STATISTICAL PREPROCESSING FOR SERVICE QUALITY ESTIMATION IN A BROADBAND NETWORK

Ernst Nordström, Olle Gällmo, Mats Gustafsson* and Lars Asplund

Department of Computer Systems
Uppsala University
Box 325, S–751 05 Uppsala, SWEDEN
Email: neuron@docs.uu.se

**Department of Technology*
Uppsala University
Box 534, S–751 21 Uppsala, SWEDEN
Email: mg@csg.teknikum.uu.se

Abstract: This paper describes a statistical preprocessing technique for a supervised neural network used for quality of service (QOS) estimation in an Asynchronous Transfer Mode (ATM) communication network. The preprocessing is based on standard queueing theory results and yields a good description of aggregate link traffic. The link statistics are computationally easy to obtain and comply with ATM real time requirements. The neural network target QOS is derived by an accurate link performance model to allow for a high utilization of network resources. Experimental results verify the feature detecting ability of the link statistics, and corresponding results for some conventional QOS approximation methods examplify the approximations' degradation effect on resource utilization.

1. INTRODUCTION

The Asynchronous Transfer Mode (ATM) is the transport mode recommended for the Broadband Integrated Services Digital Network (B–ISDN) scheduled for introduction in the mid–1990's. ATM is considered capable of supporting virtually all communication services expected in the future, including multimedia services, by asynchronous multiplexing of fixed sized packets called "cells" [1]. However, traffic control is needed to maintain the quality of service (QOS) of network connections. Connection admission control is effective in the context of preventive traffic control.

The admission control decision to accept or reject connection requests is based on an estimation of the selected path's QOS, given the characteristics of the connections already sharing the path, and the characteristics of the new connection. Although very accurate analytical traffic performance models, based on queueing theory have been proposed, e.g. the fluid flow model for heterogeneous traffic [2], they are too complex to be used in a real ATM network. This is especially true if many connections of different types share the links.

Traditionally, approximations are done by superposing the connections into a single or a few traffic sources [3], or e.g. by a direct approximation of the fluid flow QOS formula [4]. The approximations enable real time admission control, but at the price of reduced network throughput.

In this paper another approach is proposed. Instead of introducing approximations directly in the QOS performance model, neural networks are used to implement an accurate performance model. A multi layer perceptron is trained to recognize the nonlinear relationship between a link state vector and a QOS measure calculated by the heterogeneous fluid flow model. The link state vector contains statistical measures of the requested aggregate link traffic and is computationally easy to obtain. The link state vector may be based on user–provided traffic descriptors or on traffic measurements that reflect actual user characteristics.

Some traffic control methods for ATM networks based on neural networks have already been proposed. For example, in [5] a neural network is used to adaptively learn the relation between offered traffic characteristics and resulting QOS, which both are obtained from actual traffic characteristics.

2. ATM TRAFFIC CONTROL

The nature of ATM makes traffic control a challenging task. In ATM, user cells are asynchronously multiplexed onto high capacity links according to actual communication needs.

Connections allocate network resources only virtually, and statistical gain is possible when users are bursty, i.e. alternate between busy and silent states. At the switching nodes, output buffers are needed to resolve switching conflicts which arise when several cells simultaneously are switched to the same destination link. However, when too many cells arrive at the same time, or if switching conflicts arise repeatedly, the buffer will saturate and subsequent cells will have to be dropped. The probability of cell loss is therefore an important QOS measure in an ATM network. The buffer queues also affect the cell delay and cell delay variation, which are important QOS measures for real time traffic.

The main objective for traffic control is to allow for a high utilization of network resources, while sustaining an acceptable QOS level. The control methods are constrained by the high real time requirements of broadband networks and the fact that retransmissions are expensive since a large amount of traffic can be in transit during a propagation delay.

Traffic control is divided into reactive and preventive traffic control. In a broadband network, preventive traffic control is believed to be most important. Preventive traffic control consists of two parts: connection admission control and enforcement of user traffic. The traffic enforcement is necessary to prevent users from violating the agreed traffic behaviour.

3. CONNECTION ADMISSION CONTROL

Connection admission control procedures decide whether a new connection request should be accepted or rejected. At connection set up, a route through the network is selected. Then, the QOS of each affected link is estimated, taking the effect of the new connection into account. The connection request is accepted if each link can offer sufficient QOS to all connections. The QOS measure is calculated from traffic descriptors provided by the users, or obtained from traffic measurements.

Users are often described as two state on/off sources. In the on/off source model, a user alternates between a busy on–state and a silent off–state. In the busy state, users produce cells at a constant characteristic cell generation rate. For bursty traffic, e.g. data and voice traffic, the duration of the busy and silent periods are often modeled as exponentially distributed, and a continuous time queueing model is used for performance evaluation. The fluid flow queueing model belongs to this category.

4. THE FLUID FLOW PERFORMANCE MODEL

If the output–buffer size is sufficiently large, it is possible to model the discrete cell flow as a continuous fluid flow. The fluid flow approximation neglects the cell scale fluctuations and considers buffer saturation solely due to fluctuations in the burst scale. A continuous time Markov chain models the flow rate into the buffer, and the maximum service rate is given by the link capacity.

A fluid flow model for a finite number of heterogeneous on/off sources has been proposed in the literature [2]. In this model, user sources are partitioned into c classes of statistically identical sources. Each class j is described by four parameters: the number of sources in the class N_j, the constant cell generation rate or peak rate p_j, and the average duration time of the busy and silent periods, $t_{on(j)}$ and $t_{off(j)}$, which characterize the exponential distributions. The buffer capacity B, and the link capacity C, are also included in the model.

An estimation of the expected cell loss probability is carried out by calculating the buffer equilibrium probability distribution. The solution is obtained by solving a set of first order differential equations, by a standard eigenvalue and eigenvector approach. Each possible busy–source configuration has its own state equation, and the solution complexity therefore increases geometrically with the number of sources, making a direct numerical real time implementation intractable.

5. STATISTICAL PREPROCESSING FOR NEURAL QOS ESTIMATION

Neural networks have several properties valuable in ATM traffic control. The parallelism enables fast control actions, the adaptability offers flexibility, and error tolerance provides robustness [9–11]. QOS estimation for admission control is particularly suitable for neural networks, and may result in higher network throughput than conventional methods. The approach taken in this paper is simply to reproduce QOS estimations (in terms of the probability of cell loss) of the fluid flow model, by using a multi layer perceptron (MLP) as function approximator.

One of the more crucial parts of the neural network approach is the selection of link state statistics to be used as neural network inputs. The statistics should contain sufficient information with respect to the target performance model to enable a functional relationship. In a previous paper [7], knowledge of the desired mapping (permutational symmetry with respect to the traffic descriptor sets) were used in a second order Taylor expansion to define 15 symmetry–invariant statistics. Here, an MLP input vector of lower dimension is defined, based on knowledge of the buffer queueing system.

In [3], six statistics for characterizing an aggregate connection of heterogeneous on/off sources are presented. All statistics but one are of direct incremental nature, which is important since connection requests arrive sequentially. Three statistics are used to describe the stationary distribution of the arrival rate: the mean arrival rate M, the variance of the arrival rate V, and the third moment of the arrival rate μ_3. The impact of the k'th connection is incrementally calculated as

$$M_k = M_{k-1} + m_k \tag{1}$$

$$V_k = V_{k-1} + m_k\,(p_k - m_k) \tag{2}$$

$$\mu_{3(k)} = \mu_{3(k-1)} + m_k\,(p_k - m_k)\,(p_k - 2m_k) \tag{3}$$

where m_k is the mean cell arrival rate of the k'th connection, defined as the peak rate multiplied with the fraction of time the user is busy: $m_k = p_k\, t_{on(k)} / (t_{on(k)} + t_{off(k)})$. Two parameters express information about the correlation structure: the slope of the autocovariance function of the arrival rate at the origin C_0, and the asymptotic variance of the arrival rate v_∞ [3]. The incremental formulas are

$$C_{0(k)} = C_{0(k-1)} + m_k\, p_k / t_{on(k)} \tag{4}$$

$$v_{\infty(k)} = v_{\infty(k-1)} + 2\,(m_k / p_k)\,(p_k - m_k)^2\, t_{on(k)} \tag{5}$$

One parameter that reflects the overall behaviour of the queueing system is obtained from the fluid flow model. The largest negative eigenvalue z_0 of the eigenvalue–egienvector solution is known to characterize the exponential decay of the buffer probability distribution. It is obtained by solving the nonlinear equation [2, 3], e.g. by Newtons method:

$$\sum_{j=1}^{c} N_j\,(1/\mathrm{t}_{\mathrm{on(j)}} + 1/\mathrm{t}_{\mathrm{off(j)}}) + \mathrm{z}_{0(\mathrm{k})}\,(\sum_{\mathrm{j}=1}^{\mathrm{c}} N_j\,\mathrm{p}_\mathrm{j} - 2\mathrm{C}) = \sum_{j=1}^{c} N_j \sqrt{(\mathrm{z}_{0(\mathrm{k})}\,\mathrm{p}_\mathrm{j} + 1/\mathrm{t}_{\mathrm{on(j)}} - 1/\mathrm{t}_{\mathrm{off(j)}})^2 + 4/(\mathrm{t}_{\mathrm{on(j)}}\,\mathrm{t}_{\mathrm{off(j)}})} \tag{6}$$

The solution complexity is dependent on the number of classes c and the number of iterations to find the root $z_{0(k)}$, which initially should be set to the previous value $z_{0(k-1)}$. The six statistics define an input state vector $\mathbf{S}_k = (M_k, V_k, \mu_{3(k)}, C_{0(k)}, v_{\infty(k)}, z_{0(k)})$ for the neural network.

The present approach consists of two parts: 1) preprocessing to obtain the vector $\mathbf{S}_k$, and 2) estimating the cell loss probability with a supervised MLP. The complexity of the target fluid flow model admittedly restricts the flexibility of the approach, in the sense that introduction of new ATM source types requires off–line training. However, by initially training the neural network on a wide range of traffic situations, the need for off–line training is diminished.

6. EXPERIMENTS

Some example experiments have been performed to study the feasibility of the approach. The traffic situations selected in the experiments are restricted with respect to the number of simultaneous source classes/types, and to the range of traffic source characteristics. The traffic situations are chosen randomly, with the restrictions c=3, $N_j \in \{1, 2, ..., 25\}$, $p_j \in \{2, 4, ..., 14\}$ Mbit/s, $m_j \in \{1, 2, ..., 14\}$ Mbit/s and $t_{on(j)} \in \{1, 2, ..., 25\}$ msec. Furthermore, only traffic situations for which the fluid flow model are needed have been considered, i.e. when the mean arrival rate is less than the link capacity $(\Sigma N_j m_j < C)$, and the maximum arrival rate is higher than the link capacity $(\Sigma N_j p_j > C)$.

In the cell loss calculations, a realistic buffer capacity B of 100 cells and link capacity C of 135 Mbit/s are used. The cell loss values are logarithmically transformed to enhance the estimation near the acceptable cell loss level of 10^{-9}. The largest negative eigenvalue z_0 is transformed by $\log(-z_0)$, which improves the feature detecting ability studied in the experiments. After standard normalization procedures, a training set of 20 000 samples and a test set of 3 512 samples are obtained.

7. RESULTS

Results characterizing the feature detecting ability of the 6 queueing system statistics, and corresponding results for the 15 Taylor expansion statistics [7], are presented in table 1. The results are given as averages over 10 different standard backpropagation [11] training sessions. For comparison, training set results for some conventional QOS approximation techniques [4, 6, 8] are presented in table 2. Of these, only the ”Equivalent capacity” method has a low computational complexity, suitable for real time implementation.

In the tables, results are given as % bad decisions, defined as the percentage of traffic situations which yields a cell loss value on the wrong side of 10^{-9}, as compared to the target fluid flow model. In table 2, the % bad decision figures are further divided into % bad accepts and % bad rejects, which characterize the degradation effect on user QOS and network throughput.

Table 1.		*%Bad decisions (sdev)*	
	Neural network preprocessing	*Training set*	*Test set*
	Queueing system statistics	*3.5 (0.4)*	*4.0 (0.4)*
	Taylor expansion statistics	*5.2 (0.6)*	*5.2 (0.7)*

Table 2.	*Conventional methods*	*%Bad decisions*	*%Bad accepts*	*%Bad rejects*
	Equivalent capacity	*12.2*	*0.6*	*11.6*
	Virtual cell loss	*15.1*	*0*	*15.1*
	Binomial multi server	*23.3*	*0*	*23.3*

The results in table 1 show that the queueing system statistics are advantageous, and that no significant degradation is obtained for the test set. A comparison with table 2 shows that the conventional methods yields a significantly higher amount of bad decisions. However, the conventional methods are ”safe” in the sense that the bad decisions mainly reduce the network throughput. The incorporation of the bad accept/reject measures in the neural network error function is subject to current research.

8. CONCLUSION

In this paper, we have presented and evaluated a statistical preprocessing technique for a supervised neural network used for accurate QOS estimation in a broadband ATM network. The QOS estimation provides a basis for connection admission control, an important tool in preventive traffic control. A set of link statistics are used to characterize the aggregate connection, which is assumed to consist of on/off sources. The statistics are derived from knowledge of the buffer queueing system [3], and are computationally easy to obtain. One statistic has a computational complexity dependent on the number of traffic classes and constraints the approach somewhat.

An accurate, but complex link performance model (which allows for a high utilization of network resources) is used to derive the neural network target QOS. A wide range of predefined traffic situations are assumed to be learned before the neural network is introduced in on–line operation, since target QOS values are unattainable in real time.

Experimental results show that the queueing system statistics have good feature detecting properties. Results for some conventional QOS estimation methods are also presented, and examplify the fact that performance model approximations may reduce network throughput.

ACKNOWLEDGEMENT

The authors would like to acknowledge Anders Törnqvist and Daniel Fagerström for their contributions in discussions of our work. We also acknowledge ELLEMTEL Telecommunications Laboratories for financially supporting this project called ”Hardware Implementation of Neural Networks for Applications in the Telecom Area”.

REFERENCES

1 Händel R. & Huber M.N, *Integrated Broadband Networks: An Introduction to ATM–based Networks*, Addison–Wesley, 1991.

2 Jacobsen S., Dittman L. & Moth K., A fluid flow queueing model for heterogeneous on/off traffic, *Proceedings 8th Nordic Teletraffic Seminar*, Otnas, Finland, Aug 1989.

3 Andersson H. & Andersson H., On analytical models for multiplexing in ATM networks, Lund Institute of Technology, Department of Communications Systems, Nov 1991.

4 Guérin R., Ahmadi H. & Naghshineh M., Equivalent capacity and its application to bandwidth allocation in high–speed networks, *IEEE JSAC*, **Vol 9**, **no 7**, pp 968–981, Sep 1991

5 Hiramatsu A., ATM communications network control by neural networks, *IEEE Transactions on Neural Networks*, **Vol 1**, **no 1**, pp 122–130, March 1990

6 Murase T., Suzuki H., Sato S. & Takeuchi T., A call admission control scheme for ATM networks using a simple quality estimate, *IEEE JSAC*, **Vol 9**, **no 9**, pp 1461–1470, Dec 1991

7 Nordström E., Gällmo O., Asplund L., Gustafsson M. & Eriksson B., Neural networks for admission control in an ATM network, *Proceedings of The First Swedish National Conference on Connectionism*, Skövde, Sweden, Sep 1992

8 Jacobsen S., Moth K. & Dittmann L., Load control in ATM networks, *Proceedings of ISS 90*, Stockholm, Sweden, **Vol. 5**, pp 131–138, May 1990.

9 Hertz J., Krogh A. & Palmer R., *Introduction to the theory of neural computation*, Addison–Wesley, 1991.

10 Kosko B., *Neural Networks and Fuzzy Systems*, Prentice Hall, 1992.

11 Rumelhart D., Hinton G. & Williams R., Learning internal representations by error propagation, *Parallel Distributed Processing*, MIT Press, **Vol 1**, pp 318–362, 1986.