B-ISDN Connection Admission Control and Routing Strategy with Traffic Prediction by Neural Networks

Joaquim E. Neves*, Luis B. de Almeida**, and Mário J. Leitão***

* DEI, Universidade do Minho, Azurém, P-4800 Guimarẽs
** DEEC-IST, Avenida Rovisco Pais, P-1000 Lisboa
*** DEEC-FEUP, Universidade do Porto, Rua dos Bragas, P-4000 Porto
INESC - Instituto de Engenharia de Sistemas e Commputadores
*,*** Largo Mompilher 22, P-4000 Porto
** Rua Alves Redol 9, P-1000 Lisboa
eneves@inescn.pt, lba@sara.inesc.pt, mleitao@inescn.pt

Abstract

The resource allocation in the Broadband Integrated Services Digital Network (B-ISDN) can be based in an overall network performance function described in this paper and named quality of operation. The quality of operation function is determined itself by bandwidth and quality of service functions. The traffic patterns of the quality of service for each call are predicted by neural networks. The applicability of the quality of operation function to connection admission control and call routing is proposed and supported by simulation results.

1 Introduction

The B-ISDN is a connection oriented network, although it may support both connectionless services and connection oriented services. The Asynchronous Transfer Mode (ATM) provides the transport and the switching of B-ISDN services in fixed size data packets called cells. The transport of the services, like interactive video telephony, high quality video and audio broadcast programs or data file transfer, requires a call establishment phase and the adaptation of the information flow in ATM cells, which can be transported in the existent Plesiochronous (PDH) or Synchronous (SDH) Digital Hierarchies or in the new cell-based transmission systems [1].

The resource allocation for ATM call connections can be made by taking the peak bit rate of ATM cell sources as reference, but no statistical gain is obtained by multiplexing many sources. If the resource allocation is made by the average bit rate of ATM cell sources, the statistical gain obtained by multiplexing many sources is maximum but the simultaneous occurrences of the peak periods of some sources can drastically increase the cell loss rate.

The strategy proposed in [3] for connection admission control is based in a cost function called Quality of Operation, which establishes a compromise between the maximum number of connections accepted and the satisfaction of the quality of services negotiated for connections established. Within this technique, traffic prediction is required, and neural networks, learning traffic patterns of the network operation in previous situations, have been used in the quality prediction in the admission phase of each new connection.

Other applications of neural networks in B-ISDN have been proposed, such as in operation and maintenance (OAM), signal processing and in service coding, namely in video and audio compression [8]. Neural networks have also been suggested in [7] to control the routing in spatial switches, and in [6] to control an ATM network using three hierarchical levels of neural networks.

This paper is organized as follows. The evaluation of the network quality of operation functions is discussed in the next section. Section 3 addresses the applicability of the quality of operation to the connection admission control and to the call routing, with traffic prediction by neural networks. Section 4 describes simulation results. The traffic of ATM services is characterized at the network interfaces, and the simulation model is briefly described. The simulation of ATM traffic with different connection admission control methods as well as the training and test of several neural network topologies are discussed and the results are compared. Section 5 presents the conclusions.

2 Quality of Operation

The Quality of Operation is a concept presented in reference [4], and is defined by a function that integrates the parameters of the quality of service negotiated by the network in the call establishment, the availability of network resources, and the equilibrium between the connection rejection rate of different ATM service classes. The quality of operation function (QO) has been defined by the following expression:

$$QO = \sum_j (\alpha_j A_j + \beta_j B_j - \chi_j X_j - \sum_i \delta_{ji} \Delta_{ji}) \qquad (1)$$

where α_j, β_j, χ_j, and δ_{ji} non-negative real control parameters, and A_j, B_j, X_j and Δ_{ji} are functions. A_j quantifies in terms of quality of operation the bit rate allocated to each service class j; B_j quantifies the bit rate free to be allocated to each service class j; and X_j quantifies the deviation of the connection rejection rate of the service class j from the average connection rejection rate of all service classes; Δ_{ji} quantifies the main quality of service requirements to each service class j, namely cell loss rate (i=0) the delay (i=1) and the delay variation (i=2).

For a given switching node or transmission link, the bit rate allocated function A_j is defined by:

$$A_j = \epsilon_j.Cap - \mid \sum_k E_{jk} - \epsilon_j.Cap \mid \tag{2}$$

where E_{jk} is the bit rate allocated to each call k of service class j, Cap is the throughput capacity and ϵ_j is the control parameter. A_j has an increasing contribution to quality of operation function until the allocated bit rate to each service class j does not reach a certain threshold (if $\sum_k E_{jk} < \epsilon_j.Cap,\ A_j = \sum_k E_{jk}$) and has a decreasing contribution if the allocated bit rate to that service class is bigger than the threshold (if $\sum_k E_{jk} > \epsilon_j.Cap,\ A_j = 2.\epsilon_j.Cap - \sum_k E_{jk}$).

The bit rate free function B_j can be calculated by:

$$B_j = E'_j.[int(\frac{Cap - \sum_k E_{jk}}{E'_j})] \tag{3}$$

where Cap is the throughput capacity, E'_j is the mean requested bit rate of the calls characterizing service class j, and int is the integer function. As the int function gives the number of calls available for service class j (calls with average bandwidth), B_j expresses the net bandwidth that can be effectively used by service class j.

The deviation of the connection rejection rate of each service class X_j to the overall connection rejection rate is evaluated by:

$$X_j = \mid \Phi - \Phi_j \mid . \tag{4}$$

Φ and Φ_j are rejection rates, respectively overall and for the service class j, which in turn are obtained from

$$\Phi_j = \frac{\Gamma_j}{H_j} \tag{5}$$

$$\Phi = \frac{\sum_j \Gamma_j}{\sum_j H_j} \tag{6}$$

where Γ_j and H_j are respectively the number of connections rejected and the number of connections requested by the network from the service class j,

The connection rejection rates can be evaluated continuously, or within a moving window of a given time length. In the first case more weight should be given to the more recent calls by multiplying periodically Γ_j and H_j by a constant $(0 < \phi < 1)$.

The function Δ_{ji}, express the contribution of the quality of service requirements (the cell loss rate I_0, the delay I_1, and the delay variation I_2). This contribution is ussumed to be proportional to the bit rate allocated to each service class j:

$$\Delta_{ji} = \frac{\sum_k E_{jk}}{\sum_j \sum_k E_{jk}}.I_i \tag{7}$$

The quantification of the control parameters has been discussed in [4], and their values are dependent of the B-ISDN operation scenarios, the predominant services and the desirable network load. The acquisition times of the quality of operation variables have to be compatible with the time constants of the services and network.

3 Connection Admission Control and Call Routing

The decision for the connections request to be accepted or rejected is based in the B-ISDN quality of operation expected, with and without the inclusion of the new connection, in each node and link of the call route. Figure 1 sketches the block diagram of the connection admission control system.

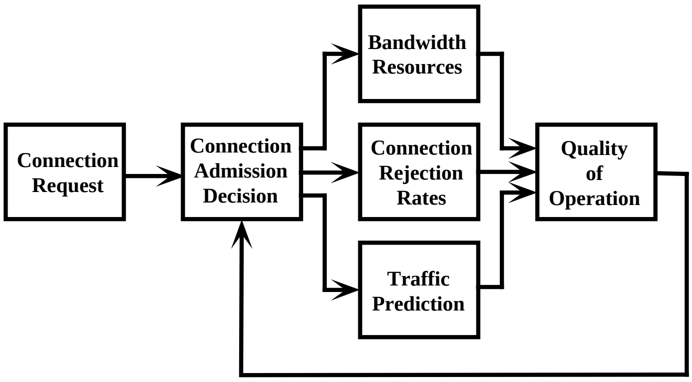


Figure 1: Connection admission control block diagram

When a request for resource allocation to a call arrives to a B-ISDN node, the quality of operation variables related to calls already established, such as the allocated bit rate, the free bandwidth and the connection rejection rate, are known to the control entity. The other variables that are related to the traffic generated if the new connection is inserted are not known but can be predicted by a neural network.

In case where a connection has available alternatives routes, the cost or quality function of the routing algorithm can include the quality of operation expected in every call path component (network node and link). A linear combination of the quality of operation of each component of the call path is one suitable routing quality function. The values of the quality of operation control parameters for the routing processing are not generally the same of those used on the connection admission control. For instance, the quantification of the allocated bit rate of each class is essential for the connection admission decision, but if it was included in the routing quality function with considerable weight, the routing of each call would have tendency follow the more loaded nodes and links. For routing the calls to paths less loaded, the overall routing quality function QO^R can be the sum of the quality of operation QO^n in each call path component n, with the allocated bit rate control parameter null ($\alpha_j^n = 0, \forall j, n$). For simplicity, the other control parameters can be made equal to those used for the connection admission. The desirable path for each call is that which maximizes the following routing quality function:

$$QO^R = \sum_n QO^n = \sum_n \sum_j (\beta_j^n B_j^n - \chi_j^n X_j^n - \sum_i \delta_{ji}^n \Delta_{ji}^n) \quad (8)$$

If the number of alternative routes, and the number of nodes of each route, is small, the best path of each call can be found in real time for each call, otherwise the best routing can be determined periodically and all calls within the same time interval follow the established route.

3.1 Bandwidth Resources and Connection Rejection Rate

When a connection request arrives to the control entity, the quality of operation is evaluated in the cases the request is accepted and is rejected. In the calculation of the connection rejection rate Φ_j for the case of accepted requests, previous consecutive rejected calls are counted as accepted, in order to give a greater chance for the service j to access the network.

Figure 2 illustrates the contribution to the quality of operation of the allocated bit rate, the free bandwidth and the connection rejection rate, when the resource allocation to a narrowband service class decrements the available net bandwidth to a broadband service class (i.e. a single narrowband call takes the bandwidth which just inhibits a broadband call to be accepted). Two situations are considered: in figure 2a) a single narrowband service class is generating calls; in figure 2b, the node is carrying a high load and the effect of the other service classes is also included.

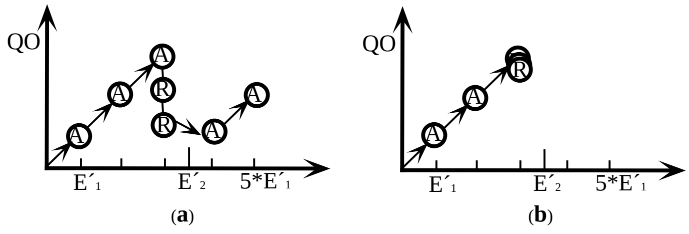


Figure 2: Quality of Operation vs. allocated bit rate to service class 1

Each time a narrow band service (class 1), with average bit rate E_1', establishes a new connection (circle A), the Quality of Operation (QO) function is incremented according to equations (2) and (3), by $(\alpha_1 - \beta_1).E_1'$. When the allocated bandwidth to the service class 1 inibits a broadband call with average bit rate E_2' to be accepted (in figure 2, $E_2' = 3.5E_1'$), the QO function is also incremented by $(\alpha_1 - \beta_1).E_1'$, but it is now decremented by $\beta_2.E_2'$, which causes the new connection to be rejected (circle R in figure 2a)). The new connection requests continue to be rejected until the decrement produced in the QO function by the deviation of the connection rejection rate of this service class X_1 to the overall connection rejection rate, evaluated by equation (4), exceeds the decrement produced by the average bit rate of the service class 2, given by equation (3) ($\chi_1.X_1 \geq \beta_2.E_2'$).

If the node is heavily loaded, as considered in figure 2b), the QO function does not decrease as much if a new connection is rejected (overlap circles R), because the connection requests are also rejected for the other service classes. This causes new connection requests to the service class 1 to be permanently rejected.

3.2 Traffic Predition by Neural Networks

Patterns of the traffic load in a B-ISDN node or link can be collected during the operation of the B-ISDN, in different traffic situations, to be used as learning patterns of neural networks. The delay and the cell loss rate that will be introduced by the call are not known at the time of the call establishment. When the resources of the new call connections are established, the vector of the allocated bit rate of each service class is stored and the traffic load pattern is evaluated later, when the new call is generating traffic. The data collected is then used in the neural network training with the backpropagation algorithm.

Neural network inputs are the allocated bandwidth to each service class, and the outputs can be the expected delay, cell loss rate, and the maximum and minimum buffer occupation, the latter leading directly to delay variation. Another output is included (the number of arrived cells) to allow a better behavior of the training process.

After the training phase, the neural network can be used in the normal operation of the B-ISDN. When a connection request arrives, each node processor asks to its neural network the expected traffic load pattern, for the node and adjacent link, with and without the inclusion of the new connection. The network answers with the expected patterns, the quality of operation is evaluated in both cases, and the resources are allocated to the connection if the expected quality of operation in every node and link of the call route is higher if the new connection was accepted.

4 Simulation Results

The B-ISDN components (transmission links, switching nodes) and procedures (routing, flow control) are simulated according to the model presented in [3], while the ATM traffic is generated by the Markovian model described in [2]. Transmission links and switching nodes are represented by delay, error rate, throughput and buffer length. Each ATM traffic source is characterized by two state spaces. The call birth of different services and users are calculated by a Markovian process with different duration and average time between call birth in each state. The cells within each connection are generated by another Markovian process, with the appropriate parameters for each service.

Three service classes have been simulated. The call generation process alternates between an activity and a silence state with a probability of leaving the state of *70%*. The quantum duration of each state is *2.5* seconds. This option gives the possibility of evaluating the behavior of the network in many combinations of the load of the services. The calls have a mean duration of *3.5* seconds and are generated in the activity state with exponentially distributed intervals with *5* milliseconds of average. The average cell rate of the services used in the simulations reported here are *1.6, 3.75* and *20.0* Kcell/s, while the peak cell rates are *10.0, 5.0* and *20.0* Kcell/s, respectively for service classes *0, 1* and *2*.

4.1 ATM Traffic Prediction

Feedforward neural networks with a layered topology, and many number of neurons in the hidden layers, have been simulated, with different activation functions. The neurons of each layer are connected by synapses to any neuron of the forward layers. For training the neural network, the backpropagation algorithm was used with adaptive learning rate parameters [5] and the sum of squared errors as cost function. Table 1 shows the average absolute error over the training and the test patterns in a neural network with different topologies and activation functions in the hidden layer neurons. The neural network has been trained with *60%* of the *3500* traffic patterns of a test set. The patterns have been collected with *5 ms* between samples during a previous simulation of a node with a buffer lenght of *100* cells. The patterns are normalized to the throughput capacity. The neural network has been trained during *2500* epochs with a moment term equal to *0.1*; the learning rate parameter was been initially set to *0.001* and the learning rate acceleration factors, using the technique proposed in [5], are *0.7* and *1.2*. The average error given in table 1 as been sampled at the end of the training, while the number of epochs reported is the number that was required to reach a stable situation, defined as less than *10%* difference from the end of the training.

The figures of table 1 show that, independently of the number of neurons in the hidden layer, the hyperbolic tangent and the inverse tangent activation functions present much better accuracy than the logistic function. Considering that the average of the normalized test pattern outputs is *0.433*, the relative error observed with the inverse tangent and with the hyperbolic tangent is about *10%*, which is suitable for traffic predictions.

Table 1: Training and test of the neural network

Number	Activation Function								
of	Logistic			Arctangent			Hyperbolic Tangent		
Hidden	Training		Test	Training		Test	Training		Test
Neurons	Error	Epoch	Error	Error	Epoch	Error	Error	Epoch	Error
00	0.160	815	0.161	0.162	68	0.163	0.161	67	0.162
02	0.169	267	0.168	0.074	174	0.073	0.150	80	0.151
05	0.152	327	0.153	0.057	628	0.056	0.054	413	0.053
10	0.138	387	0.140	0.054	461	0.053	0.050	786	0.050
20	0.125	1824	0.126	0.051	547	0.050	0.046	648	0.046

The results reported in the next subsection were obtained with a *3* layer neural network with *10* neurons in the hidden layer, linear activation function in the output layer, hyperbolic tangent activation function in the internal neurons, and was trained with all *3500* traffic patterns used in the test reported in table 1.

4.2 Performance Simulation

The connection admission control method based in the requested average cell rate and the peak cell rate for each service, have been simulated and compared with the quality of operation connection admission control technique. Figure 3 shows the allocated bandwidth during the *25 seconds* of simulation time in one node with a buffer capacity of *100* cells. The results are normalized to the node capacity and are shown in the three cases: allocation based on average cell rate; allocation based on the peak cell rate; allocation based on the proposed technique, with the following values for the control parameters: $\alpha_j = \epsilon_j = 1.0$, $\beta_j = \chi_j = 0.1$, $\delta_{j0} = 0.4$, $\delta_{j1} = 0.4$ and $\delta_{j2} = 0.2$, $\forall j$.

As seen in the figure, with the allocation by the average and the peak cell rate only the narrowest band service class (narrowest average and narrowest peak, respectively) can access the network resources, namely during the significantly loaded periods. With the proposed technique, the figure shows that all the service classes can share the available resources even when demand is higher. The histogram of the buffer occupation shows that, with the average allocation method, a full buffer occupation is reached, while with the peak allocation

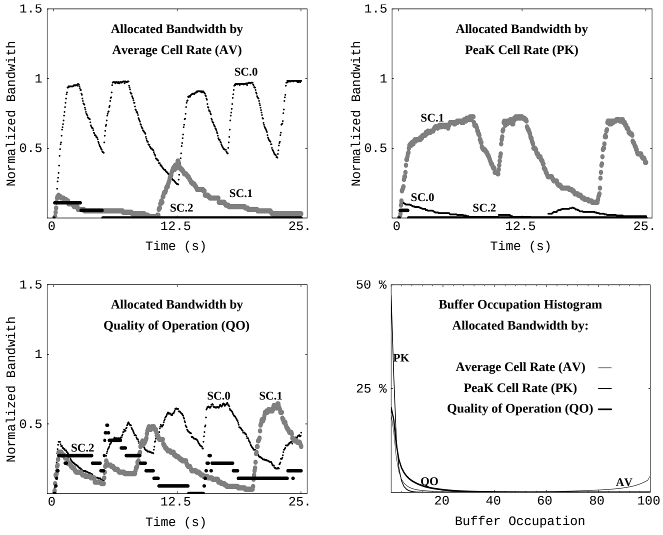


Figure 3: Connection admission control - allocated bit rate and buffer occupation for different strategies

method the buffer is lightly loaded. With the proposed technique, the average load in the node does not show any of the problems of the two other methods. Moreover the statistical distribution of the buffer occupation can be controlled by the QO function parameters.

Figure 4 presents a routing simulation, showing the average allocated bandwidth of a network composed of five nodes interconnected by six links as sketched in figure 5. The simulation time is *25* seconds.

All nodes generate traffic towards node *4*: node *0* generates traffic from the three service classes; node *1* generates only service classes *1* and *2*, while node *2* generates classes *0* and *2*, and node *3* generates only classes *0* and *1*.

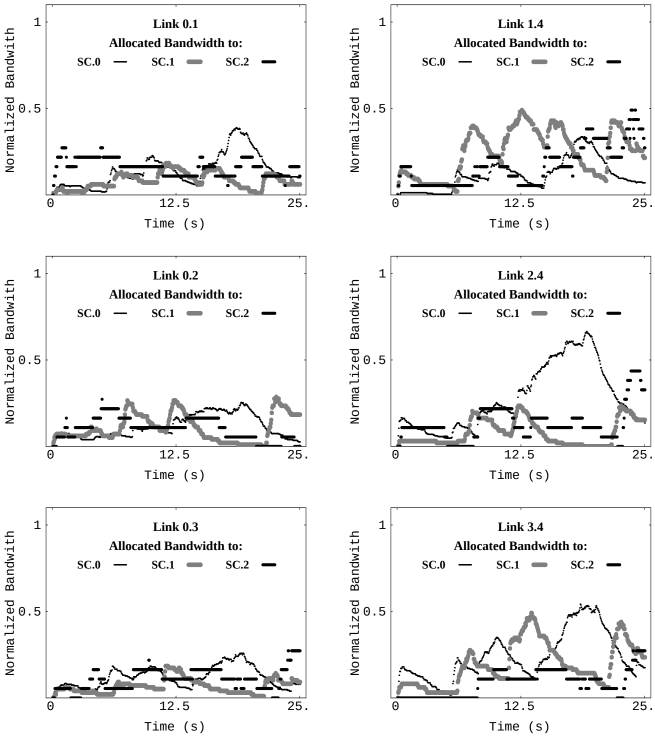
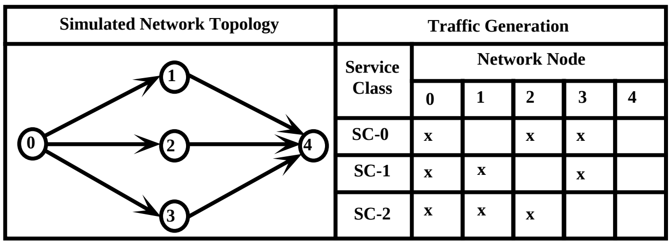


Figure 4: Routing - allocated bit rates in different routes



Traffic Generation					
Service Class	Network Node				
	0	1	2	3	4
SC-0	x		x	x	
SC-1	x	x		x	
SC-2	x	x	x		

Figure 5: Routing - topology of the simulated network

For routing calls from node *0*, the routing quality function given by equation (8) was used, with the following control parameter values: $\beta_j = \chi_j = 0.1$, $\delta_{j0} = 0.4$, $\delta_{j1} = 0.4$ and $\delta_{j2} = 0.2$, $\forall j$. The results show a good balance between usage of links *0.1, 0.2* and *0.3*, confirming that the quality of operation, with suitable values of the control parameters for routing purposes, has capabilities to find a suitable route for the calls.

5 Summary

The Quality of Operation is an overall B-ISDN quality function which incorporates the allocated bandwidth, the connection rejection rate and ATM traffic related variables. During the previous operation of the B-ISDN, patterns of the traffic load in nodes and links are collected to be used as training patterns of neural networks, for predicting the ATM traffic related variables of the new connections. The Quality of Operation function can be used as a decision criterion to control the resource allocation to the ATM call connections. When a connection request arrives to an interface node, the QO function is evaluated in every B-ISDN components involved in the call path, for admission control and for routing purposes, in case the new connection is accepted and in case it is rejected. The resources are allocated to the new connection if the quality of operation is higher, if the new connection is accepted, in every component of one available call path.

References

[1] "B-ISDN draft recommendations," *Rec. I Series*, Geneva: CCITT, May 1990.

[2] J. E. Neves and M. J. Leitão, "A markovian model for ATM traffic generation," *IEEE-Malaysia International Conference on Communications - MICC 93*, (Kuala Lumpur), Nov. 1993.

[3] J. E. Neves and M. J. Leitão, "Modeling of B-ISDN for performance simulation of ATM services," *International Symposium on Modular Information Systems and Networks - ICSNET 93*, (St. Petersburg), Sept. 1993.

[4] J. E. Neves, L. B. Almeida, and M. J. Leitão, "ATM call control by neural networks," *International Workshop on Aplications of Neural Networks to Telecommunications*, (Princeton, NJ), Oct. 1993.

[5] F. M. Silva and L. B. Almeida, "Acceleration techniques for the backpropagation algorithm," *Neural Network EURASIP Workshop Proceedings*, (Springer-Verlag), 1990.

[6] T. Takahashi and A. Hiramatsu, "Integrated ATM traffic control by distributed neural networks," *IEEE Transactions on Neural Networks*, vol. 1, Mar. 90.

[7] T. Troudet and S. Walters, "Neural network architecture for crossbar switch control," *IEEE Transactions on Circuits and Systems*, vol. 38, pp. 42–56, 1991.

[8] Yi-Tong, R. Chellapa, A. Vaid, and B. K. Jenkins, "Image restoration using a neural network," *IEEE Trans. on Acoustics Speech, and Signal Processing*, vol. 36, no. 7, pp. 1145–1151, July 1988.