Adaptive Technique for ATM Call Admission and Routing Control using Traffic Prediction

by Neural Networks

Joaquim E. Neves
DEI, Universidade do Minho, Guimarães
INESC, Largo Mompilher, 22 P-4050 Portol
Joaquim.Neves@dei.uminho.pt

Mário J. Leitão
DEEC-FEUP, Universidade do Porto, Porto
INESC, Largo Mompilher, 22 P-4050 Porto
mleitao@inescn.pt

Luís B. Almeida
DEEC, Instituto Superior Tcnico, Lisboa, Portugal
INESC, Rua Alves Redol 9, P-1017 Lisboa,
lba@inesc.pt

Abstract

This paper discusses a technique for call admission and routing control, based on a global quality function, which is dependent on the allocated bandwidth, the free network capacity and the call rejection rate, and incorporates quality of service functions, predicted by neural networks. The superior capability of this technique to support admission and routing decisions, according to the characteristics of the traffic generated by admitted calls, is demonstrated by simulation results carried out using suitable traffic and network models, which are equally discussed.

It is also shown that the proposed technique, being based on several observed traffic parameters, offers better results than methods based only on declared bandwidth parameters.

1. Introduction

The Broadband Integrated Services Digital Network (B-ISDN), based on the Asynchronous Transfer Mode (ATM), is able to support a mix of connectionless and connection oriented services, with a great variety of throughput and quality of service requirements.

Traffic control in ATM networks is mainly assured by preventive mechanisms, such as connection admission, but reactive methods acting on the information flow can also be used at the user network interfaces, in order avoid network congestion (e.g. cell discard).

Many publications in the last years have proposed feedforward or recurrent neural networks for several ATM traffic control applications and for network management [4, 5, 8]. As the ATM traffic and the behavior of B-ISDN components are non-linear and complex in nature, artificial neural networks are more suitable than analytical techniques to carry out such control functions. They not only lead to a high flexibility, allowing the satisfaction of specific service needs, in the user and network operator perspectives, but are also very efficient to adapt resource allocation to stringent network load conditions.

This paper discusses the use of a neural network based technique for connection admission control and call routing [5], presenting simulation results showing its capabilities for call admission and routing, according to the characteristics of the traffic generated within the B-ISDN. The following section presents the call control process, based on a global quality function which aims to represent an operational objective. The admission and routing decisions are made according to the predictions made by neural networks for a few quality of service parameters, expected for the new connections. In section 3, the simulation models of ATM traffic sources and B-ISDN components are briefly described and, based on these models, the simulation results are presented in section 4, showing the good performance of this technique, and the advantage over other mothods. The conclusions are summarized in section 5.

2. Call Control

Most of the published call admission methods are based either on the availability of bandwidth effectively necessary to guarantee the negotiated quality of service, or on the prediction of the values of the quality of service parameters expected for a given allocated bandwidth.

The initial approaches of the first type were based on the use of the peak or average cell rate as basic admission criteria. More elaborated methods introduced the concept of equivalent capacity [3] or the effective bandwidth [2, 9], which give a value between the peak and average, depending on the type of traffic and desirable maximum cell loss rate. However, as these methods use only simple bandwidth parameters, they have limited capabilities to represent the complexity of the competitive process which they attempt to control.

Several methods based on analytical predictions of the quality of service parameters, such as the cell loss rate, have also been proposed for call admission control [1]. However these methods are usually based on convolution operations, which become very complex for realistic traffic and network models.

It is preferable to perform the prediction by processing a set of samples of traffic parameters, which can be easily done by neural networks [4, 8, 7]. This is the approach used in the technique for call admission and routing control discussed in [5], which manages the competitive access of new calls from different services, and considers quality of service objectives established in terms of time and semantic transparency. The call control entities obtain traffic prediction from neural networks associated to each node and link of the call paths to decide the admission of each new connection.

2.1. Quality of Operation

The proposed method introduces a *quality of operation (QO)* function [5] as a measure of network performance, and defines this function 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} are non-negative real control parameters, and A_j, B_j, X_j and Δ_{ji} are functions of the network state.

To manage the competitive access of new connections belonging to different service classes, the quality of operation (QO) incorporates, for each class j, functions of the allocated bandwidth (A_j), the free transmission capacity that can be allocated (B_j) and the deviation of the connection rejection rate from the average connection rejection rate of all service classes (X_j). The time and semantic transparency is also included in the quality of operation (Δ_{ji}): cell loss rate (i=0), delay (i=1) and delay variation (i=2).

The decision to accept or reject a new connection is based on the expected quality of operation, with and without the inclusion of the new connection, in each node and link of the path. With this technique, a new connection is accepted if it increases the quality of operation (e.g. increasing the allocated bandwidth without any large increase of delay and cell loss). Figure 1 sketches the block diagram of the connection admission control process.

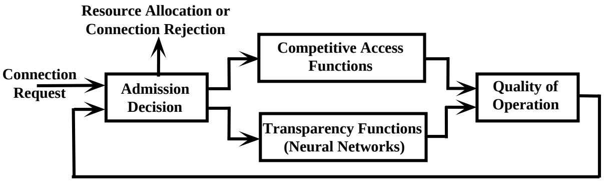


Figure 1. Block diagram of the connection admission control process.

When a node or link is asked to allocate resources to a connection, the quality of operation variables related to connections already established are known to its control entity. The variables that are related to the traffic that will be generated if the new connection is accepted can be calculated (for competitive access functions) or predicted by the neural network (for time and semantic transparency functions). The quality of operation can then be computed with and without the inclusion of the new connection. Finally, the resources are allocated to the call if the expected quality of operation in every B-ISDN node and link of the call route is higher with the new connection than without it.

For routing a call to alternative paths, the QO of each network node and link may be included in the cost or quality function of a conventional routing algorithm. Since routing objectives may be different from call admission ones, the values of the control parameters may not be the same for the two situations.

2.2. Traffic Prediction by Neural Networks

In principle, any neural network paradigm suitable for supervised training could be used to predict the ATM traffic variables that are used in the transparency functions. Among the supervised training paradigms, multilayer perceptrons are the most frequently used, since they usually yield the simplest and most accurate solutions.

As discussed in [7], fully connected multilayer perceptrons are able to predict ATM traffic parameters in the switching nodes and in the transmission links, with a reasonable training time. These neural networks must have at least one hidden layer and the neurons of each layer can be connected by synapses to any neuron of the contiguous layers. A linear activation function is recommended for the neurons of the output layer and several symmetric activation functions, such as the hyperbolic tangent and the inverse tangent, are suitable for the neurons of the input and hidden layers.

For training the neural network, the backpropagation algorithm can be used with adaptive learning rate parameters and with the sum of squared errors over the training patterns as cost function.

Figure 2 presents the block diagram of the real time neural network training to perform the prediction of ATM traffic parameters, on a given B-ISDN switching node or transmission link.

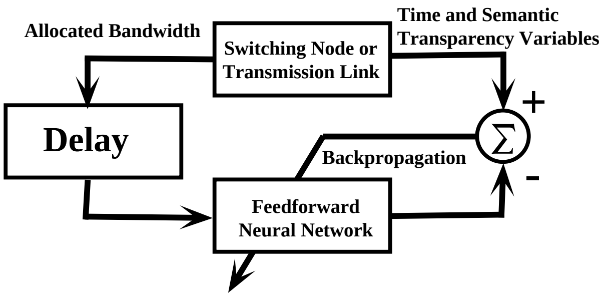

Allocated Bandwidth
Switching Node or Transmission Link
Time and Semantic Transparency Variables
Delay
+
Σ
Backpropagation
-
Feedforward Neural Network

Figure 2. Block diagram of the neural network training.

The bandwidths allocated to the connections of different types of service classes are used as the input traffic variables for training process. The sampling of the output traffic variables is delayed with respect to the input traffic variables. The value of this delay must be chosen according to the B-ISDN operation environment to guarantee the statistical significance and stability of the measured variables. For instance, to predict cell loss rate, delay and delay variation, the sampling delay needs to be of the same order of magnitude as the time constants which characterize the traffic sources and the measurement window of these parameters.

3. Simulation Models

As the B-ISDN is not available yet, a suitable choice of real time (i.e. discrete events) simulation models for ATM traffic sources and network components is unavoidable, since the reliability of the results is strongly dependent on the accuracy of such models.

Most of the analytical or real time simulations reported in the literature are carried out with ATM traffic models characterized by two alternating states with exponentially distributed durations. The cells are generated during the active state at constant (peak) rate, while in the silent state no cells are generated.

Because the state durations are exponentially distributed in these models, the number of cells generated in the active state is also exponentially distributed. Due to a high number of states of short duration and a small number of states of longer duration, and also due to the discrete nature of the cell generation process, the average and the peak cell rate observed in the traffic deviate from the values declared in the analytical law, especially when the average state duration becomes short compared to the interval between consecutive cells. As a rule of thumb, the parameters of these models are only valid if the average peak duration is at least ten times greater than the time between cells; but even in this case, the model is quite limited because it is too simplistic and incapable of emulating short burst sources.

To overcome the drawbacks of the traffic models with exponentially distributed state durations, the model presented in [6] defines three functional levels to control the generation of ATM traffic (i.e. connection requests or cells) at B-ISDN user network interfaces.

In the generation level, a traffic source is specified by a Markovian state space, with geometrically distributed state durations and, for each state, different timing distribution functions of the generated traffic (i.e. time between events and event durations). The synchronization level incorporates the timing characteristics of the environment, such as the slotted nature of the ATM traffic. The adaptation level performs the low pass filter functions for the cell stream, to guarantee at the B-ISDN user network interface the peak cell rate, negotiated at the establishment of the connection; it affects other characteristics, such as the burstiness and, if disabled, it allows the introduction of violations of peak parameters by the traffic sources.

The B-ISDN simulation model discussed in [7] incorporates in its components the traffic effects of the cell loss due to buffer overflows or transmission errors, the propagation delay and the delay variation due to accumulation of cells in buffers. According to this model, B-ISDN components are Switching Nodes and Transmission Links, modeled by buffers in which the cells are always read by a first in first out (FIFO) discipline. Each node or link is characterized by the following parameters: Buffer Length, Throughput Capacity, Transmission Cell Loss Rate and Minimum Delay.

Switching nodes can be connected, without restrictions, by transmission links, according to the desired network topology. Within the switching nodes there is a routing table that addresses, for each call, the outgoing link of the call path, while in each transmission link, the routing table addresses the destination node.

4. Simulation Results

Several simulations have been carried out to show the capability of the quality of operation function to adapt admission and routing decisions to the traffic characteristics of three service classes (SC_0, SC_1 e SC_2), which attempt to gain access to the network resources.

Each service class generates call requests, each of which requires only one connection in a single direction. The call generation process has one active state with mean intervals between call requests of 5, 15 and 100 ms, respectively for the service classes SC_0, SC_1 e SC_2, and one inactive state with no call requests.

For accepted calls, the cell stream of each service class is characterized by the traffic parameters presented in table 1 (the equivalent capacities are given for cell loss rates of 10^{-8}, according to [3]). Service classes were chosen to produce high burstiness and low average bit rate (SC_0), small burstiness and medium average bit rate (SC_1) and high constant bit rate (SC_2).

Table 1. Main traffic characteristics of the service classes SC_0, SC_1 e SC_2.

Service Class	Average Cell Rate (kcell/s)	Equivalent Capacity (kcell/s)	Peak Cell Rate (kcell/s)	Peak Duration (ms)
SC_0	1.60	2.02	10.00	0.167
SC_1	3.75	4.09	5.00	2.000
SC_2	20.00	20.00	20.00	-

Within each node and link, the time and semantic transparency variables of the cell stream generated by the service classes SC_0, SC_1 e SC_2 were predicted by a neural network with *3* layers and *10* neurons in the hidden layer, linear activation function in the output layer, hyperbolic tangent activation function in the internal neurons. The training was supervised by a set of *3500* traffic patterns (sampled in previous simulations) with the backpropagation algorithm, during approximately *800* epochs.

4.1. Admission Control

Two simulations, illustrating the performance of the quality of operation function on call admission control, are presented next, for the case of a single node with throughput capacity of 155 520 $kbit/s$ and a buffer length of 100 cells. In these simulations, the following values have been chosen for the control parameters of the QO function: $\alpha_j = 1.0$, $\beta_j = \chi_j = 0.1$, $\delta_{j0} = 0.6$, $\delta_{j1} = 0.5$ and $\delta_{j2} = 0.2$, $\forall_{j=0,1,2}$. These figures where chosen by performing successive simulations, with the objective of achieving a regulating effect of each transparency variable [5], which guarantees competitive access between services.

In the call generation process, for both states of each service class, the quantum duration was fixed to 2.5 seconds while the transition probability between states was set to 0.7. The duration of the calls is exponentially distributed, with an average of 3.5 seconds for all service classes, which can easily be shown to be a value long enough to reach overflow situations if all connections of only one class were accepted.

The cell stream of each service class was generated by two different traffic models.

Model A: It is a widely used model in the literature in which the cell generation process has two alternated states with exponentially distributed durations. The cells are generated during only one state (active state) at constant (peak) rate, according to the generation parameters presented in table 1 (for SC_2 the generation process has, in fact, only the active state, since the duration of the other state is null).

Model B: This is a more elaborated traffic model proposed by the authors [6], in which the cell generation process can have a different number of states for each service, with constant or geometrically distributed state durations. The parameters of the generation process were chosen to give various load conditions.

The cell stream for SC_0 has a minimum interval between cells of 2.7 μs, but the majority of cells are generated with intervals multiple of 0.1 ms. The duration of the silence intervals is multiple of 1.0 ms.

In case of SC_1, there are only 2 active states with equal average durations and equal probability (50%) of staying or leaving from one to the other. The cells are generated in both states with exponentially distributed intervals between cells.

The generation process from SC_2 is simply based on one state and the interval between cells is exponentially distributed.

Figure 3 presents simulation results showing, for all combinations of 2 service classes, the number of calls observed with the quality of operation method, with the cell stream of calls generated by $Model\ A$ (top pictures) and by $Model\ B$ (bottom pictures). For comparison, the figure also shows the maximum number of calls that can be theoretically reached by the average cell rate, the equivalent capacity and the peak cell rate admission methods. Simulations results of these three methods have already been presented in [7], where it was shown that the service class with the smallest requested bandwidth (average, peak and equivalent bandwidth, respectively) dominates the competition for network resources, especially during the significantly loaded periods. With the quality of operation method, all the service classes share the available resources even when demand is higher.

It can be seen in the figure that SC_0 and SC_2 have more calls accepted with the cell stream generated by $Model\ A$ than with the one generated by $Model\ B$. With SC_1 there is a little difference between the two simulations, which results from the competition with the other services. The explanation for this behavior seen for SC_0 and SC_2 is given next.

In the simulation with the cell stream generated by $Model\ A$, there are periods with even more calls accepted from SC_0 than the maximum admitted by the average cell rate method. The reason of this result is that, with this generation model, the actually observed traffic parameters (such as the average cell rate and the peak duration) of the cell stream from SC_0 are lower than the corresponding declared values (the peak duration was only 1.65 times the interval between cells). This shows that the quality of operation method manages to allocate bandwidth in an efficient way, which is possible since the admission process is not only based on declared values, but it also adapts to the observed traffic (i.e. transparency variables). In the other simulation, with generation by $Model\ B$, the number of calls accepted from SC_0 is lower than by the average method and approximately equal to the equivalent capacity method, since this model guarantees the declared traffic parameters even for very short peak durations.

As mentioned before, the cell stream of SC_2 is generated at constant rate by $Model\ A$ and with exponentially distributed intervals between cells by $Model\ B$. In this case, there is no difference between declared and actual values of the average cell rate as it happens with SC_0. The exponential distribution used in $Model\ B$ implies higher fluctuations on the node buffer, which in turn imposes more restrictions on the number of accepted calls.

Figure 4 shows the total amount of allocated bandwidth, the cell generation rate and the maximum delay, observed with intervals of 50 ms during the period of 25 s of each previously described simulation. The results are normalized to the node capacity (allocated bandwidth and cell rate) and full buffer (maximum delay).

Although the allocated bandwidth, on the simulation with the cell stream generated by $Model\ A$ is higher than with the cell stream generated by $Model\ B$, the cell generation rate is equivalent on both simulations. This is due to the fact already mentioned that $Model\ A$ is not able to generate the cell stream according to the declared parameters, specially in very short bursty sources, as it happens with SC_0.

The highest values of the maximum delay measured in the 50 ms windows are approximately equal in both simulations, but the average of the maximum delays is slightly higher in the simulation with the cell stream generated by $Model\ B$. This small difference was expected, since the exponential distribution of the interval between cells in $Model\ B$ leads to higher fluctuations on node buffer occupation, which means higher delay variations.

As a conclusion, these results show that the proposed technique can adapt the admission decision to the node traffic characteristics, in order to accept the highest possible number of connection requests, from different service classes, without degradation of the time and semantic transparency parameters (i.e. quality of service) from accepted calls.

4.2. Routing Control

A simulation was carried out to show the capability of the quality of operation function to adapt the combined admission and routing decisions to the load of each node. The state durations of the call generation process were fixed to 5 s, with unitary probability of transition between both states. The call durations have been fixed to exactly half of the state duration.

The cell streams of the service classes SC_0, SC_1 e SC_2, have been generated by $Model\ B$ and the control parameter values of the QO function, have been suitably chosen for admission ($\alpha_0 = 0.9$, $\alpha_1 = 0.8$, $\alpha_2 = 0.7$, $\beta_j = \chi_j = 0.1$, $\delta_{j0} = 0.4$, $\delta_{j1} = 0.5$ and $\delta_{j2} = 0.2$, $\forall_{j=0,1,2}$) and routing control ($\alpha_j = \beta_j = \chi_j = 0.0$, $\delta_{j0} = 0.4$, $\delta_{j1} = 0.5$ and $\delta_{j2} = 0.2$, $\forall_{j=0,1,2}$).

The topology of the simulated network is sketched in the center of figure 5. Nodes *0, 1* and *2,* generate, on active states, traffic from service classes SC_0, SC_1 and SC_2, towards node *3*. In node *0*, the call generation process is always active, while in nodes *1* and *2* it alternates between active and inactive states. The allocated bandwidth within each node and the cell generation rate within each link, during the 20 s of simulation time, are also shown in Figure 5.

The results show that all nodes and the links *1.3* and *2.3*, are almost full, but not overloaded, which means that admission control objectives (e.g. the satisfaction of quality of service in terms of time and semantic transparency variables) have been taken into account.

The cell generation rate observed on links *0.1* and *0.2* shows that the path chosen to route the calls, from node *0* to node *3*, was always the opposite of the path which included the node in the active state. This shows that the call control process has adapted the admission and routing decision to the network traffic load characteristics.

5. Summary

This paper has discussed the application of a novel technique for connection admission control and call routing, based on a global quality function, in which the admission decision is made according to the prediction, by neural networks, of a few quality of service parameters, expected for the new connections.

The simulation models of ATM traffic sources and B-ISDN components have been briefly described and, based on these models, results of resource allocation simulations have been presented. The tests carried out have shown the capability of this technique to adapt admission and routing decisions according to the traffic load characteristics.

Since the ATM call control based on the quality of operation incorporates variables which reflect the competitive access and the time and semantic transparency, it presents, for a variety of load situations, better results than other methods based only on the required bandwidth.

References

[1] C. Courcoubetis, G. Fouskas, V. Friesen, and S. Sartzetakis. Real-time issues in call acceptance management for ATM networks. In *Proceedings of the Fifth RACE TMN Conference*, page III.1/3, 1991.

[2] A. I. Elwalid and D. Mitra. Effective bandwidth of bursty, variable rate sources for admission control to B-ISDN. In *IEEE International Conference on Communications - ICC'93*, pages 1325–1330, Geneva, CH, May 1993.

[3] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications*, 9(7):968–981, Sept. 1991.

[4] A. Hiramatsu. Integration of ATM call admission control and link capacity control by distributed neural networks. *IEEE Journal on Selected Areas in Communications*, 9(7):122–130, Sept. 1991.

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

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

[7] J. E. Neves, M. J. Leitão, and L. B. Almeida. Neural networks in B-ISDN flow control: ATM traffic prediction or network modelling? *IEEE Communications Magazine*, 33(10):50–56, Oct. 1995.

[8] E. Nordström, J. Carlström, O. Gällm, and L. Asplund. Neural networks for adative traffic control in ATM networks. *IEEE Communications Magazine*, 33(10):43–49, Oct. 1995.

[9] E. D. Sykas, K. M. Vlakos, K. Tsoukatos, and E. N. Protonotarios. Congestion control - effective bandwidth allocation in ATM networks. In *4th IFIP Conference on High Performance Networking*, Liége, Dec. 1992.

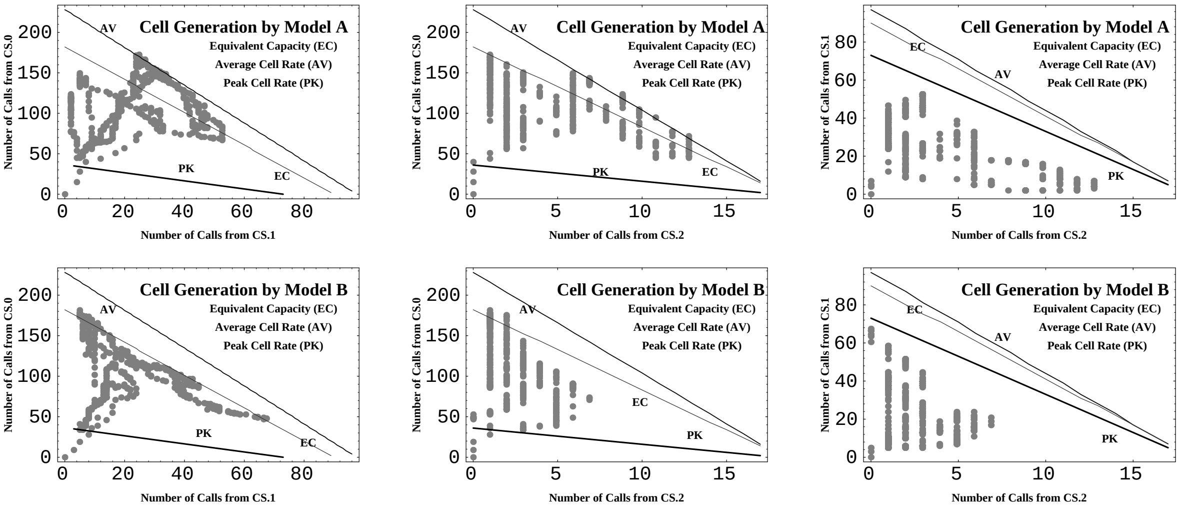


Figure 3. Number of calls from the service classes SC_0, SC_1 e SC_2, with cell generated by $Model\ A$ and by $Model\ B$. Straight lines: upper bounds corresponding to three classical admission control criteria. Grey dots: situations observed in simulations with the ”quality of operation” admission control criterion.

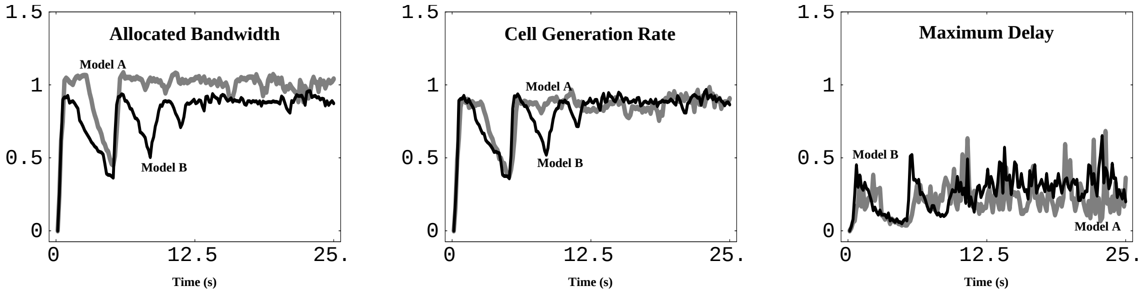
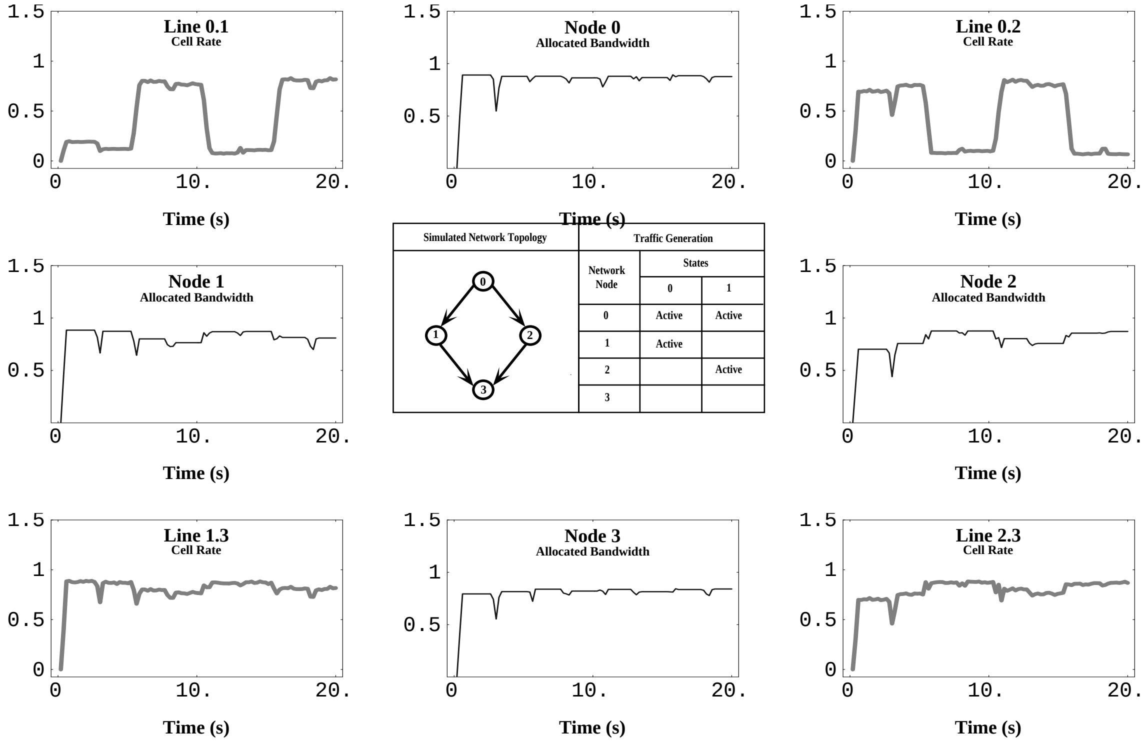


Figure 4. Parameters for the total amount of traffic, with cell generated by $Model\ A$ and by $Model\ B$.



Traffic Generation		
Network Node	States	
	0	1
0	Active	Active
1	Active	
2		Active
3		

Figure 5. Normalized allocated bandwidth within each node and cell rate within each link of the simulated network (topology and traffic activity shown in the center picture).