Architecture and Configuration of Broadband Access Networks Supporting Multimedia Applications

András Kern, Mátyás Martinecz, Zalán Heszberger and Gyula Sallai
Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics
2. Magyar tudósok körútja Budapest, Hungary, H-1117
{kern, martinecz, heszberger, sallai}@tmit.bme.hu

Abstract

Recent Internet access speeds have made it possible even for home users to reach and use previously unavailable bandwidth consuming (e.g. Video on Demand) services. However, the rapid spread of these applications is hindered by the lack of QoS guarantees in packet networks. In this article a QoS architecture is presented which have been designed for access networks and applies a logical overlay network to make network management tasks easier. The novelty of the architecture lies in the application of two key procedures adopted for dimensioning logical channels between edge nodes and mapping them to the physical routes. The upper load thresholds of the applied logical channels are also determined, which may form the basis of a more sophisticated call admission control policy.

1 Introduction

Recent Internet access speeds have made it possible even for home users to reach and use previously unavailable value-added, bandwidth consuming (e.g. Video on Demand) services. The rising demand for these applications urge the solution of QoS provisioning in packet networks, as without guaranteed service quality level the rapid spread of these novel applications is unlikely. To solve this problem network management mechanisms should be implemented into the network.

The first step towards the aim of QoS provision for Internet services should be the provision of QoS guarantees in the access network (AN). Therefore a robust access network should be developed implementing an efficient QoS architecture capable of guaranteeing a proper level of transmission quality for value-added services.

A framework called Enhanced Resource and Information Control (ENRICO) has been developed to solve this problem [1]. According to this concept a central bandwidth broker manages the access network’s resources by making call admission control (CAC) decisions and bandwidth allocations. To make network management tasks easier a simpler logical overlay network is defined onto the physical topology.

In our research work the ENRICO model was regarded as a starting point, but it has been altered in several aspects. Instead of the centralized CAC entity a distributed traffic load control scheme is recommended which is realized by the edge nodes of the AN. Furthermore, new concepts have been introduced in the implementation of the logical overlay network. The main contribution of the paper is built around these points.

The realization of the logical overlay network in the AN raises two important questions: special algorithms are needed to the dimensioning of the logical channels and also methods are to be developed to map these channels to the physical routes. The importance of determining the maximum capacity of each logical channel is also emphasized, as this value can form the basis of a more sophisticated CAC policy. Such procedure is also suggested in the paper.

The rest of this paper is organized as follows. In the following section the QoS architecture to be applied in the AN is introduced in details. The third section contains the description of a method that may be used to determine the required capacity of the logical channels. In the fourth section the technique that maps logical channels onto physical routes and determines maximal capacity of those is presented. The fifth section contains an illustrative, numerical example which aids the understanding of the benefits and operation of the presented methods. Our concluding remarks can be found in the last section.
2 ENRICO Based Model

In general, network traffic belonging to the session of a value-added application may travel along three different networks until reaching the destination located at the subscriber’s home network: For first, as a typical source, the application service provider (ASP) network, that possibly have its own storage area subnetwork consisting of multiple interconnected servers. For second, the network of the internet service provider (ISP), which typically consists of routers, AAA server(s), mail server(s), etc. Accordingly the main tasks of the ISPs are to offer connectivity for ASPs to the AN, to authorize and to authenticate access. Finally, the AN, which consists of the local loop and the aggregation network. However, this paper focuses only on a segment of the aggregation network, the boundaries of which are defined by the DSLAMs (DSL Access Multiplexers) and the BASs (Broadband Access Servers).

The QoS architecture introduced in this article is based on the ENRICO concept [1] and has been designed for QoS provisioning primarily in DSL access networks. In this concept an interface is provided between the NAP and the ASP through which QoS- and resource reservation-related information may be exchanged in a manner that is fully transparent to subscribers. Load control functions (call admission control, traffic policing) are realized by the edge nodes of the AN in a distributed manner. Resource reservation operations are carried out however by a central configurator entity, which defines a logical overlay network onto the physical one in advance. This predefined overlay network consists of logical point-to-point trunks between the edge nodes of the AN. For gathering the needed information for the operation of the QoS architecture certain functions of currently available protocols such as the RSVP or the DHCP are also exploited.

To make the whole concept clearer let us see its operation step by step (see Figure 1). The following example depicts a value-added service provisioning through the DSL-based AN applying the above detailed architecture:

Step 0. The central configurator entity defines the logical overlay network, that is it configures a logical channel between each DSLAM and BAS per QoS class. These QoS pipes can be realized either by PVPs (Permanent Virtual Paths) in case the L2 technology of the aggregation is ATM, or by exploiting the ELS (Ethernet Line Service) in case L2 technology is Ethernet. For more details see [1].

Step 1. A new subscriber would like to connect to the Internet, therefore it requests an IP address from its ISP using the DHCP. This DHCP communication is monitored by the BAS through which the subscriber can reach its ISP and important pieces of information (subscriber's IP address, the DSLAM to which the subscriber’s DSL is connected) are stored by the BAS.

Step 2. The subscriber initiates a value-added service (e.g. VoD) through the web page of the ASP by simply clicking on the appropriate hyperlink.

Step 3. The ASP registers the request and sends an RSVP packet to the subscriber. This RSVP message describes the desired QoS level of the service and also contains the peak transmission rate of the traffic generated by the initiated session.

Step 4. The BAS captures the RSVP packet and according to the encapsulated information checks whether there is enough capacity available in the appropriate QoS pipe to fulfill the QoS requirement of the newcomer and already admitted sessions at the same time.

Step 5.a If there is enough available capacity the newcomer session is admitted: it is registered by the BAS and the policers of the BAS and the subscriber’s DSLAM are set to the transmission rate of the traffic generated by the initiated session as well.

Step 5.b In case the predefined QoS pipe has not enough available capacity for the newcomer session then the capacity of the pipe should be expanded if possible. As the maximum sizes of the logical channels are previously determined by the configurator entity this question can be easily answered. In case the currently allocated capacity for the QoS pipe is less than its maximum size it should be expanded to a higher value so that it could cope with the new session as well.

Step 5.c In case the capacity of the appropriate QoS pipe is not enough to cope with the currently active plus the newcomer session and the pipe cannot be expanded the service request must be rejected and the ASP’s server should be informed about this decision.

Figure 1. The model of the recommended QoS architecture

Obviously the predefined logical overlay network plays a central role in this architecture. For its realization two main questions have to be answered. The first one deals with the dimensioning of the logical channels, that is how much capacity should be allocated for each QoS pipes provided...
that an anticipated traffic load and a target QoS level per QoS class is defined. The other question is how to map these logical channels onto physical routes.

Besides these two questions it is also worth investigating what the maximum capacities of the logical channels are. This result can be useful for the network operator in two fields. First, in the already mentioned case when the traffic load gets higher than anticipated, it provides information about how many spare capacities could be allocated to expand the predefined logical channels so that they could cope with the extra traffic. Second, if the AN operator would like to expand its subscriber population without extending the currently available physical capacities, then from the maximum logical channel capacities the maximum number of subscribers at each DSLAM could be deduced.

In the next two sections we recommend methods that can be used for realizing the above mentioned three functions of the central configurator entity.

3 Logical Channel Dimensioning

One of the most important questions arising in connection with the logical overlay network deals with the required capacities of the logical channels. To answer this question a method will be presented that bases only minimal given parameters: it is supposed that only the number of subscribers per DSLAM, their rate of service initiation, the mean holding time of the sessions and the peak and mean transmission rates of the service are known.

Let us suppose that the AN contains  k DSLAMs and the number of subscribers connected to DSLAMi, denoted by Ni, i = 1...k is known. Let us also assume that the subscribers request a certain type of service independently from each other and with probability \( p \). Thus service requests arriving at DSLAMi can be modelled with a Poisson process of rate \( \lambda = N_i \ast p \). The mean value of the holding time of the service is denoted by \( h \). Using Erlang’s B formula the minimal number of channels that is required to stay below a target blocking probability \( \varepsilon \) can be computed as follows:

\[
c_{i,min} = \inf \left\{ c : E(\rho_i, c) = \frac{\rho_i^c}{\sum_{j=0}^c \rho_i^j / j!} \leq \varepsilon \right\}
\]

where \( \rho_i = \lambda_i \ast h \) is the offered traffic at DSLAMi, traditionally measured in Erlangs.

The question is now that how much capacity is needed to transmit \( c_{i,min} \) simultaneous sessions while fulfilling their QoS requirements composed in terms of delay, delay variation and information loss. To answer this question a simple modelling approach may be applied with the aid of which quantitative guarantees can be provided for the PLR (Packet Loss Ratio, e.g. the ratio of lost and sent packets) while keeping delay and jitter at as low level as possible. In the Bufferless Fluid Flow Multiplexing (bffm) framework multiplexed traffic sources are regarded as an aggregated fluid flow flowing through a bufferless link with capacity \( C \). This means that information loss only occurs when the instantaneous arrival rate of the aggregated flow exceeds \( C \). This approach does not describe real life perfectly, as real packet schedulers start serving a packet only after it has totally arrived and only one packet is served at a time, however using this framework a conservative bound on the expected PLR can be determined and only minimal buffer sizes are needed to avoid packet drops occurring at simultaneous packet arrivals.

In the applied model \( n \) fluid flows are aggregated onto a communication link (i.e. a QoS pipe) with capacity \( C \). Let \( X_i \) (\( i = 1...n \)) be random variables denoting the instantaneous arrival rate of the \( i \)-th stationary flow. It is assumed that the composition (i.e. ratio of certain traffic source types) of the anticipated aggregated traffic is known however only limited amount of data is available about each type of traffic sources: they are characterized by their \( p_i \) peak and \( m_i \) mean arrival rate. Let \( X \) random variable denote the arrival rate of the aggregated traffic \( X = \sum_{i=1}^n X_i \). Thus the value of packet loss ratio can be expressed: \( PLR \overset{\text{def}}{=} \frac{E[(X-C)^+]}{E[X]} \), where \( E[.] \) stands for the expectation value operator and \( (X-C)^+ = \max(X-C,0) \).

The goal is to determine the equivalent capacity of \( X \), which is by definition:

\[
C_{equ,PLR} \overset{\text{def}}{=} \inf \{ C : WLR \leq e^{-\gamma} \}.
\]

It can be seen that for the evaluation of (2) the expected PLR should be determined somehow. As only two parameters of the traffic sources are known it can not be calculated exactly however it can be bounded from above using the well-known Chernoff inequality [2, 3]:

\[
PLR \leq \inf_{s>0} \exp \left( \Lambda_X(s) - sC - \log(sM) \right).
\]

where \( \Lambda_X(s) \overset{\text{def}}{=} \log E[\exp(sX)] \) is the logarithm of the moment generating function (MGF) of \( X \).

The MGF of \( X \) can not be determined exactly, as only the mean and peak rate of a traffic source are known, however if the original traffic sources are substituted with their ON/OFF equivalents and the possibly inhomogeneous sources are homogenized a simple upper bound for the MGF of the original aggregated flow’s rate distribution function can be constructed [4]:

**Theorem 1** Let \( X_1, \ldots, X_n \) indicate \( n \) independent random variables with \( 0 \leq X_i \leq p_i \), \( X = \sum_{i=1}^n X_i \), and
\[ M = \mathbb{E}[X]. \] Let \( Y_{i}^{\text{on/off}} \) be independent homogeneous on-off sources with identical peak rate \( p = \max(p_i, i = 1, \ldots, n) \), 
\[ n_Y = \left| \sum_{i=1}^{n} p_i / p \right|, \] and identical mean rate \( m = \sum_{i=1}^{n} m_i / n_Y. \) Then for \( s > 0 \),
\[ \mathbb{E}[\exp(sX)] \leq \left( 1 - \frac{M}{n_Yp} + \frac{M}{n_Yp} e^{sp} \right)^{n_Y}. \] (4)

For proof of Theorem 1 see [5]. Using (3) and (4) the following direct equivalent capacity bound can be constructed [6]:

\[ \hat{C}_{\text{equ,PLR}} = \inf_{s>0} \frac{\hat{X}_{X,\text{so}}(s) + \gamma - \log(sM)}{s}, \] (5)

where \( \hat{X}_{X,\text{so}}(s) \) is the logarithm of the MGF bound presented in (4). It is worth noting here that this expression contains only one optimization, which can be performed rapidly.

4 Network Configuration Task

The main goal of this task is to allocate capacities and determine the maximal sizes for the logical pipes in the physical networks. Since a central entity performs this operation, the global view of the network state is known.

A possible solution is to tune sharp the sizes of the pipes applying an iterative method [7]: the sizes of the pipes are decreased if the configurator cannot solve the problem and increased if it can and free capacities remain in the network.

On the contrary, in the proposed approach, these two tasks (path allocation and pipe sizing) are handled at the same time. The goal is to maximize the throughput of the system instead of minimize the allocated network capacity. To achieve this, elastic pipes are assumed, i.e. the size of the pipe can be changed between bounds. The required traffic determined during the logical channel dimensioning task appears as lower bound for the pipe. Is is possible to define an upper bound for the pipe but it is not necessary.

However, the problem of how to distribute the bandwidth among the demands brings up the issue of fairness. Vast number of different fairness policies can be defined. In [7] the authors present different fairness considerations and detail the relation among the fairness, the available throughput and the required computational time. Our aim is to maximize the network throughput to serve unexpected demands, however unequal capacity distribution and pipe starvation should be avoided. These requirements meet the definition of the max-min fairness presented in [7].

The throughput maximization alone is not able to avoid the unequal capacity distribution. For instance, if three pipes (with 10 Mbit/s minimal bandwidth) are routed on a common 155 Mbit/s link. For two pipes only 10-10 Mbit/s are assigned and the remaining capacity is allocated (135 Mbit/s) for the third pipe. To prevent this distribution, the minimum of bandwidth of these pipes should be also maximized. However, this latter policy does not distribute the remaining free network capacities after it maximized this bandwidth minimum. This problem becomes crucial, when the network has bottlenecks. Therefore, we have to introduce an additional policy, which punishes the longer pipes.

4.1 Formal model

Let \( G(V, L, C) \) be a directed graph, where \( V \) is the set of routers, \( L \) is the set of links. \( V_{\text{DSLAM}} \in V \) and \( V_{\text{BAS}} \in V \) define the sets of the DSLAMs and BASs. Full-duplex links are assumed; however, the graph \( G \) remains directed in order to describe asymmetric traffic. Therefore a link is modelled with two opposite edges. The amount of allocated capacity on the link will be the maximum of the load of the two opposite link.

In the model presented above, a pipe is assigned to each QoS class between every pairs of DSLAMs and BASs. The requirements of a pipe are also called as demand. Let \( D \) be the set of these demands. Each demand \( d \in D \) has the following attributes: source \( (s(d) \in V_{\text{DSLAM}}) \), destination \( (t(d) \in V_{\text{BAS}}) \), minimal required bandwidth \( (B^d) \) and traffic unit size \( (b^d) \). The minimal required bandwidth is determined with the aid of (5) for each QoS pipe. The method can allocate integer multiple of this traffic unit. The Integer Linear Programming (ILP) formulations is based on network flow models [8]. The notation is as follows:

<table>
<thead>
<tr>
<th>Input parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D ) set of demands</td>
</tr>
<tr>
<td>( N ) set of nodes</td>
</tr>
<tr>
<td>( L ) set of links</td>
</tr>
<tr>
<td>( c_l ) available capacity of link ( l \in L )</td>
</tr>
<tr>
<td>( B^d ) minimal required bandwidth for each demand ( d \in D ) (according to Eq. 5)</td>
</tr>
<tr>
<td>( b^d ) size of the traffic unit for each demand ( d \in D )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Solution (variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n^d ) number of traffic units allocated for demand ( d \in D )</td>
</tr>
<tr>
<td>( x^d ) number of traffic units routed for demand ( d \in D ) routed on link ( l \in L )</td>
</tr>
</tbody>
</table>

The objective is to maximize the following function derived from the combination of fairness policies presented above:

\[ n + \sum_{d \in D} n^d b^d - c - \alpha \sum_{l \in L} \sum_{d \in D} x^d l \] (6)

Two additional variables \( (n \) and \( c \)) are introduced to im-
implement the previously defined fairness policies. The constraints are the following.

**Flow conservation constraints** The aim of these constraints are to ensure that traffic demands reach their destinations. Constraint (7) is the well known flow conservation equation. The model supports load sharing, i.e., may define more than one routes between source and destination nodes. The inequalities in (8) says that on a single link we cannot allocate more traffic units for a demand than its total traffic units.

\[
\sum_{j \in N^i} x_{ij}^d - \sum_{j \in N^i} x_{ki}^d = \begin{cases} 
 n^d & \text{if } i = s(d) \\
 -n^d & \text{if } i = t(d) \\
 0 & \text{otherwise}
\end{cases}, \\
\forall i \in N \text{ and } d \in D
\]

(7)

\[
x^d_l \leq n^d, \quad \forall d \in D \forall l \in L
\]

(8)

**Demand size constraints** Since elastic pipes are assumed introduction of bounds are necessary. In (5), for each demand, an exact lower bound is determined during pipe dimensioning task. This bound is defined with constraint (9). On the contrary, there is no given upper bound, however the available capacity on the links defines such a bound. It is obvious that more bandwidth on a link cannot be allocated than its total capacity. This rule is described in (12).

\[
b^d \cdot n^d \geq B^d, \quad \forall d \in D
\]

(9)

\[
\sum_{d \in D} x^d_l \cdot b^d \leq c_l, \quad \forall l \in L
\]

(10)

**Fairness description** To implement the fairness policy defined above, additional variables were introduced. Parameter \(n\) is responsible for preventing unequal capacity distribution. The inequalities of (11) indicate that the number of allocated traffic units has to be greater than \(n\). Formula (12) describes that parameter \(c\) is an upper bound for the allocated capacity of a demand.

\[
b^d \cdot n^d \geq n, \quad \forall d \in D
\]

(11)

\[
\sum_{l \in L} x^d_l \cdot b^d \leq c, \quad \forall d \in D
\]

(12)

**Type restriction**

\[
x^d_l \geq 0 \text{ integer}, \quad \forall d \in D \text{ and } \forall l \in L
\]

(13)

\[
n^d \geq 0 \text{ integer}, \quad \forall d \in D
\]

(14)

5 Numerical Example

In previous sections we presented a framework for ensuring QoS in broadband access networks and presented techniques for defining a logical overlay network onto the physical one. In this section we try to shed more light on the process of the configuration task through a simple numerical example. Extensive simulations and comparison with other methods are planned to be performed later.

Figure 2 shows the physical topology of the sample network. On the left side there are 11 DSLams. The DSLams can be categorized into three classes: city (DSLAM #12–#18), suburban (DSLAM #22–#24) and rural (DSLAM #25–#26). For the sake of simplicity now the BASs are the content provider servers: a VoD server and a VoIP gateway. The link capacities are also shown on the figure.

For the subscribers three services are provided: voice (VoIP), video on demand (VoD) and best effort. Capacity for unexpected VoIP calls are not ensured, thus, they can be described as rigid traffic (the lower and upper bound of the QoS pipes transporting VoIP traffic are the same). The best effort traffic (BET) uses the remaining free capacity in the network. Therefore, between every pairs of DSLAMS and BASs two logical channels are defined: one for the video and an other for the voice traffic.

The properties of the DSLAMS depend on its class. First of all, the numbers of connected subscribers are 250, 150 and 100 for city, suburban and rural classes respectively. Moreover, in the first two cases, the probability of initiating a new VoD demand is higher. All parameters of the DSLAM classes are presented in Table 1.

The first task is to determine the number of simultaneous VoD and VoIP demands. To calculate them Eq. 1 is applied. In the current example the blocking probabilities of VoD and VoIP are set to 0.01 and 0.001. For instance 33 VoD channels are needed in case of suburban DSLAMS.

Although the number of simultaneous channels are de-
Table 1. Required number of simultaneous VoD/VoIP channels and amounts of equipment capacities.

<table>
<thead>
<tr>
<th>DSLAM class</th>
<th>Number of Users (n)</th>
<th>p</th>
<th>h</th>
<th># of channels</th>
<th>C_{Req} [kbit/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>250</td>
<td>0.2/120</td>
<td>90</td>
<td>30</td>
<td>61624</td>
</tr>
<tr>
<td>Suburb</td>
<td>150</td>
<td>0.2/120</td>
<td>90</td>
<td>33</td>
<td>45146</td>
</tr>
<tr>
<td>Rural</td>
<td>100</td>
<td>0.1/120</td>
<td>90</td>
<td>15</td>
<td>26094</td>
</tr>
</tbody>
</table>

Table 2. Offered number of simultaneous VoD and VoIP calls (PLR = 1%)

<table>
<thead>
<tr>
<th>DSLAM type</th>
<th>C_{Req} [kbit/s]</th>
<th>Max Cap. [kbit/s]</th>
<th>Max VoD</th>
<th>Call intensity</th>
<th>Max Pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>61624</td>
<td>158208</td>
<td>159</td>
<td>0.75/120</td>
<td>934</td>
</tr>
<tr>
<td>Suburb</td>
<td>45146</td>
<td>126528</td>
<td>123</td>
<td>0.93/120</td>
<td>699</td>
</tr>
<tr>
<td>Rural 1</td>
<td>25094</td>
<td>52608</td>
<td>40</td>
<td>0.38/120</td>
<td>386</td>
</tr>
<tr>
<td>Rural 2</td>
<td>25094</td>
<td>105472</td>
<td>98</td>
<td>1.095/120</td>
<td>1095</td>
</tr>
</tbody>
</table>

determined, the size of the logical pipes still have to be calculated. To perform the calculation (5) is to be used. If 1% packet loss ratio (PLR) is supposed, the equivalent capacities are shown in Table 1.

Now, all the demand parameters are defined that are needed for the network configuration; except the size of the traffic units. Both for VoD and VoIP traffic bandwidth is allocated in 64 Kbit/s units.

The next task is to solve the network configuration problem according to the model presented in Section 4.1. The solution consists of the physical paths, the reserved capacities on the links, and last but not least, the allocated and the maximal available bandwidths for the logical pipes.

Table 2 shows the results. It can be seen that a lot more capacity can be provided for the pipes than previously calculated. This means that the network is capable to transport a lot more video sessions with proper QoS level. In the case of Suburb DSLAM class 3.6 times more (122 instead of 33) VoD demand can be served if necessary, so it is possible to serve unexpected VoD demands by expanding the VoS pipes. This could happen if service initiation intensity increases e.g. more customer want to see a popular movie. On the other hand, this result shows that a larger population may be served using the assumptions presented above.

6 Conclusion

In this article a QoS capable architecture of broadband access networks was presented. To simplify the call admission control tasks the config entity defines a logical overlay network model by establishing logical channels between the edge nodes for each QoS class. We presented a method capable of determining the equivalent capacities of the logical channels that is the minimum required capacity needed to fulfill the QoS requirements of the aggregated traffic belonging to the channel.

To be able to determine the maximum available bandwidth values for the logical pipes and map these pipes onto physical paths an integer linear programming based network configuration method has been introduced. The suggested methods were illustrated through a detailed numerical example, in which we could see that the investigated network would be able to deal with more simulations demands than formerly supposed, thus, either a bigger population or a larger percentage of the same population can be served.

References


