurement architecture which makes it possible using the passive end–to–end measurement based CAC not only in IP telephony gateways but also in VoIP terminals.

A new entity called Call Admission Control Broker is introduced which is responsible for collecting network measurement reports from terminals. The broker arranges these reports into a database and it is responsible for making the call admission control decisions upon a new call is initiated by the terminals.

If a VoIP terminal would like to start a new telephone call it sends a Call Request message to the Call Admission Control Broker. The message contains at least the IP address of the calling party and the called party. The broker makes a call admission control decision and the result is reported back to the terminal (Call Accept message). If the decision is positive the terminal can proceed with the voice session. During the voice sessions the terminals monitor the transmission quality. They collect loss and delay statistics. Right after the termination of the telephony call the terminal reports the measurement results to the Call Admission Control Broker in a Measurement Report message. This message contains the packet loss ratio and the average delay measured during the session and informs the broker also about the IP address of calling and called party.

The Call Admission Control Broker implements the clustering method described in [23]. It classifies all the IP addresses received in Call Request messages into clusters. It maintains the state variables (loss, delay) that are used by the call admission control method for each pair of clusters. Upon arrival of a new Measurement Report message about the call quality measured for a voice session between a particular pair of clusters, the state variables corresponding to the cluster pair are updated. If a Call Request message arrives to the Call Admission Control Broker, the broker checks the IP address of the called and calling party, and determines the clustering of both VoIP terminals. The network state variables corresponding to the cluster pair are used in the call admission control process. The outcome of the call admission control decision is reported back to the VoIP terminal.

Even if the IETF standard Session Initiation Protocol (SIP) [24] is used by the VoIP terminal for establishing the voice session, and the actual signalling takes place end–to–end without the involvement of any intermediate proxy, there is now a way for blocking the call if the required resources can not be set aside for ensuring outstanding voice quality. [25] discusses how network QoS and security establishment can be made a precondition to sessions initiated by SIP, and

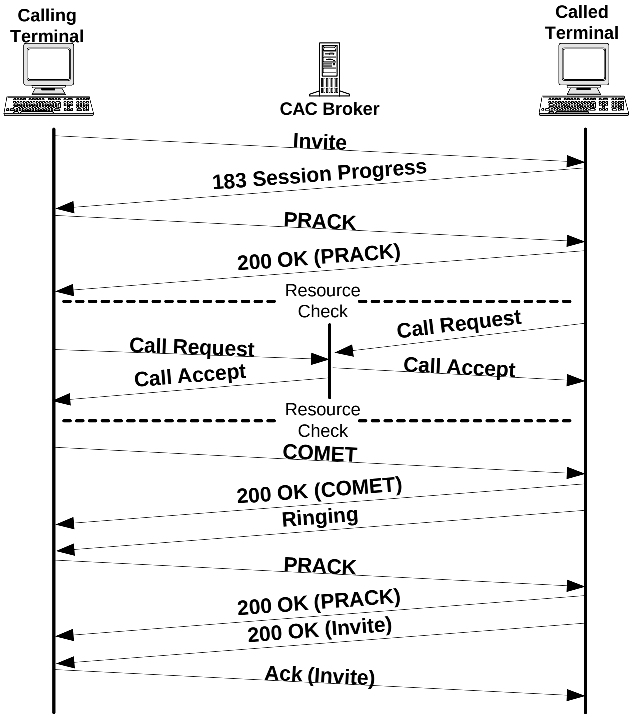

Figure 8: Integrating the Resource Checking with SIP Signalling

described by SDP. These preconditions require that the participant reserve network resources before continuing with the session. These preconditions require a participant to use an existing resource reservation mechanisms before beginning the session. This results in a multi–phase call–setup mechanism, with the resource management protocol interleaved between two phases of call signalling. Figure 8 depicts an example of integrating the proposed call admission control mechanism into SIP end–to–end session control signalling using the approach of [25]. This chart covers the case of successful session establishment.

Thesis 3 : Performance Evaluation of the End–to–end measurement Based Call Admission Control Solution

I have carried out a comprehensive performance evaluation using simulation. Two main networking scenarios were simulated. The first scenario assumes a Diffserv capable IP network where voice flows are carried in a separate traffic class. The second scenario assumes a pure best–effort IP network. In the second scenario, voice traffic is mixed with TCP controlled traffic, and it is not just the voice quality which needs to be evaluated but also the interaction between the two traffic types. That makes the second scenario a distinct case.

Thesis 3.1 : Performance Evaluation Assuming a Diffserv Capable IP Network [C3]

I have validated the benefits of the proposed call admission control

solution by analysing the results of an extensive set of simulations which covered all relevant traffic types and network scenarios.

The following cases were simulated:

- Two —widely accepted— voice models were used.

  CBR produces fixed length packets at a constant rate.

  ON/OFF produces also fixed length packets at a constant rate, but only in ON periods, while in OFF periods no packet is sent. The distribution of the length of ON and OFF periods is exponential.

- Background load was generated using three types of source models: Pareto ON/OFF sources, admission controlled voice traffic, unregulated voice traffic.

- Many network topologies were evaluated.

  - One pair of gateways connected by a single bottleneck link.
  - Two gateway pairs sharing a bottleneck link.
  - Multiple gateway pairs operating in a 15 node network with multiple bottleneck links.

- Simulations were run with modest, high, and dynamically changing offered traffic.

- The Call Admission Threshold, the size of the bottleneck buffer, the frequency of feedback, and the capacity of the bottleneck link was also varied.

I introduced a new measure for characterising the performance of the proposed method. Instead of using the so–called "loss–load" curve [26], which shows the average loss suffered by the flows subjected to the admission control at a certain load level, I collect per–flow statistics, and I calculate how many percent of the individual voice flows exceed a certain loss threshold. This is a better measure, because state of the art voice codecs can completely conceal quite high loss rates (typically up to 1%) and the user–perceived quality is better represented by the per–flow performance.

Based on the simulation results, the proposed method can be characterised as follows:

1. It is a viable option for performing call admission control in IP telephony systems, because it is capable of ensuring very low (below 1%) per–flow packet loss ratio with acceptable level of link utilisation (around 70%).

2. Increasing the buffer capacity results in more and more calls being admitted into the network, and those admitted calls experience a better loss performance as well.[2]

3. The optimal Call Admission Threshold depends on the call intensity, smaller call inter–arrival time requires stricter threshold to yield the same performance.[3] This problem can be handled by using the adaptive CAC thresholds.

[2]The blocking probability decreased by 12% and the fraction of calls suffering more than 1% loss decreased by 14% in a particular simulation setup.

[3]A loss threshold of 0.0005 instead of 0.1 is required in a particular simulation setup.

4. If two gateway pair uses the same CAC Threshold, they experience the same performance, and each gateway receives its fair share of the bandwidth of the bottleneck link. However, in the asymmetric case, the gateway with the higher CAC threshold monopolizes the link, and gets almost all of its calls through, while the per–call performance in the two gateways is indistinguishable.

5. Simulations with dynamically changing load reveal that the CAC method reacts to the increased level of congestion as the increase of the blocking probability indicates. The reaction is however not aggressive enough, since the per–call performance degrades.[4]

6. Experiments with multiple bottleneck links show that the CAC method negatively discriminates the gateway, which sends its traffic via a longer path.[5]

As presented in the previous list(3.), the simplest version of the end–to–end measurement based CAC approach is not stable enough. It can handle wide range of offered traffic but the optimal CAC threshold is different for high and low traffic load, especially if one is concerned with not only the communication quality but also the network utilisation. Simulations reveal that the measured loss ratio shall be kept around one order of magnitude below the loss target. The adaptation methods presented in Thesis 2.1 have been simulated.

The busy hour based method nicely smoothes the network load, but it is applicable only if the network operator can determine the periods of heavy network load in advance.

The continuous adaptation of the CAC threshold leads to ∼4.4% more blocking, and ∼6% decrease in utilisation over the bottleneck link for the normal load period, but the performance during high load periods improves significantly. (The fraction of calls suffering more then 1% loss decreases by ∼20%.)

Thesis 3.2 : On the Interaction of End–to–end Measurement Based Call Admission Control with TCP Traffic [J1, C4]

I have characterized the voice quality, the throughput achievable by TCP traffic, fairness of the resource sharing between the traffic types, and the effect of congestion on the feedback path if flow based admission control is used simultaneously with in-flow congestion control in an IP network.

I used simulation to study the interaction between traffic flows subjected to end–to–end measurement based admission control and TCP flows. The following cases are simulated:

- TCP Reno version is used because it is the most widely deployed today.
- Network topologies with single and multiple bottlenecks are simulated.
- Varying number of TCP flows and varying bottleneck link capacities are tested.

[4]The blocking probability is increased by ∼25%, and the fraction of calls experiencing more then 1% loss increased by ∼26%.

[5]Calls flowing through three bottleneck links experience ∼3–times more blocking compared to the calls running through only a single bottleneck.

- Many different configurations of the admission control method are tested.
- The effect of RED vs. FIFO buffer management is assessed.

Two questions need to be answered:

1. How does the admission controlled voice traffic effect the throughput of simultaneous TCP sessions?
2. To what extent it is possible to ensure acceptable voice quality when there is competing TCP traffic present?

Increasing the number of TCP flows inevitably results in complete expelling of the admission controlled traffic, simply because TCP is not that concerned with the packet loss ratio, and can work nicely with much higher packet loss ratio then what can be tolerated by the voice traffic. The proposed flow level, end–to–end measurement based call admission method however has a very definite benefit:

- It blocks all those voice calls, which would anyway be useless because of the very high packet loss ratio, and in this way it ensures much higher throughput for TCP traffic.

In the range, where the number of TCP flows present in the network is modest, that is the resulting packet loss ratio is not more then what can be tolerated by the voice traffic, the method is beneficial for two reasons:

- It protects TCP traffic from non–responsive flows, and results in an approximately fair share[6] of the bandwidth between the two traffic classes.
- It ensures the required (<10%) packet loss ratio for the voice flows.

The method can ensure that the delay bound of the voice packets is not exceeded. Taking however delay statistics also into account makes the call admission control method even more conservative[7] when fighting for bandwidth against TCP flows.

The receipt of regular feedback messages is crucial for the proposed CAC method to work: if update packets are lost, the gateway can not update its estimate of current conditions in the network.

I experimented with the backoff mechanism proposed in Subthesis 2.1. The backoff mechanism improves the loss performance in each simulation setting[8]. In return for good results regarding packet loss, the link utilization is usually lower[9] with the backoff technique, but in most cases providing QoS is more important than utilisation especially when it comes to supporting a telephony–like service.

[6]Fair share means that the two traffic classes get equal fraction of the bottleneck bandwidth. The measured bandwidth share of the TCP traffic is between 40% and 60% of the bottleneck bandwidth in a large operating region.

[7]24 TCP source is enough to monopolise a 8 Mb/s bottleneck link when the delay threshold is 50ms, while 32 flows can only grab ∼77% of the bottleneck bandwidth when the delay threshold is set to 80 ms.

[8]The improvement is most noticeable when the update loss ratio is high (0.45). The fraction of calls suffering more then 1% packet loss is halved by setting the backoff value to 4.

[9]The decrease is 1%–5% depending on the loss rate and the backoff parameter.

5 Application of New Results

I contributed with the details of the AAL2 Signalling protocol to Study Group 11 of ITU–T [S1]–[S24]. My proposals have been incorporated into new ITU–T Recommendation Q.2630.1 [27] and Q.2150.1 [28]. Q.2630.1 has later been standardised by 3GPP [29] to be used to control AAL2 connections in UMTS Terrestrial Radio Access Networks. [J2] is cited in five independent publications [Cit1]–[Cit5].

The Internet is on its way to become a universal service platform offering real–time multimedia services as well as simple file transfer, web browsing and e–mail. However, for the time being most of the routers deployed in the network are capable of providing simple best effort packet forwarding, and it will take significant time and money to upgrade these routers to support sophisticated QoS differentiation. There is a need for a sound migration path. In this transition phase there is a definite window of opportunity for solutions which offer some sort of QoS guarantees, therefore represent a step forward from the best effort service paradigm, but operate without requiring any involvement from core routers. The method I propose in Thesis 2.1 can be realized by upgrading only the IP telephony gateways, and can still provide reasonable QoS for voice sessions under wide range of network conditions.

6 A Quick Reference Guide to the Theses

Thesis 1 : A Signalling Protocol for Supporting Switched AAL type 2 Connections in UMTS Terrestrial Radio Access Networks

Thesis 1.1 : AAL2 Signalling Protocol [C1, J2, J3]

It is discussed in Section 2.2 of the dissertation.

Thesis 1.2 : Bearer Independent Signalling Protocol Architecture [P1, C1]

It is discussed in Section 2.3 of the dissertation.

Thesis 1.3 : Optimising the Performance of AAL2 Signalling [C2]

It is discussed in Section 2.4 of the dissertation.

Thesis 2 : A Core–stateless End–to–end Measurement Based Call Admission Control Method for Supporting IP Telephony

Thesis 2.1 : Call Admission Control for IP Telephony Based on Passive End–to–end Measurements [C3, P2]

It is discussed in Section 3.2 of the dissertation.

Thesis 2.2 : A Distributed Measurement Architecture for Performing Call Admission Control Based on End–to-end Measurements

It is discussed in Section 3.4.2 of the dissertation.

Thesis 3 : Performance Evaluation of the End–to–end measurement Based Call Admission Control Solution

Thesis 3.1 : Performance Evaluation Assuming a Diffserv Capable IP Network [C3]

It is discussed in Section 3.3.3 of the dissertation.

Thesis 3.2 : On the Interaction of End–to–end Measurement Based Call Admission Control with TCP Traffic [J1, C4]

It is discussed in Section 3.3.4 and Section 3.3.5 of the dissertation.

References

[1] *Introduction to 3G Mobile Communications.* Artech House, Norwood, USA, 2001.

[2] *Third Generation Mobile Systems, http://www.ericsson.com/3G.*

[3] *Wideband CDMA for Third Generation Mobile Communications.* Artech House, Norwood, USA, 1998.

[4] ITU–T Recommendation I.363.2. *B-ISDN ATM Adatpation Layer Type 2*, 1997.

[5] IETF RFC 2210. *The Use of RSVP with IETF Integrated Services*, 1997.

[6] IETF RFC 2475. *An Architecture for Differentiated Services*, 1998.

[7] Internet draft, Work in progress. *Load Control of Real Time Traffic*, 1999.

[8] Viktória Elek, Gunnar Karlsson, and Robert Rönngren. Admission control based on end–to–end measurements. In *IEEE Infocom*, Tel-Aviv, Israel, April 2000.

[9] Ion Stoica and Hui Zhang. Providing guaranteed services without per flow management. In *ACM Sigcomm*, Cambridge, USA, August 1999.

[10] Kevin Thompson, Gregory J. Miller, and Rick Wilder. Wide-area internet traffic patterns and characteristics. *IEEE Network*, 11(6), November/December 1997.

[11] Art Mena and John Heidermann. An empirical study of real audio traffic. In *IEEE Infocom*, Tel-Aviv, Israel, April 2000.

[12] Sally Floyd and Kevin Fall. Promoting the use of end–to–end congestion control in the internet. *IEEE/ACM Transactions on Networking*, August 1999.

[13] Jerome H. Saltzer, David P. Reed, and David D. Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems*, 2(4), November 1984.

[14] *UCB/LBNL/VINT Network Simulator ns (version 2) can be found at: http://www-mash.cs.berkeley.edu/ns/ns.html.*

[15] Göran Eneroth and Martin Johnsson. ATM transport in cellular networks. In *International Switching Symposium (ISS'97)*, Toronto, Canada, September 1997.

[16] Jonas Reinius. Cello — an ATM transport and control platform. *Ericsson Review*, (2), 1999.

[17] *Trillium's AAL2 Solutions, http://www.trillium.com/products/broadband/aal2/solutions.*

[18] *Extreme AAL 2 Emulation Option, http://adtech.spirentcom.com/datasheets/T4159.asp.*

[19] András Valkó, András Rácz, and Gábor Fodor. Voice QoS in 3rd generation mobile systems. *IEEE Journal on Selected Areas in Communications*, 17(1), January 1999.

[20] *Queuing Systems.* John Wiley and Sons, New York, USA, 1976.

[21] IETF RFC 2598. *An Expedited Forwarding PHB*, 1999.

[22] Giuseppe Bianchi, Antonio Capone, and Chiara Petrioli. Throughput analysis of end–to–end measurement–based admission control in IP. In *IEEE Infocom*, Israel, April 2000.

[23] Balachander Krishnamurthy and Jia Wang. On network–aware clustering of web clients. In *ACM Sigcomm*, Stockholm, Sweden, August 2000.

[24] IETF RFC 2543. *SIP: Session Initiation Protocol*, 1999.

[25] Internet draft, Work in progress. *Integration of Resource Management and SIP*, 2001.

[26] Lee Breslau, Sugih Jamin, and Scott Shenker. Comments on the performance of measurement–based admission control algorithms. In *IEEE Infocom*, Tel-Aviv, Israel, April 2000.

[27] ITU–T Recommendation Q.2630.1. *AAL type 2 signalling protocol (Capability Set 1)*, 1999.

[28] ITU–T Recommendation Q.2150.1. *Signalling Transport Converter on MTP3 and MTP3b*, 2001.

[29] 3GPP Technical Specification TR 25.426. *UTRAN Iur and Iub Interface Data Transport & Transport Signalling for DCH Data Streams*, 1999.

Publications

Journal Papers

[J1] I. Szabó, *On Call Admission Control for IP Telephony in Best Effort Networks*, Computer Communications, Vol. 26, No. 4, pp. 304-313, March 2003.

[J2] G. Eneroth, G. Fodor, G. Leijonhufvud, A. Rácz, I. Szabó, *Applying ATM/AAL2 as a Switching Technology for 3^{rd} Generation Mobile Networks*, IEEE Communications Magazine, Vol. 37, No. 6, pp. 112-122, June 1999.

[J3] A. Rácz, I. Szabó, *Az ATM/AAL2 alkalmazása UMTS hálózatokban*, Magyar Távközlés, pp. 9-14, 2000 február.

[J4] I. Cselényi, R. Szabó, I. Szabó, N. Björkman, A. Latour-Henner, C. Gisgård, *Experimental Platform for Telecommunication Resource Management*, Computer Communications, Vol. 21 pp. 1624-1640, November 1998.

[J5] R. Ozvald, I. Szabó, *Jelölő algoritmusok teljesítményvizsgálata differenciált szolgáltatást nyújtó IP–hálózatokban*, Híradástechnika, pp. 37-40, 2001 április.

Conference Papers

[C1] I. Szabó, *A Flexible Signalling Protocol for Supporting Switched AAL type 2 Connections in UMTS Terrestrial Radio Access Networks*, IEEE International Conference on ATM, Colmar, France, June 1999.

[C2] I. Szabó, S. Székely, I. Moldován, *Performance Optimisation of AAL2 Signalling for Supporting Soft Handoffs in UMTS Terrestrial Radio Access Networks*, IEEE International Symposium on Computers and Communications, Antibes, France, July 2000.

[C3] I. Szabó, *Performance Evaluation of a New End-to-end Measurement Based Call Admission Control Scheme for Supporting IP Telephony*, SCS, International Symposium on Performance Evaluation of Computer and Telecommunication Systems, Orlando, Florida, July 2001.

[C4] I. Szabó, *A Call Admission Control Method for Supporting Telephony Sessions in a Best Effort IP Network*, COST Workshop on Quality of Future Internet Services, Coimbra, Portugal, September 2001.

[C5] I. Cselényi, C. Gisgård, P. Haraszti, A. Latour-Henner, I. Szabó, G. Záruba, *Multipurpose Middleware for Broadband Multimedia Applications*, ConTEL Broadband & Multimedia Workshop, Zagreb, Croatia, November 1996.

[C6] I. Cselényi, I. Szabó, P. Haraszti, N. Björkman, C. Gisgård, *A Versatile Multimedia Platform*, IASTED International Conference on Applied Informatics, Innsbruck, Austria, February 1997.

[C7] B. Benkovics, I. Szabó, *Provisioning QoS in 3G Networks With RSVP Proxy*, International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, September 2001.

[C8] A. Jüttner, I. Szabó, Á. Szentesi, *On Bandwidth Efficiency of the Hose Resource Management Model in Virtual Private Networks*, accepted for publication, IEEE Infocom 2003.

Patents

[P1] I. Szabó, G. Eneroth, *Bearer Independent Signalling Protocol*, US Patent and Trademark Office patent application No. 09/064,830, April 1998.

[P2] I. Szabó, *A Lightweight Core–stateless Call Admission Control Method for IP Telephony Gateways*, European Patent Office patent application No. 00850110.8, June 2000.

[P3] I. Szabó, *Supporting QoS Differentiation in RTP Multiplexing Scenarios*, European Patent Office patent application No. 00850111.6, June 2000.

[P4] I. Szabó, *A Method for Improving the Bandwidth Efficiency of RTP Multiplexing while Maintaining Tolerable Delay*, European Patent Office patent application No. 00850112.4, June 2000.

[P5] S. Rácz, Sz. Malomsoky, I. Szabó, *Extending the Content of the Link Characteristics Parameter in AAL2 Signalling to Enable the Use of Much More Reliable Connection Admission Control Methods*, US Patent and Trademark Office patent application No. 10/143,826, May 2002.

Contributions to International Standardisation

[S1] *AAL2 Link Identifier*, Delayed Contribution D.966, ITU-T Study Group 11 Meeting, Geneva, Switzerland, May 1998.

[S2] *Addressing in AAL2 Networks*, Delayed Contribution D.971, ITU-T Study Group 11 Meeting, Geneva, Switzerland, May 1998.

[S3] *Addressing of AAL2 Nodes in a Switched ATM Network*, Delayed Contribution D.972, ITU-T Study Group 11 Meeting, Geneva, Switzerland, May 1998.

[S4] *Making the AAL2 Signaling Protocol Independent of Underlying Signaling Bearer*, Delayed Contribution D.973, ITU-T Study Group 11 Meeting, Geneva, Switzerland, May 1998.

[S5] *AAL2 Signaling Protocol Architecture Reference Figure*, HEL.29, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S6] *AAL type 2 Addressing Considerations*, HEL.34, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S7] *AAL2 CID Contention Handling*, HEL.35, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S8] *Generic Primitives Provided for the AAL type 2 Signaling Protocol*, HEL.37, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S9] *AAL type 2 Signaling Message Flows for Basic Connection Handling*, HEL.39, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S10] *AAL type 2 Path Identifier*, HEL.40, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Helsinki, Finland, August 1998.

[S11] *Coding of SSCS Specific Information in Q.AAL2*, Delayed Contribution D.1103, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S12] *Blocking Procedure*, Delayed Contribution D.1105, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S13] *Basic Connection Control*, Delayed Contribution D.1109, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S14] *Proposed Modifications to Appendix 1 in Draft Q.AAL2*, Delayed Contribution D.1112, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S15] *Service Provided by the AAL Type 2 Signaling Entity*, Delayed Contribution D.1113, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S16] *Timers & Cause Codes in Q.AAL2*, Delayed Contribution D.1114, ITU-T Study Group 11 Working Party 1 Meeting, Ostend, Belgium, November 1998.

[S17] *Reuse of Cause Values Defined in Q.2610*, BER.20, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Bern, Switzerland, January 1999.

[S18] *Layer Management Interface Definition and Related Corrections*, BER.25, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Bern, Switzerland, January 1999.

[S19] *State Transition Diagrams for the Maintenance Protocol Procedures*, BER.26, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Bern, Switzerland, January 1999.

[S20] *A Correction in the Normal Connection Release Procedure*, BER.29, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Bern, Switzerland, January 1999.

[S21] *Draft New ITU-T Recommendation Q.sbcmtp AAL type 2 Signalling Bearer Converter on Broadband MTP*, BER.34, ITU-T Study Group 11 AAL2 Special Rapporteur's Meeting, Bern, Switzerland, January 1999.

[S22] *State Transition and SDL Diagrams for Signalling Bearer Converter on Broadband MTP*, Delayed Contribution D.1334, ITU-T Study Group 11 Meeting, Geneva, Switzerland, March 1999.

[S23] *Aligning the Cause Values Used in Q.AAL2 width Q.2610 Based on the Agreement Made in Bern*, Delayed Contribution D.1239, ITU-T Study Group 11 Meeting, Geneva, Switzerland, March 1999.

[S24] *Modifications to the GIT Recommendation to Support Identification of Q.AAL2 Users*, Delayed Contribution D.1240, ITU-T Study Group 11 Meeting, Geneva, Switzerland, March 1999.

[S25] *Delay Performance of AAL2 over a 2 Mb/s I_{ub} Link*, 3GPP TSG-RAN Working Group 3 meeting, Berlin, Germany, August 2000.

Independent Citations

[J2] is cited in the following publications:

[Cit1] S. Nananukul, Y. Guo, M. Holma, S. Kekki, *Some Issues in Performance and Design of the ATM/AAL2 Transport in the UTRAN*, IEEE Wireless Communications and Networking Conference, Chicago, USA, September 2000.

[Cit2] A-F. Canton, S. Tohme, D. Zeghlache, T. Chahed, *Performance Analysis of AAL2/ATM in UMTS Radio Access Network*, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Lisboa, Portugal, September 2002.

[Cit3] A. B. Garcia, M. Alvarez-Campana, E. Vázquez, J. Berrocal, *"Quality of Service Support in the UMTS Terrestrial Radio Access Network*, 9^{th} Workshop of the HP OpenView University Association, June 2002.

[Cit4] S. Dixit, Y. Guo, Z. Antoniou, *Resource Management and Quality of Service in Third-Generation Wireless Networks*, IEEE Communications Magazine, Vol. 39, No. 2, pp. 125-133, February 2001.

[Cit5] J. F. Huber, D. Weiler, H. Brand, *UMTS, the Mobile Multimedia Vision for IMT-2000: A Focus on Standardization*, IEEE Communications Magazine, Vol. 38, No. 9, pp. 129-136, September 2000.