26

Performance Evaluation of an Inter-Stream Adaptation Algorithm for Multimedia Communications

Alaa Youssef, Hussein Abdel-Wahab, and Kurt Maly
Department of Computer Science
Old Dominion University
Norfolk, VA 23529, USA
(youssef,wahab,maly)@cs.odu.edu

Abstract

Controlling the quality of a collaborative multimedia session, which employs multiple streams, is a challenging problem. In this paper, we present and analyze the performance of an inter-stream adaptation algorithm, which dynamically allocates the shared resources reserved for a session among the streams belonging to it. The objective of this dynamic allocation is to optimize the overall session quality. The traffic characteristics of the streams are specified using the M-LBAP (Modified Linear Bounded Arrival Processes) model. The M-LBAP model provides tight characterization for the traffic while maintaining the simplicity and linearity of the LBAP model. Delay bounds for streams sharing a group reservation are analytically derived using the M-LBAP model. Degradation paths specified using the M-LBAP model are used as the basis for a dynamic rate based algorithm for inter-stream adaptation (RISA). The performance of RISA is contrasted to static resource allocation policies, and it is shown that higher utilization and acceptance ratios are achievable by RISA. These achievable results are reflected on and summarized by a proposed metric for judging the effectiveness of resource allocation on the overall quality of session.

Keywords

Inter-stream adaptation, Quality of Session, Quality of Service, resource allocation, traffic characterization.

1 INTRODUCTION

Continuous media streams represent a major component of new distributed collaborative systems, that is as equally important as the conventional data streams, for providing effective collaboration. These continuous media streams have some inherent characteristics that are not found in other data streams: they have timing and throughput requirements that must be met. In order to

Performance of Information and Communication Systems U. Körner & A. Nilsson (Eds.)

meet those requirements, several quality of service (QoS) architectures, that rely on reservation of resources, such as buffers and transmission bandwidth, were proposed. Examples can be found in (Campbell *et al.* 1994, Ferrari *et al.* 1990, Zhang *et al.* 1993).

These collaborative multimedia systems typically rely on a multi-point to multi-point communication pattern, in which multiple sources, which change over time, generate a group of streams that are multicasted to all members of the session. This group of streams cooperate to present an integrated view to the users. These streams have two main dynamic characteristics: they are activated and deactivated at any instant in time throughout the lifetime of the session; and their relative priorities with respect to each other change over the session lifetime. The first characteristic implies that traditional resource reservation techniques, which treat different streams independently, may either over-allocate resources or allow for potential rejection of connections requested on-the-fly during a session. This motivated several researchers to propose resource sharing through *group reservations* by which an application can reserve collective amounts of resources to be dynamically shared by its streams (Gupta *et al.* 1995, Zhang *et al.* 1993).

The above mentioned characteristics, together with the fact that most multimedia encoders can provide multiple grades of service, suggest that there is a need for overall control, beyond the level of QoS as pertaining to individual streams in isolation of others, for a particular application. For this purpose, we proposed the concept of *Quality of Session (QoSess)* (Youssef *et al.* 1997). The quality of the session as perceived by the end user, can best be determined by the end application. At every instant in time, the quality of the session depends on the priorities of the on going streams, from the application's perspective, as well as on the QoS offered for each of these streams. A *QoSess control layer* controls the allocation of the resources reserved for the application among the streams belonging to it, in a way that stems from the application semantics.

In this paper, we focus on the inter-stream adaptation (ISA) techniques used by the *QoSess* control layer to allocate the reserved bandwidth among the cooperating streams belonging to an application. This allocation is changed over time to match the dynamic nature of the streams. We use the M-LBAP traffic characterization model to specify the characteristics of the streams. The M-LBAP model is a variation of the LBAP model (Anderson 1993) that gives tighter characterization for the traffic of individual streams.

The rest of this paper is organized as follows. In section 2, the M-LBAP traffic characterization model is described. In section 3, delay bounds for streams sharing the same group reservation are derived analytically. A framework for controlling the quality of a multimedia session is described in Section 4, and a strategy for allocating resources among a group of cooperating streams is presented in Section 5. A unified metric for measuring the quality of a session and comparing the effectiveness of resource allocation strategies is proposed

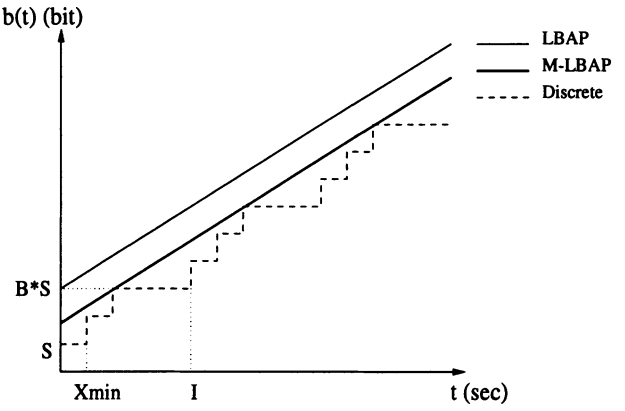

Figure 1 Bounding functions of 3 traffic characterization models

in section 6, and simulation results are discussed in section 7. Finally, our conclusions and future work are discussed in Section 8.

2 TRAFFIC CHARACTERIZATION MODEL

A key issue for providing QoS guarantees is the ability to characterize the traffic of the stream for which guarantees are being provided. A traffic characterization model must be tight enough to avoid excessive allocation of resources, and simple enough for the application to use in its specification and for the network to be able to support, as well as for the analysis to be tractable. In addition, the model should allow for the aggregate characterization of the traffic of a group of streams sharing the same resources which are reserved for the group.

In (Cruz 1991) bounding techniques, based on a fluid traffic model (σ, ρ), were developed. Central to the analysis is the concept of *traffic constraint function b(t). b(t)* is defined to be the maximum number of bits that can arrive during any interval of length t. For the (σ, ρ) model, $b(t) = \sigma + \rho t$.

The linear bounded arrival processes *(LBAP)* model (Anderson 1993), characterizes the traffic using three parameters *(R, B, S)*, where R is the average rate in bits/sec, B is the maximum burst size is packets, and S is the maximum packet size in bits. It can be easily shown that the LBAP model is simply a (σ, ρ) model with $\sigma = BS$ and $\rho = R$. The LBAP model has the advantage of being simple for the application to use in its specification as well as for the network to use in its implementation in order to support the specified characteristics for the streams.

In (Ferrari *et al.* 1990), the discrete model *(Xmin, Xave, I, S)* is described, where $Xmin$ is the minimum packet inter-arrival time, $Xave$ is the average packet inter-arrival time, I is the averaging interval, and S is the maximum packet size. In (Zhang *et al.* 1994), the bounding function $b(t)$ for the discrete

model is given by $\left(min\left(\lceil\frac{t \bmod I}{Xmin}\rceil, \lceil\frac{I}{Xave}\rceil + \lceil\frac{t}{I}\rceil \lceil\frac{I}{Xave}\rceil\right)\right) S$. The discrete model is tighter in characterizing streams but lacks a lot of the simplicity of the LBAP model. Also, determining the optimum value of I is not a trivial task and may be impossible for real-time traffic.

The *Modified-LBAP (M-LBAP)* model, which we use, is derived from the LBAP model. It strikes a balance between the simplicity of specification and analysis of the LBAP model and the accuracy of representation of the discrete model. In M-LBAP, a stream is characterized by four parameters *(R, B, S, PAR)*, where the first three parameters are the same as the LBAP original parameters, and PAR is the peak-to-average-rate ratio or the burst ratio. Figure 1 shows a graphical representation for the bounding functions of the different models. It can be easily shown that for M-LBAP, the bounding function $b(t)$ is given by $BS\left(1 - \frac{1}{PAR}\left(1 - \frac{1}{B}\right)\right) + Rt$ M-LBAP, is also a (σ, ρ) model with $\sigma = BS\left(1 - \frac{1}{PAR}\left(1 - \frac{1}{B}\right)\right)$ and $\rho = R$. This model provides a tighter characterization for the burstiness of a stream than the LBAP model and hence avoids the excessive allocation of resources.

One of the main advantages of having a linear model derived from the (σ, ρ) model is the ability to characterize a group of streams, as a single aggregate stream. It can be easily shown that the aggregate traffic of K streams, each satisfying (σ_k, ρ_k), $k = 1, 2, .., K$, satisfies $(\sum_{k=1}^{K} \sigma_k, \sum_{k=1}^{K} \rho_k)$. This characteristic of the M-LBAP model makes it adequate for characterizing the streams sharing a group reservation, and regarded by the underlying network as a single aggregate stream.

3 BOUNDING DELAYS

In a packet-switching network, the end-to-end delay of a packet consists of: (1) *link delay*, which includes the propagation delay and other delays incurred in intermediate subnetworks if some of the links are subnetworks; (2) *switching delay*, which depends on the implementation of the switches; (3) *transmission delay*, which is a function of the packet length and link speed; and (4) *queuing delay* at each switch.

Under the assumption that there are no intermediate subnetworks, or alternatively that all intermediate nodes have reservation capabilities, the link delay is constant and equal to the propagation delay. The switching delay is fixed. Knowing the link speed and the maximum packet length makes the transmission delay fixed as well. The queuing delay is the component that can be affected by controlling the load or using an appropriate service discipline, and hence is the major concern.

3.1 Bounding Delays in a FCFS Scheduler

The following theorem was stated and proven in (Zhang *et al.* 1994).

Theorem 1:

Let there be n channels multiplexed on a link with a FCFS scheduler and link speed l. If for $j = 1, ..., n$, the traffic on channel j is bounded by b(.), then the delays of packets on all the channels are bounded by d, where d is defined by

$$d = \frac{1}{l} \, max_{\forall u \geq 0} \left\{ \sum_{j=1}^{n} b_j(u) - lu \right\} + \frac{Smax}{l}$$

where, Smax is the maximum packet size that can be transmitted over the link.

Including $\frac{Smax}{l}$ accounts for the fact that a lower priority, non-real time, packet may be in transmission and cannot be preempted.

In (Youssef *et al.* 1997), we prove the following theorem, which defines the delay bounds for a FCFS scheduler and a group of streams whose traffic obey the M-LBAP model.

Theorem 2:

Let there be n channels multiplexed on a link with a FCFS scheduler and link speed l. If for $j = 1, ..., n$, the traffic on channel j obeys the M-LBAP traffic specification (R_j, B_j, S_j, PAR_j), and if $\sum_{j=1}^{n} R_j \leq l$, then the delays of packets on all the channels are bounded by d, where d is defined by

$$d = \frac{1}{l} \left\{ \sum_{j=1}^{n} B_j S_j \left(1 - \frac{1}{PAR_j} \left(1 - \frac{1}{B_j} \right) \right) \right\} + \frac{Smax}{l}$$

where, Smax is the maximum packet size that can be transmitted over the link.

Based on Theorem 2, we show in (Youssef *et al.* 1997) that for n channels sharing a group reservation at the rate of R_{tot}, on a link with speed l. If for $j = 1, ..., n$, the traffic on channel j obeys the M-LBAP traffic specification (R_j, B_j, S_j, PAR_j), and if $\sum_{j=1}^{n} R_j \leq R_{tot}$, then the delays of packets on all the channels are bounded by d, where d is defined by

$$d = \frac{1}{R_{tot}} \left\{ \sum_{j=1}^{n} B_j S_j \left(1 - \frac{1}{PAR_j} \left(1 - \frac{1}{B_j} \right) \right) \right\} + \frac{Smax}{l}$$

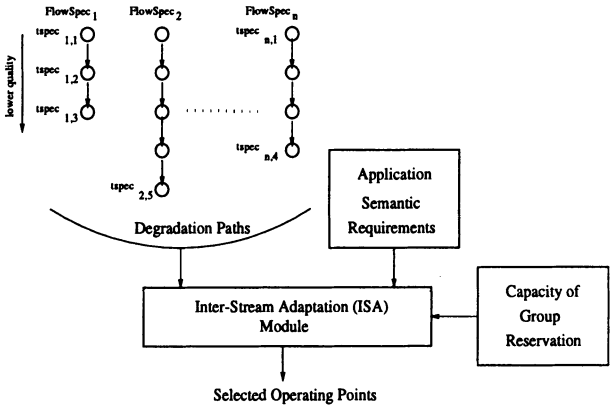

Figure 2 Using degradation paths in inter-stream adaptation

4 A FRAMEWORK FOR SESSION QUALITY CONTROL

In a networking environment where group reservation is provided to applications to support sharing of resources, the QoSess control layer must allocate fractions of the total amount of reserved resources to each stream. As shown in Figure 2, an inter-stream adaptation *(ISA)* module uses its knowledge of the degradation paths of the streams, the semantic requirements of the application, and the amount of resources reserved for the application to dynamically determine the operating point of each stream, and informs the application for enforcement.

Each operating point for a continuous media stream can be mapped from encoder specific parameters, e.g., frame rate or size, number of quantization levels, encoding technique,...etc., into traffic specific parameters. Arranging more than one operating point for a stream in the form of a degradation path, as shown in Figure 2, gives flexibility for the ISA module in adapting to availability of resources or changes in application requirements. The flow specification (*FlowSpec*) for a stream is composed of a traffic specification (*TSpec*) and a QoS requirements specification (*RSpec*). *TSpec* represents an ordered list of operating points. Using the M-LBAP model parameters, $TSpec = \{(R_1, B_1, S_1, PAR_1), ..., (R_m, B_m, S_m, PAR_m)\}$. *RSpec* represents the delay, jitter, and loss constraints for the stream as well as the relative importance of each of these factors. $RSpec = \{D, J, L, W_R, W_D, W_J, W_L\}$, where, D, J, and L are the maximum allowed delay, jitter, and loss ratio, respectively, and W_x is the weight of factor x.

In a collaborative multimedia application, the group of streams belonging to the application have a highly dynamic nature. A stream may be started/stopped at any instant. Moreover, the relative priority of a stream with respect to the other streams varies with time. In addition to priorities, other types of relation-ships between groups of streams may be implied by the semantics of the application. For example, a pair of streams, e.g., audio and video from

the same source, may be required to be always in the same active/inactive state. More complex relation-ships include inter-stream synchronization and synchronization of multiple views of the same stream.

5 RATE BASED INTER-STREAM ADAPTATION

The degradation paths of n streams belonging to an application represent an n-dimensional space. A valid point in this space represents an operating point for each of the n streams. Knowing the *FlowSpec* of each stream allows for computing the total resources required for the selected point. The rate based approach we introduce here assumes continuous values in the range $[Rmin, Rmax]$ for the parameter R of the M-LBAP model, while the other parameters are fixed. This is equivalent to changing only the sampling rate or alternatively the frame rate of the encoder while keeping all other precision and quality parameters of the encoder constant. For *RSpec*, we consider a delay bound constraint that must be respected, with no losses, and jitter is ignored. The *RISA (Rate-based Inter-Stream Adaptation)* algorithm is run whenever a change in priority occurs or a stream is activated/deactivated. It uses the above information to select an optimal point in the n-dimensional space of degradation paths. This is done in two phases:

1. Selection Phase: All streams are scanned in descending order of priority, granting each its requested minimum rate if the available bandwidth permits and the delay bound constraints for all selected streams are not violated, based on Theorem 2.

2. Enhancement Phase: The remaining non-allocated bandwidth is divided among the selected streams. The objective is to make the share of each active stream as close as possible to its specified $Rmax$, while maximizing the overall benefit gained by the session from this allocation. This resource allocation problem is formulated as an optimization problem that reduces to the well known knapsack problem (Coreman *et al.* 1990).

$$\textit{Maximize} \sum_{i=1}^{n} (p_i * f_i)$$
$$\textit{Subject to:}$$
$$\sum_{i=1}^{n} [f_i(Rmax_i - Rmin_i)] \leq R_{tot} - \sum_{i=1}^{n} Rmin_i$$
$$0 \leq f_i \leq 1 \quad for\ i = 1, 2, ..., n$$

where n is the number of streams selected in phase 1, p_i is the priority of stream i, and f_i is the computed fraction of $(Rmax_i - Rmin_i)$ which should be granted to stream i.

The knapsack problem is a special linear optimization problem for which an optimal solution can be obtained by traversing the list of streams in the order of $p_i/(Rmax_i - Rmin_i)$ and giving each stream its maximum need until all resources are exhausted (Coreman *et al.* 1990).

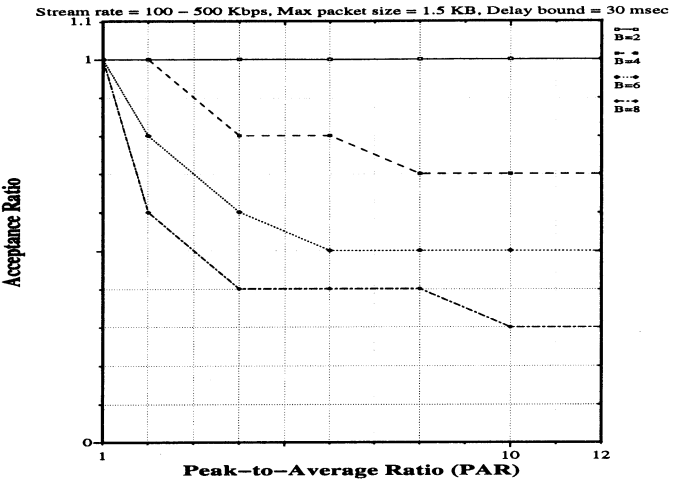

Figure 3 Effect of Peak-to-average ratio on Acceptance ratio

6 A METRIC FOR COMPARING RESOURCE ALLOCATION STRATEGIES

To compare the behavior of the system under the application of different resource allocation strategies, we propose to use a unified metric that reflects the overall performance of the system for a given allocation. At a certain instant in time, t, we define Q_i as the degree of satisfaction of stream i. The QoSess metric is the weighted arithmetic mean of these.

$$Q_i = \left\{ \begin{array}{ll} R_i/Rmax_i & if\ i\ is\ active \\ -1 & if\ i\ is\ not\ active \end{array} \right\}$$

$$QoSess = \frac{\sum_{i=1}^{n} (p_i * Q_i)}{\sum_{j=1}^{n} p_j}$$

where, R_i is the current level of resource allocation to i. The system is penalized by -1 for each inactive stream. The value of QoSess lies in the interval [-1,1]. Since the QoSess is intended for comparing different strategies, the best attainable QoSess may be below one sometimes.

7 SIMULATION STUDY

In this section we present results from simulation experiments conducted for a single node (switch). The purpose of this simulation was to investigate the effect of the M-LBAP model parameters on the performance of RISA, and to evaluate the benefits obtained from using degradation paths over static resource allocation policies, for traffic characterized by M-LBAP. In each experiment, the session was composed of identical streams with operating rate in the range from 100 to 500 Kbps. The number of streams requested to be activated was set to the maximum number that can be admitted based on the rate constraint alone (i.e. $\sum_{i=1}^{n} Rmin_i = R_{tot}$ for the requested streams),

Figure 3 shows the effect of the PAR parameter on the Acceptance ratio for the RISA approach. The Acceptance ratio is defined in (Gupta *et al.* 1995) as the number of accepted (activated) streams divided by the number of streams requested to be activated. It is clear from the figure that the effect of the PAR parameter stabilizes for values roughly above 5. This relaxes the requirement for exact calculation of PAR, which is an advantage for using PAR instead of the peak rate as the fourth parameter for the M-LBAP model.

In order to evaluate the benefits of employing degradation paths in inter-stream adaptation, the RISA approach is compared to three static resource allocation strategies that do not employ the concept of degradation paths in inter-stream adaptation. These three cases cover all the extremes. The first case (*Fixed-Min*) represents a conservative system that is designed for worst case scenarios, where $R_i = Rmin_i$ always, for each stream i. The second case (*Fixed-Max*) represents an aggressive design where $R_i = Rmax_i$. In between lies the average case (*Fixed-Avg*), with $R_i = \frac{Rmax_i + Rmin_i}{2}$.

Figures 4 and 5 show that while some of the static strategies achieve high utilization and others achieve high acceptance ratios, RISA strikes the balance of achieving both goals. This is summarized by the QoSess metric in Figure 6. Since the number of streams requested to be activated is equal to the maximum number that can be admitted based on the rate constraint only, the QoSess values for Fixed-Min are close to those for RISA. Typically, during a session there will be periods where the number of requested streams is smaller and hence significantly higher QoSess values will be obtained using RISA relative to Fixed-Min.

8 CONCLUSION AND FUTURE WORK

The need for controlling the quality of collaborative multimedia sessions that employ multiple streams, by means of inter-stream adaptation decisions, has been recently established (Youssef *et al.* 1997). In this paper we focused on sessions that operate in the presence of a reservation protocol which allows for a group reservation of a collective amount of bandwidth to be shared by the streams of a session. Degradation paths defining the possible operating levels

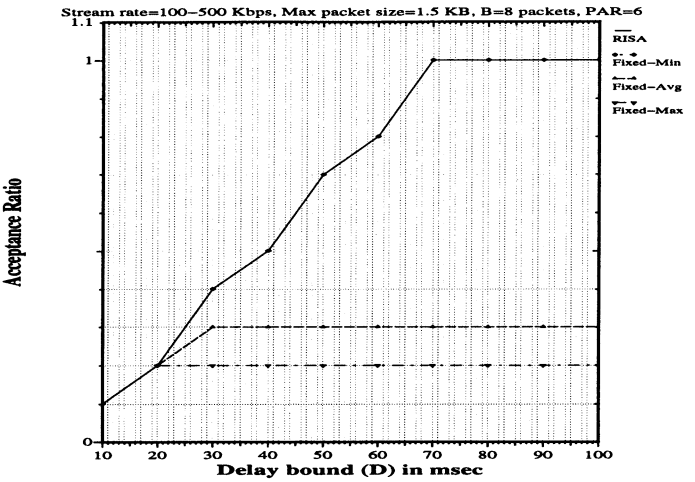

Figure 4 Effect of Delay bound on Acceptance ratio

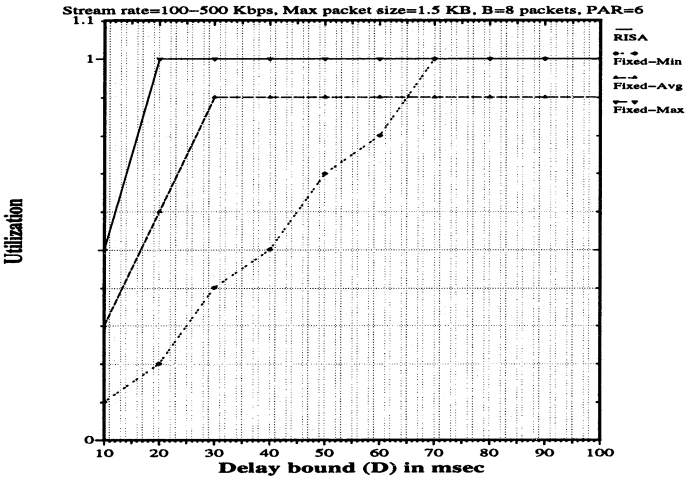

Figure 5 Effect of Delay bound on Utilization

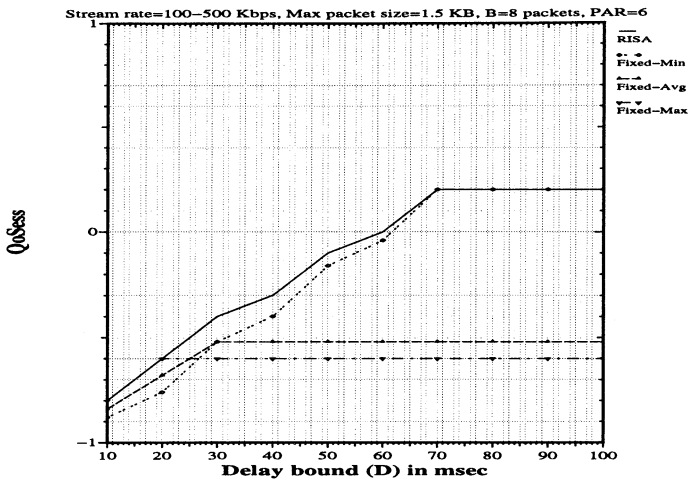

Figure 6 Effect of Delay bound on QoSess

of each stream were used as the basis for a dynamic rate based algorithm for inter-stream adaptation, RISA. Simulation results showed that higher utilization and acceptance ratios are achievable by RISA over typical fixed static policies. These results were reflected on and summarized by the introduced quality of session *(QoSess)* metric.

Currently, we are implementing the inter-stream adaptation (ISA) schemes, for more intensive experimental investigation, using the IRI (Interactive Remote Instruction) system (Maly *et al.* 1997) as a testbed. Also, we are investigating ISA policies that are based on discrete degradation paths, and situations where no group reservation is provided by the underlying network, but rather capacity estimation and resource usage monitoring techniques are used to estimate the resources available to the application.

REFERENCES

Anderson, D.P.(1993) Meta-scheduling for distributed continuous media. *ACM Transactions on Computer Systems*, **11(3)**, August 1993.

Campbell, A., Coulson, G. and Hutchinson, D.(1994) A Quality of Service Architecture. *Internal Report MPG-94-08, Lancaster University*, 1994.

Coreman, T., Leiserson, C. and Rivest, R.(1990) Introduction to Algorithms. *McGraw-Hill and MIT Press*, 1990.

Cruz, R.(1991) A Calculus for Network Delay, Part I: Network Elements in Isolation. *IEEE Transactions on Information Theory*, **37(1)**, 1991.

Ferrari, D. and Verma, D.(1990) A Scheme for Real-Time Channel Establishment in Wide-Area Networks. *IEEE Journal on Selected Areas in Communications*, **8(3)**, 1990.

Gupta, A., Howe, W., Moran, M. and Nguyen, Q.(1995) Resource Sharing for Multi-Party Real-Time Communication. *Proceedings of INFOCOM'95.*

Maly, K., Abdel-Wahab, H., Overstreet, C.M., Wild, C., Gupta, A., Youssef, A., Stoica, E. and Al-Shaer, E.(1997) Interactive Distance Learning over Intranets. *IEEE Internet Computing*, **1(1)**, January 1997.

Youssef, A., Abdel-Wahab, H. and Maly, K.(1997) Inter-Stream Adaptation over Group Reservations. *TR_97_37, Old Dominion University*, April 1997.

Youssef, A., Abdel-Wahab, H., Maly, K. and Gouda, M.(1997) Inter-Stream Adaptation for Collaborative Multimedia Applications. *Proceedings of the Second IEEE Symposium on Computers and Communications (ISCC'97)*, Alexandria, Egypt, July 1997.

Zhang, H. and Ferrari, D.(1994) Improving Utilization for Deterministic Service in Multimedia Communication. *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, 1994.

Zhang, L, Deering, S., Estrin, D., Shenker, S. and Zappala, D.(1993) RSVP: A New Resource ReSerVation Protocol. *IEEE Network Magazine*, September 1993.

9 BIOGRAPHY

Alaa Youssef is a PhD candidate and research assistant in computer science at Old Dominion University. His research interests include networking support for multimedia systems, resource management in heterogeneous distributed systems, and multimedia collaborative and tele-teaching systems. Youssef received an MSc in computer science from Alexandria University.

Hussein Abdel-Wahab is a professor of computer science at Old Dominion University, an adjunct professor of computer science at the University of North Carolina at Chapel Hill, and a faculty member at the Information Technology Lab of the National Institute of Standards and Technology. His main research interests are collaborative desktop multimedia conferencing systems and real-time distributed information sharing. Abdel-Wahab received a PhD in computer communications from the University of Waterloo. He is a senior member of IEEE Computer Society and a member of the ACM.

Kurt Maly is a Kaufman professor and chair of computer science at Old Dominion University. His research interests include modeling and simulation, very high performance network protocols, reliability, interactive multimedia remote instruction, Internet resource access, and software maintenance. Maly received a PhD in computer science from the Courant Institute of Mathematical Sciences, New York University. He is a member of the IEEE Computer Society and the ACM.