Queueing Network Modelling with Distributed Neural Networks for Service Quality Estimation in B-ISDN Networks

Alex Aussem†, Antoine Mahul†, Raymond Marie‡

†LIMOS/ISIMA - University Blaise Pascal, Clermont Ferrand II
B.P. 125 63173 AUBIERE Cedex FRANCE - E-mail: alex@isima.fr
‡IRISA - Campus de Beaulieu 35042 RENNES Cedex FRANCE

Abstract

We discuss an original scheme based on distributed feedforward neural networks (NN), aimed at modelling several queueing systems in cascade fed with bursty traffic. For each queueing system, a neural network is trained to anticipate the average number of waiting packets, the packet loss rate and the coefficient of variation of the packet inter-departure time, given the mean rate, the peak rate and the coefficient of variation of the packet inter-arrival time. The latter serves for the calculation of the coefficient of variation of the cell inter-arrival time of the aggregated traffic which is fed as input to the next NN along the path. The potential of this method was sucessfully illustrated on several single server FIFO queues in [3]. We now apply this technique to model a small queueing network made up from a combination of queues in tandem and in parallel fed by a superimposition of OnOff sources. Our long-term goal is the design of preventive control strategy in a multiservice communication network.

Keywords - Neural networks, B-ISDN networks, traffic control, queueing systems.

1 Introduction

Broadband Integrated Services Digital networks (B-ISDN) are intended to allow a flexible integration of a wide spectrum of services (including data, voice and video traffics) with different traffic characteristics and very specific quality of service requirements. While ATM and IP networks are now converging towards the same goal, they are not yet able to meet this difficult challenge [8]. Indeed, congestion is still experienced due to variation in network conditions. The origin of this problem is to be found in the bursty character of the stream of data cells arriving at a buffer. Arrivals are strongly correlated and cannot be accurately modeled by a Poisson process any longer. Therefore, switching gives rise to new problems in queueing theory which cannot be solved using the mathematical tools in common use in conventional queueing theory. Hence the need for traffic control methods, such as neural networks, to adapt in real-time to time-varying traffic conditions [12].

Neural networks have been employed in the literature to predict: i) the short-term behavior of the traffic [2, 12], ii) the parameters of a source model that best match the statistical descriptors of a real bursty source [6], iii) the transient leaky bucket characterization of data traffic [7], iv) the packet loss rate in a multiplexer to perform Call Admission Control (CAC) [5, 13, 14, 15, 16, 17], routing and/or bandwidth allocation [4] and others (see references). A somewhat different tack is taken in this study. Instead of paying attention to a single buffer, we investigate, in this paper, a way to model a *complete* queueing system by means of a simple feedforward NN fed with three descriptors of the aggregated input traffic. The NN is trained to output a characterization of the *average* queue load and a burstiness descriptor of the flow leaving the queue. The latter is fed as input to the next NN along the path. In this way, the overall communication network may be modelled by means of several distributed connectionist networks located on the switches with a view to achieving global resource management. The validation phase for various single-server queues has been completed in [3], we now assess the usefulness of the method on queues in tandem and in parallel with a superposition of packetized voice traffic.

In Section 2, the traffic descriptors are discussed. Section 3 addresses how the learning capabilities of NN's are used to model small queueing systems fed with packetized voice traffic.

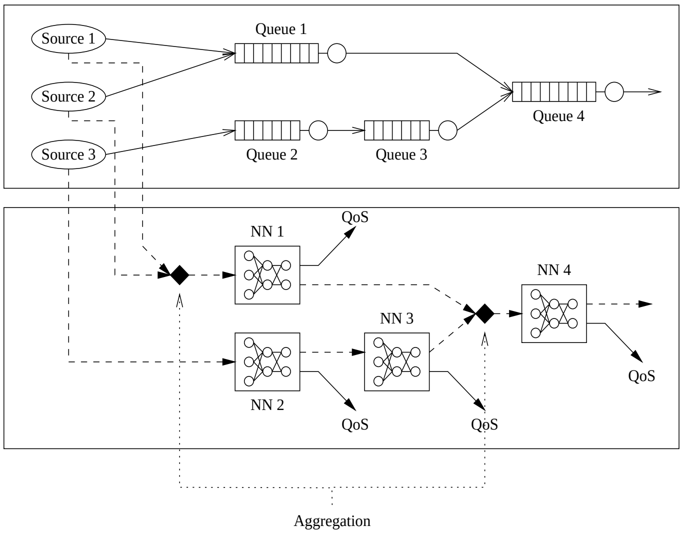

Figure 1: The queue system is represented by distributed NNs interconnected but trained separately.

2 Traffic descriptors

In a network allowing the statistical multiplexing of bursty sources, it is essential to be able to characterize the cell generation process within a call. Traffic descriptors should ideally be: i) understandable by the user, ii) significant for resource allocation, and iii) verifiable by the network. Overall, it should be also possible to deduce an easy characterization of the aggregate traffic streams by combining the descriptors of the individual sources. With the above objective in mind, we limit ourselves to: i) the average bit rate, ii) λ, the peak rate, p, and iii) the coefficient of variation of the cell inter-arrival time, $Cv^2 = var(X)/E^2(X)$, where X is the length of time between the beginning of the transmission of a given cell and the beginning of transmission of the previous cell. The peak rate and the squared coefficient of variation of the inter-arrival time are relevant measures when considering cell loss [8]. Variability in cell arrivals has been connected to the queueing delays cells are subject to. The general rule is that more variability corresponds to longer delays.

Now, let class i represent a set of *independent*, *stationary* sources with common statistical characteristics, J is the number of classes, N_i is the number of active sources of class i. Then λ_a, p_a and Cv_a^2, the aggregate traffic descriptors are approximated by:

$$\left[\sum_{i=1}^{J} N_i\lambda_i, \quad \sum_{i=1}^{J} N_i p_i, \quad \frac{1}{\lambda_a}\sum_{i=1}^{J} N_i\lambda_i Cv_i^2.\right] \tag{1}$$

This amounts to considering the aggregated trafic as a renewal process [10, 11]. This is a good approximation when J, the number of superimposed sources, goes to infinity.

3 Experiments

In a previous paper [3], we have shown that a single NN can accurately model a range of single server queueing systems, including OnOff/D/1/c, OnOff/OnOff/1/c and multi-OnOff/D/1/c queueing systems which are ideal representations of the queueing systems encountered in telecommunication networks and switching devices [1]. In this section, we report on experiments carried out to model several interconnected FIFO single server queueing systems in tandem and in parallel.

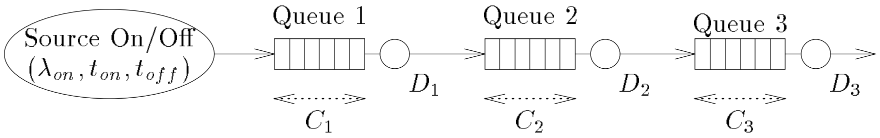

Figure 2: Three queues in tandem fed with a single OnOff source.

3.1 Simulation & Training set-up

The data sets used for training and validation purposes were obtained by discrete-event simulation and selected such that they cover as many traffic patterns as possible in a critical range where congestion is observed. 80% of the data was randomly selected for training and the remaining data was set aside for performance assessment. The target values, namely $\bar{N}$, CLR, and Cv^2_{out}, were estimated from the simulation. Since small CLR values (around 10^{-4}) are related to rare events, and thus subject to uncertainties, we implicitly assumed that the average CLR calculated over 10 replications was the correct answer. We used a logarithm conversion to map all parameter values in the $(0, 1)$ interval. The log transformed values were then linearly rescaled such that for each input variable X, $\bar{X} = 0$ and $Var(X) = 1$. Notice that the NN are trained to output the average logarithm, instead of the logarithm of the average, and the NMSE values given in this paper refer to squared differences of log-transformed data.

Fully-connected feed-forward NN with a single hidden layer were trained with the standard back-propagation algorithm. Momentum, weight decay and adaptive learning step were used to prevent local minima and avoid overfitting. For simplicity, each NN has 3 output neurons for the simultaneous prediction of the 3 targets. A standard measure of fit for the NN is given by the *normalized mean squared error*, that is, the mean squared error divided by the empirical variance of the simulation values calculated over all the N training patterns. A value of the NMSE = 1 thus corresponds to predicting the unconditional mean.

3.2 Queues in tandem

The NN scheme is applied to model an overall communication path made up from several queues in tandem, see Fig 2. We consider three traffic scenarii: a single OnOff source, homogeneous and non homogeneous superimposition of OnOff sources. Since the peak rate at the input of queue 2 and 3 is the maximum service rate of the prior queue, the latter is constant. Also, it is not taken as input variable to the NN modelling queue 2 and 3.

3.2.1 Single OnOff source

The system is first fed by a single OnOff source, a simple and common traffic scenario modelling packetized *voice* source by a succession of exponentially distributed *On* and *Off* states of average duration t_{on} and t_{off} seconds respectively. In the *On* state, cells are generated every $T = 1/\lambda_{on}$ seconds while an *Off* state corresponds to a period of silence. λ is the average rate. The descriptors of the OnOff source are directly derived by the formula: $\lambda = (\lambda_{on} t_{on})/(t_{on} + t_{off})$ and $Cv^2 = (2\lambda_{on} t_{on} - 1)/(1 + t_{on}/t_{off})^2$. The queues 1, 2 and 3 have the following deterministic service rates and capacities: (100, 13 000); (80, 12 000) and (20, 12 800). The parameters for the OnOff source were chosen at random in the following intervals: $\lambda \in [6500, 11000]$, $\lambda_{on} \in [15000, 17000]$, $t_{on} \in [10^{-3}, 10^{-2}]$. The NN has 15 hidden neurons. 500 tuples are used for training and are 100 left apart for testing. Results are shown in Table 1. Note that our attempts to replace the peak rate by the average burst length (contrary to the peak rate, the average burst length keeps varying as it is propagated further on) did not prove to be improve the results.

3.2.2 N-homogeneous OnOff sources

N-homogeneous OnOff sources (their parameters are identical) are now multiplexed before entering the system. The parameters are chosen at random: average rate of aggregated traffic $\lambda \in [6000, 11000]$, average rate of k-th source is $\lambda^k = \lambda_{tot}/N$, $\lambda^k_{on} \in [5000, 8000]$, $t^k_{on} \in [10^{-3}, 10^{-2}]$,

	NMSE		
	Cv^2	$\bar{N}$	$Log(CLR)$
NN 1	$2,07.10^{-5}$	$9,43.10^{-4}$	$1,82.10^{-3}$
NN 2	$1,15.10^{-5}$	$1,93.10^{-3}$	$5,10.10^{-3}$
NN 3	$5,24.10^{-5}$	$1,72.10^{-3}$	$3,40.10^{-3}$

Table 1: NMSE on test set when a single OnOff source is used.

and $N = 5$. Again, all sources are identical. The queues 1, 2 and 3 have the following deterministic service rates and capacities: (100, 16 500); (60, 15 000) and (40, 14 000). With these values, the CLR lies mostly between 10^{-4} and 10^{-2}. As before, the NN has 15 hidden neurons. 500 tuples are used for training and are 100 left appart for testing. Again, the peak rate, which equals the sum of the individual peak rates, is only used as input to the first NN. Results are shown in Table 2. A slight decrease of performance is observed on queues 2 and 3 compared to the first queue in terms of CLR. Interestingly, when the true Cv^2 (obtained by simulation) is replaced by analytical one, no difference is observed on the performance. This shows that the predicted burstiness descriptor is reliable.

	NMSE		
	Cv^2	$\bar{N}$	$Log(CLR)$
NN 1	$9,38.10^{-5}$	$2,62.10^{-3}$	$7,07.10^{-3}$
NN 2	$1,81.10^{-4}$	$8,14.10^{-3}$	$3,88.10^{-2}$
NN 3	$1,94.10^{-4}$	$6,34.10^{-3}$	$1,64.10^{-2}$

Table 2: NMSE on test set when N-homogeneous OnOff sources are used.

3.2.3 Two heterogeneous OnOff sources

So far we have obtained positive results with simple traffic models. We show however that the performance slightly degrades as we move up to more realistic traffic situations. Here we consider 2 independent and heterogeneous OnOff sources. The queues 1, 2 and 3 have the following deterministic service rates and capacities: (100, 13 000); (60, 12 000) and (40, 12 500). The CLR lies mostly between 10^{-4} and 10^{-2}. The traffic parameters are chosen at random in the following intervals: average rate of aggregated traffic $\lambda \in [6500, 11000]$; λ^1_{on} and $\lambda^2_{on} \in [8000, 9000]$; t^1_{on} and $t^2_{on} \in [10^{-3}, 10^{-2}]$. The average rate of source 1 is taken as $\lambda^1 \in [3000, \min(\lambda, 8000)]$, and $\lambda^2 = \lambda - \lambda^1$.

	NMSE		
	Cv^2	$\bar{N}$	$Log(CLR)$
NN 1	$1,39.10^{-3}$	$7,65.10^{-2}$	$2,44.10^{-1}$
NN 2	$1,2.10^{-3}$	$6,28.10^{-2}$	$1,26.10^{-1}$
NN 3	$1,42.10^{-3}$	$3,17.10^{-2}$	$6,95.10^{-2}$

Table 3: NMSE on test set when two heterogeneous OnOff sources are used.

Results are shown in Table 3. As far as the loss rate is concerned, the quadratic error has increased owing to the intrinsic limitations of the traffic descriptors to characterize the correlation structure of the input traffic (which is implicitly neglected in the recombination formula). However, the overall peformance is still acceptable especially for Cv^2 and $\bar{N}$.

	NMSE		
	Cv^2	$\bar{N}$	$Log(CLR)$
NN 1 & 2	$1,64.10^{-4}$	$5,57.10^{-4}$	$2,88.10^{-3}$
NN 3	$6,54.10^{-3}$	$6,94.10^{-3}$	$2,85.10^{-2}$

Table 4: NMSE on test set. Queues 1 and 2 feed queue 3.

3.3 Two parallel queues feeding a third queue

Consider now two queues in parallel fed with a single OnOff sources each Fig. 3. Their output traffics are combined to feed a third queue of larger capacity. This scheme illustrates the affectiveness of the recombination formula of the individual burstiness forecasts. Both OnOff sources are independent and heterogeneous; their parameters are chosen at random as in section 3.2.1 for construction of the training set. The queues 1, 2 and 3 have the following deterministic service rates and capacities: (100, 13 000); (100, 13 000) and (120, 22 000). The descriptor, C_v^2, for the third queue is approximated from Eq. 1. Results are shown in Table 4.

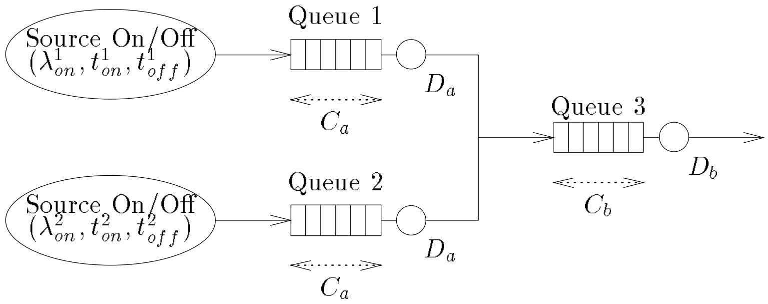

Figure 3: Two queues in parallel feeding a third one. Traffic is OnOff.

4 Discussion and Conclusion

The results presented here are useful in proving that the NN approach provides a reliable estimate of $\bar{N}$, CLR and Cv_{out}^2. The latter is transmitted to the NN modelling the next queueing system on the route and so on. Therefore, a complete communication path could be modelled by means of several NN distributed on the queueing systems. Satisfactorily tested on simple settings, the approach shows promise for a variety of applications to QoS-based traffic control, routing and congestion avoidance in multiservice telecommunication networks. However, the model may be criticized on several accounts.

First, the method, as it is, is off-line and was not yet implemented on-line with actual (possibly non-stationary) traffic. Experiments using huge TCP traffic traces (recorded on a packet by packet basis at an Ethernet gateway to the Internet, [7]) are being conducted at present in order to provide short-term forecasts of the CLR over time windows of a few seconds (ranging from 1s up to 10s) and will be reported in due course. Second, an important question is whether the variance of the interarrival time can be calculated in real-time. This raises the question whether we are modelling the network at the cell-scale or at the burst-scale. This is not yet clearly defined. In addition, the proposed traffic descriptors provide no information on the *correlation structure* of the input traffic (which is implicitly neglected in the recombination formula). This is probably the main weakness of our approach but so far, we don't know any correlation descriptor that can easily be aggregated. Other descriptors have been proposed in the litteraure [5] to take into account the correlation but, unfortunately, they are specific to the (ideal) On/Off source model. Last but not least, we are only dealing here with *average* behavior of the buffers (e.g., the average end-to-end cell delay can easily be deduced from the

Litlle formula [10]). This is not yet sufficient to guarantee QoS in terms maximum end-to-end delay which requires some understanding of the *transient* behavior of the queueing systems.

Finally, we emphasize that the methodology presented here should be viewed as a step in the direction of modelling the overall state of the telecommunication network conveying multimedia traffics. With this objective in mind, the modelling principles discussed here offer a great deal of potential for future research.

References

[1] A. Aussem, "Call admission control in ATM networks with the random neural network," *IEEE ICNN'94*, Vol. 4, pp. 2482-2487, 1994.

[2] A. Aussem and F. Murtagh, "A neuro-wavelet strategy for Web traffic forecasting," *Research in Official Statistics*, Vol. 1, No. 1, pp. 65-87, 1998.

[3] A. Aussem, Sébastien Rouxel and Raymond Marie, "Neural-based queueing system modelling for service quality estimation in B-ISDN networks", *ICANN'99*, Edinburgh, Scotland, pp. 970-975, 1999.

[4] R. Bolla, F. Davoli, P. Maryni, T. Parsini, "An adaptive neural network admission controller for dynamic bandwidth allocation," *IEEE Trans. Syst., Man, and Cyb.*, Vol. 28, No. 4, 1998.

[5] H. Brandt, E. Nordström, O. Gällmo and M. Gustafsson and L. Asplund, "A hybrid neural network approach to ATM admission control", *ISS'95*, Berlin, pp. 283-287, 1995.

[6] E. Casilari, A. Alfaro, A. reyes, A. Diaz-Estrella, F. Sandoval, "Neural modelling of Ethernet traffic over ATM networks", *EANN'98*, June, Gibraltar, pp. 304-307, 1998.

[7] F. Clérot, P. Gouzien, R. Feraud, A. Gravey and D. Collobert, "Using neural networks for predicting transient leaky bucket characterizations of data traffic. Application to dynamic resource allocation in ATM networks", *IFIP'98*, Iskley, U.K., 1998.

[8] COST 242, "Methods for the Performance Evaluation and Design of Broadband Multiservice Networks," Parts I Traffic Control, Final Report, Seminar version, June, 1996.

[9] J.J. Custodio, M. Tascon, M. Merino and Y.A. Dimitriadis , "A new neuro-fuzzy system for efficient ATM traffic control", *ICANN'99*, Edinburgh, Scotland, 1999, pp. 964-969.

[10] E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks*, 2nd Ed., Wiley, 1998.

[11] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersion" *IEEE J. Sel. Areas Commun.*, No.2, Vol.9, pp. 203-211, 1991.

[12] I.W. Habib, "Applications of neurocomputing in traffic management of ATM networks", *Proceedings of the IEEE*, No. 10, Vol. 84, Pp. 1430–1441, 1996.

[13] A. Hiramatsu, "Integration of ATM Call Admission Control and Link Capacity Control by distributed neural networks," *IEEE J. Sel. Areas Commun.*, Vol. 9, pp. 1131-1138, 1991.

[14] V. Lemaire, "Connection admission control in ATM network using an artificial neural networks", *HELNET'97*, Montreux, Switzerland, 1997.

[15] V. Lemaire and F. Clérot, "Estimation of blocking probabilities in an ATM network node using artificial neural networks for connection admission control", *ITC16*, 1999.

[16] E. Nordström, J. Carlström, O. Gällmo and L. Asplund, "Neural networks for adaptive traffic control in ATM networks," *IEEE Communications Magazine*, Oct. 1995.

[17] P. Tran-Gia and O. Gropp, "Performance of a neural net used as admission controller in ATM systems," *Proc. IEEE GLOBECOM'92*, Orlando, Florida, pp. 1303-1309, 1992.