December 1993

Performance Tradeoffs for Scheduling pre-orchestrated Multimedia Information Over Broadband Integrated Networks

Zafar Ali
Purdue University School of Electrical Engineering

Arif Ghafoor
Purdue University School of Electrical Engineering

PERFORMANCE TRADEOFFS FOR SCHEDULING PRE-ORCHESTRATED MULTIMEDIA INFORMATION OVER BROADBAND INTEGRATED NETWORKS

ZAFAR ALI
ARIF GHAFOOR

TR-EE 93-48
DECEMBER 1993

SCHOOL OF ELECTRICAL ENGINEERING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1285

Performance Tradeoffs for Scheduling Pre-orchestrated Multimedia Information Over Broadband Integrated Networks

Zafar Ali, Arif Ghafoor
Distributed Multimedia Systems Laboratory
School of Electrical Engineering, Purdue University
West Lafayette, Indiana *47907*
Tel. *(317) 494-0638*
FAX *(317) 494-6440*
E-mail: ghafoor@ecn.purdue.edu

Abstract

In this report we present a framework for evaluating performance of scheduling pre-orchestrated multimedia information over broadband integrated networks. We propose a set of *Quality* Of *Presentation* (QOP) parameters which quantify the quality of multimedia presentation from user's point of view. The communication of mulltimedia data in a networked environment can affect the desired QOP parameters due to jitter delays in the network. We evaluate trade-offs between QOP parameters and the system resources including channel utilization and buffering at the destination. These trade-offs can be used to develope an optimal transmission schedule for multimedia information.

Key words: *multimedia communication, multimedia presentation, temporal synchronization, pre-orchestrated multimedia information, quality of presentation, fluid flow models, timed Petri-nets.*

1 Introduction

Advances in high-speed networking technology and increasing demand for tirnely distribution of information have resulted in a tremendous interest in variety of services which will be available in the future Distributed Multimedia Information Systems (DMIS). Most of these services will use some form of pre-orchestrated stored objects, e.g., video-on-demand, digital libraries, virtual reality, education/training, CAD/CAM, etc. The pre-orchestrated nature of the data poses a different set of challenges in management and communication than those faced in dealing with the data generated in real-time. The major challenge is to satisfy some pre-specified temporal constraints among multiple objects at the time of their playout. These constrains need to be met in spite of the heterogeneity of data, varying quality of service available over the network connections, and vastly different storage architectures. For such a purpose, it, is important that the data objects are delivered in time so that they can meet their individual deadlines. As the data may be transmitted over multiple virtual channels from the source to the destination, one possible way to meet these deadlines is to use intra-stream synchronization mechanisms [15]. These mechanisms can provide flexibility in presentation of multimedia information by adjusting the information generation and consumption rates.

For presentation of pre-orchestrated stored multimedia information in a DMIS, data must be retrieved in bulk and "well ahead" of their playout deadlines [10]. A scheduling algorithm which can generate the deadlines for the transmission and presentation of the multimedia objects is presented in [10]. The idea is to carefully orchestrate the transmission schedule at the source site according to the constraints specified by the playout schedule at the destination. This requires scheduling the transmission of all the objects involved in the presentation at some time (called ***control time)*** prior to their deadlines. The method proposed in [10] determines the control time using the largest affordable delay that can be sustained to individual objects. This approach has a drawback since it requires extensive

buffering at the destination as the end-to-end delay experienced by an object on a channel may vary widely in magnitude due to differences in quality of service (QOS) parameters [9].

The objective of this report is to present performance tradeoff between the quality of presentation of multimedia objects and the resources needed to maintain this quality. We consider two important network resources, the channel utilization and buffering requirement at the destination. In order to quantify the presentation quality, we propose two Quality Of Presentation (QOP) parameters, namely the maximum tolerable probability of buffer overflow ($\mathcal{P}_i^b$) at the destination and the probability that an object misses its deadline ($\mathcal{P}_i^d$) when transmitted from its source to the destination. These parameters directly affect the presentation process. For example, $\mathcal{P}_i^d$ is a measure of synchronization failure at the destination and $\mathcal{P}_i^b$ indicates the information loss due to finite buffering capability at the destination.

The individual values for these parameters must be determined from the delay/loss characteristic of the type of data involved in the presentation. For example, audio and video data are isochronous in nature, having stringent delay and delay jitter requirements. Therefore, for this data, low values of $\mathcal{P}_i^d$ are desirable. On the other hand, for the case of traditional data such as text or graphics, large delays can be tolerated but their delivery is very sensitive to cell losses, hence low values of $\mathcal{P}_i^b$ is needed. Our analysis provides an interplay among these parameters and the network resources (channel utilization and the destination buffer requirements). We show how the resource consideration including destination buffering, and the desired QOP parameters dictate the control time. An "optimal control time" for scheduling transmission of objects is found which guarantees both the quality of presentation and the *best* utilization of the resources. This result can be easily extended for designing an *optimal* transmission schedule for pre-orchestrated multimedia information consisting of multiple objects over broadband integrated networks. We assume that the pre-orchestrated multimedia information is stored according to Object Composition Petri Net (OCPN), which is a temporal specification model [11].

Specifically, following are the major contributions of this report.

- We propose a set of QOP parameters which can directly characterize the presentation of multimedia information.
- We present analytical results which show the effect of control time on the resource requirement for a specified QOP. These results provide the upper and the lower bound on the destination buffer requirement and the channel utilization.
- We then determine the control time which achieves the best possible presentation quality when the system resources are fixed. Accordingly, we establish a condition to determine whether or not the transmission of an object is possible.
- Finally, we find the optimal control time which minimizes the resource requirements and guarantees the specified QOP.

Our model is based on the fluid flow approximation which has been widely recognized as a powerful analytical tool for the analysis of queueing systems in packet voice communication networks and ATM (Asynchronous Transfer Mode) networks [1], [4]. We consider the information flow through the network to be uniform rather than in discrete packets/ cells. Although we assume that the underlying network is ATM based, our analysis is not restricted to these networks.

This report is organized as follows. In the next section we introduce rate flow model for an OCPN. The analysis regarding bounds on the destination buffering for given QOP parameters is given in Section 3. Section 4 establishes a condition about scheduling of multimedia objects. Section 5 provides results related to determining optimal control time for best utilization of system resources. Section 6 conclude the report.

2 Concept of Scheduling the OCPN and the Rate Model

In this section we propose a rate model for the presentation process of multimedia object. We characterize the transmission and the display processes as constant rate fluid flow, and the overall process of presentation as a "work conserving" system [8]. Consequently we describe a destination buffer occupancy function which is used to estimate the destination buffering requirements. We begin our discussion with the presentation of the new extension to the OCPN model which incorporates the performance considerations. Subsequently, we elaborate the various processes involved in the transmission and presentation of an object of an OCPN.

We need few notations for our analysis, which are summarized in Table: **(2).** All parameters with subscript i refer to object O_i. Bold faced letters are used to denote the random variables and random processes.

2.1 The OCPN Based Temporal Specification Model

Various models for the specification of temporal relationship among multimedia objects have been recently proposed in the literature. This include graphical model [11], [15], language based model [14], and object oriented model, etc. Among these models, OCPN has been used by many researchers [10], [11], [15]. It is a time-augmented Petri-Net model for specifying temporal requirements among multimedia objects. It can model concurrency, asynchrony, and logical precedence relations among various multimedia data objects in a simple way. Formally, the OCPN can be defined as follows

Definition *2.1 An OCPN is a timed augmented Petri-Net based model with places P (with cardinality | p |), transitions T (with cardinality | t |), input and output arcs A, an initial*

Table 1: Notations

Symbol	Explaination
$\mathcal{P}_i^d$	Maximum tolerable probability of deadline misses
$\mathcal{P}_i^b$	Maximum tolerable probability of destination buffer overflow
K_i	Buffer space available at the destination for O_i
s_i	Size of object O_i
$k_i = \frac{K_i}{s_i}$	Fraction of the object that can be buffered at the destination
λ_i	Transmission rate of object O_i, $\lambda_i \leq C_i$ (C_i=channel capacity)
μ_i	Consumption rate of object O_i
$\eta_i = \frac{1}{\rho_i} = \frac{\lambda_i}{\mu_i}$	Transmission link utilization factor
τ_i	Presentation duration of object O_i
τ_i^s	Transmission duration of object O_i
π_i	Playout deadline of object O_i
π_i^s	Transmission start time of object O_i
T_i^c	Control time for object O_i
$\mathbf{D}_i$	End-to-end delay
$F_{\mathbf{D}_i}(\cdot)$	End-to-end delay distribution

marking M_o, duration of the action D, and the set of resources R.

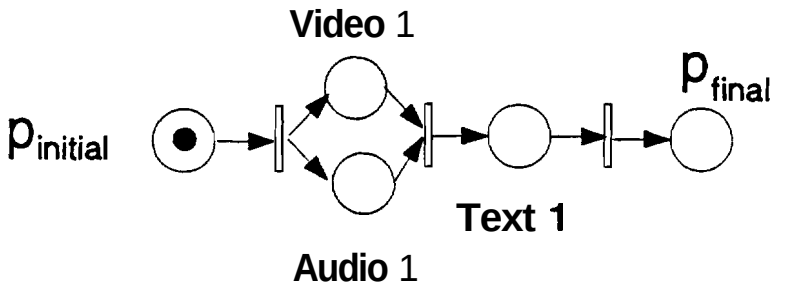
$$\begin{aligned} OCPN &\triangleq (P, T, A, M, D, R) \\ P &= \{p_1, p_2, \cdots, p_{|p|}\} \\ T &= \{t_1, t_2, \cdots, t_{|t|}\} \\ A &= \{P \times T\} \cup \{T \times P\} \\ M_o &= \{m_1^o, m_2^o, \cdots, m_{|p|}^o\} \\ D &= \{\tau_1, \tau_2, \cdots, \tau_{|p|}\} \\ R &= \{r_1, r_2, \cdots, r_{|p|}\} \end{aligned}$$

The initial marking M_o is a $(|p|, 1)$ column vector. It represents a mapping from the set of places P to the natural numbers:

$$M_o : P \rightarrow IN \quad where \quad M_o(p_i) = m_i^o \quad for \quad i = 1, 2, \cdots |p|$$

The D is a mapping from a set of places to the positive real numbers:

$$D : P \rightarrow \mathbb{R}^+ \quad where \quad D(p_i) = \tau_i \quad for \quad i = 1, 2, \cdots |p|$$



(a) An OCPN

Figure 1: Temporal Specification using the OCPN

with τ_i representing the duration of the action associated with the place p_i.

The R is a mapping from the set of places to a set of resources:

$$R : P \rightarrow \{r_1, r_2, \cdots, r_{|p|}\} \quad \textit{where} \quad R(p_i) = p_i \quad \textit{for} \quad i = 1, 2, \cdots | p |$$

where r_i denotes the collection of resources required to perform action associated by place pi.

A place in OCPN represents the actual object to be presented. These objects may be continuous media or discrete media type. An example of OCPN specification for the multimedia presentation is shown in Fig. 1 which represents the concurrent presentation (with start time π_1) of an audio and a video clips for the duration τ_1 and τ_2 respectively, followed by text (with deadline π_2) for the τ_3 time units.

The OCPN model has limitation that it cannot be used directly for specifying transmission scheciule for objects over a network and also does not render itself easily for analyzing the performance of multimedia communication over a network. We propose an extension to OCPN which can be used for evaluating the performance of network for transmission of multimedia objects and to generate an "optimal" schedule for objects. For this purpose, we differentiate two types of places in an OCPN: continuous media (CM) object places and

discrete **media** *(DM) object places. Furthermore, noting the dependence of the buffer size on the specified QOP, the object size and its display rate, we augment each place in the existing* OCPN *definition with three new attributes; the desired QOP parameters, the size of the object and its display rate. The new model is referred to as the Rate-based OCPN* (ROCPN) *which is formally defined* **as** *follows:*

Definition *2.2 An* $ROCPN \triangleq (OCPN, Q, F, S, Y, R)$ *is an eztended OC'PN with the following new mapping functions added to the original definition of the* ***OCPN.***

$Q = \{q_1, q_2, \cdots, q_{|p|}\} : P \rightarrow QOP = \{'P^d, 'P^b\}$, *is a mapping from the set of places to a set defining the desired quality of presentation, where q; defined by the* $\mathcal{P}_i^d$ *and the* $\mathcal{P}_i^b$ *represents the* ***desired*** *QOP of the object represented by the place* p_i.

$F = \{f_1, f_2, \cdots, f_{|p|}\} : P \rightarrow IF$, *is a mapping from the set of places to a set of functions, where* f_i *represents the rate of consumption of the object represented by the place p;. The consumption rate function is defined in (1).*

$S = \{s_1, s_2, \cdots, s_{|p|}\} : P \rightarrow \mathbb{I} = \{0, 1, 2, \cdots\}$, *is a mapping from the set of places to the nonnegative integers, where s; denotes the size of the object associated* ***with*** *the place* p_i.

$Y = \{y_1, y_2, \cdots, y_{|p|}\} : P \rightarrow \{CM, DM\}$, *is a mapping from the set of places to a set of types.*

$R = \{r_1, r_2, \cdots, r_{|p|}\}$, *is a mapping from the set of places to a set of* ***resource*** *types [11]. We propose the following tuple for* r_i

$$r_i \triangleq \{C_i, K_i\}$$

Where C_i, K_i *respectively denote the capacity of the virtual link available for the transmission and the* ***destination*** *buffer space available for the object represented* ***by*** *place* p_i.

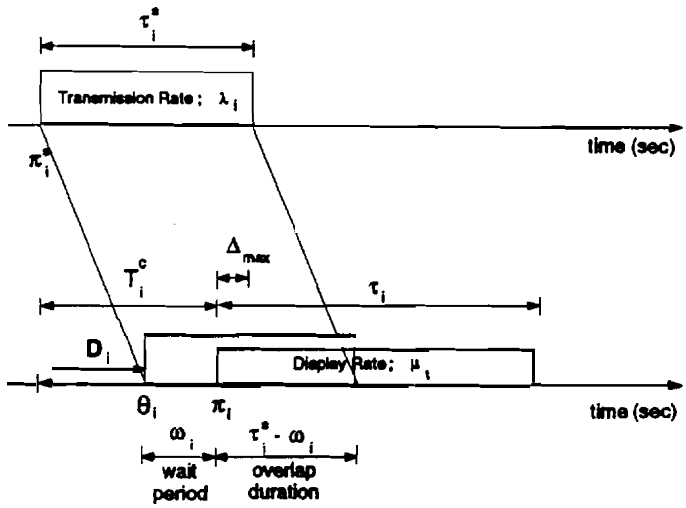


Figure 2: Timing diagram for Object Synchronization

2.2 Synchronization Consideration Due to Network Delays

For synchronization of multimedia objects over the network, we use our previously proposed prescheduling strategy [10], which requires that for an object O_i to meet its deadline π_i, sufficient control time (T_i^c) must be allowed to overcome the end-to-end. delay ($\mathbf{D}_i$) from the transmitter to the destination. If network delays are deterministic, then it is trivial to ensure safe scheduling. However, network and jitter delays are random and introduce unpredictable latencies in the playback of media objects. This is undesirable, especially for continuous media presentation where data must arrive at the destination at an almost constant sate. Even in the case of discrete media, due to the synthetic temporal relationships, a fixed arnount of data must arrive at the destination at some specified time. Hence, it is necessary to choose T_i^c to be at least equal to the end-to-end delay, $\mathbf{D}_i$, such that scheduling of an object for transmission at time, $\pi_i^s = \pi_i - T_i^c$ guarantees successful synchronization as illustrated in Fig. 2.

The total end-to-end delay $\mathbf{D}_i$ has the following main components:

- Propagation delay (D_i^{prop}): It represents the time it take energy to move from the

source to the destination (also called the latency of the channel). If we assume that the transfer of information take place at the speed of light (C), then this delay can be approximated by

$$D_i^{prop} = L_i^{sd}/C$$

Where, L_i^{sd} represents the distance between the source of the object O_i and its destination. Clearly, this delay is constant for a given object.

- Transmission delay (D_i^{trans}): This denotes the time it take to pump one unit of consumable information into the link. We define the consumable information as an entity that can be played out at the destination, e.g. one video frame, an image. Let, s_i^{min} is the minimum size of one unit of consumable information and λ_i is the transmission rate., then

$$D_i^{trans} = \frac{s_i^{min}}{\lambda_i}$$

Hence, the transmission delay depends on media type of the object and the available channel capacity.

- Jitter delay (D_i^j): This variable portion of the end-to-end delay, is caused by the queueing within network buffers which resolve the output contention occurring in switching and multiplexing stages. Let $F_{D_i^j}(\cdot)$ denotes its distribution function. Several results have been reported (e.g, [3]-[16]) stating approximate close form expression of $F_{D_i^j}(\cdot)$ in virtual circuits.

Hence, the total end-to-end delay (D_i) can be characterized as

$$D_i = D_i^{prop} + D_i^{trans} + D_i^j$$

Let $F_{\mathbf{D}_i}(\cdot)$ is the distribution function of $\mathbf{D}_i$. As the components, D_i^{prop} and D_i^{trans} constitutes constant delays for a given object, therefore

$$F_{\mathbf{D}_i}(\xi) = F_{\mathbf{D}_i^j}(\xi - (D_i^{trans} + D_i^{prop}))$$

2.3 The Rate Model of Presentation Process

The rate model is based on the fluid flow assumption that the information flow in and out of the destination buffers is uniform rather than in discrete packets/ cells. The fluid assumption generally provides accurate results for the case when the packet size is small as compared to the transmission rate [8]. Fortunately this is true for ATM based integrated,networks as these networks can operate at rates which approach gigabits/sec, while they have fixed size cells that are only 53 octets long [2]. Hence, it is possible to ignore the discrete nature of the data and treat; it as a continuous bit stream [1]. Based on fluid flow assumption, we can divide the overall process of object retrieval and presentation into three distinct processes, namely; transmission process, arrival process and consumption process. Each process conserves the fluid involved in the presentation.

2.3.1 The Transmission Process

We model the transmission of the object O_i as a continuous bit stream flowing at a constant rate, λ_i, such that $\lambda_i \leq C_i$. Hence, the transmission process, $t_i(t, \pi_i)$ can be described by the following equation.

$$t_i(t, \pi_i) = \begin{cases} \lambda_i(t - \pi_i^s) & \text{if } \pi_i^s \leq t \leq \pi_i^s + \tau_i^s \\ 0 & \text{otherwise} \end{cases}$$

where τ_i^s denote the object transmission duration and π_i^s is the start transmission time for the object. Object transmission duration depends on its size and the transmission rate $(\tau_i^s = s_i/\lambda_i)$. The transmission process is represented by a line $(t_i(t, \pi_i))$ in Fig. (3) which

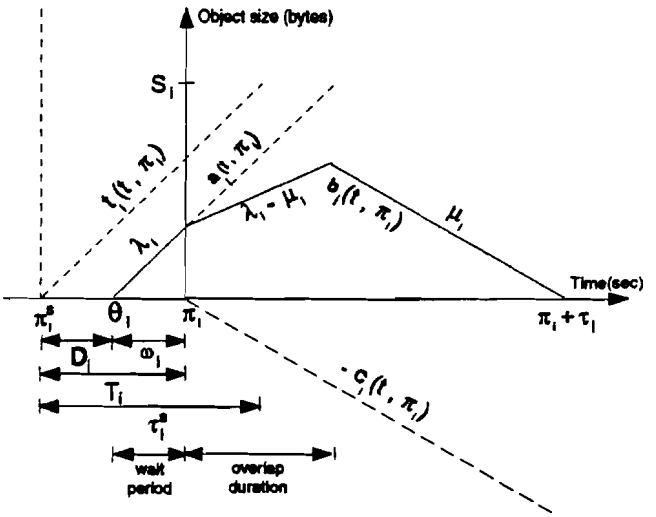


Figure 3: Buffer occupancy diagram for the object O_i.

indicates that the transmission of the object starts at some time π_i^s and it continues for the duration τ_i^s at a constant rate λ_i.

2.3.2 The Arrival Process

The arrival process, $\mathbf{a}_i(t, a_i)$, describes the stochastic behavior of arrival of the traffic, generated by the transmission process, at the destination. The arrival process depends on the transmission process and the end-to-end delay distribution. In our analysis, we assume that the input traffic is controlled by the adequate traffic rate control mechanisms. Under this assumption, it is known **[6]** that the end-to-end delay jitter is in the range of the maximum transfer delay of one queueing node on the channel. Hence, we can consider that the end-to-end jitter distribution is identical for all the subsequent cells of O_i (as argued in [12]) and the object arrival rate at the destination is the same as its transmission rate. Note, we ***do not*** assume the same delay distribution for all objects, as we consider heterogeneous channels for the transmission of the different objects. These assumptions, along with the fluid assumption, are needed to make the analysis tractable. Under these assumptions, the

arrival process $\mathbf{a}_i(t, \mathrm{R};)$ can be described by the following stochastic fluid flow equation:

$$\mathbf{a}_i(t, \pi_i) = \begin{cases} \lambda_i(t - \pi_i + \omega_i) & \text{if } \pi_i - \omega_i \leq \mathrm{t} \leq \pi_i - \omega_i + \tau_i^s \\ 0 & \text{otherwise} \end{cases}$$

where ω_i denotes the random wait time for O; at the destination before its presentation can start, and is given by

$$\omega_i = T_i^c - \mathbf{D}_i \qquad ; \omega_i \in \mathbb{R}$$

For the case when $\omega_i < 0$, destination has to wait for O; to arrive. This results in "starvation" at the destination. Without loss of generality, we assume that for this case, some default presentation continues until the object is received.

The arrival process for the case when w; $>$ 0 is shown in Fig. (3) by the line $\mathbf{a}_i(t, \pi_i)$ which indicates that the arrival of O_i at the destination starts at some random time θ_i and is maintained at a constant rate thereafter for the duration τ_i^s.

2.3.3 The Consumption Process

The consumption process represents the outflow of O_i from the destination buffer for its presentation. We model this process as a constant rate process for CM objects, therefore if $c_i^{CM}(t, \pi_i)$ denotes consumption of the CM object O;, then

$$c_i^{CM}(t, \pi_i) = \begin{cases} \mu_i(t - \pi_i) & \text{if } \pi_i \leq t \leq \pi_i + \tau_i \\ 0 & \text{otherwise} \end{cases}$$

Here μ_i represents the rate of consumption of O_i. This process is shown in Fig. (3) by the line $C_i(t, \pi_i)$. For DM objects, we can view this process as an instantaneous process because the whole object is a single entity. Therefore, we model consumption rates for DM objects as a Dirac delta function,

$$c_i^{DM}(\mathrm{t}, \pi_i) = \delta(\pi_i)$$

Formally, we define a consumption rate function f_i associated with place p; in an **ROCPN** as:

$$f_i = \begin{cases} \mu_i & \text{if } y(p_i) \equiv CM \\ \delta(\pi_i) & \text{if } y(p_i) \equiv DM \end{cases} \quad (1)$$

2.3.4 The Buffer Occupancy Function

We define the buffer occupancy function for O; as the difference of the arrival function and the consumption function, i.e.,

$$\mathbf{b}_i^j(t, \pi_i) = \mathbf{a}_i(t, \pi_i) - c_i^j(t, \pi_i) \quad \forall t, \forall j \quad ; \mathbf{b}_i^j(t, \pi_i) \in \mathbb{R}$$

Again, $\mathbf{b}_i^j(t, \pi_1) < 0$ represents the case of starvation at the destination.

Fig. (3) shows the buffer occupancy function $\mathbf{b}_i(t, \pi_i)$ when $0 \leq w; \leq \tau_i^s$. Since $\theta_i \leq \pi_i$, the object has to wait in the destination buffer which results in a linear increase in $\mathbf{b}_i(t, \pi_i)$ at rate A; up to the playout start time of the object, π_i. As $w; \leq \tau_i^s$, i.e. the random wait time is less than the transmission duration of the object, therefore for a duration defined by $[\pi_i, \pi_i + \tau_i^s - \omega_i]$, the object arrival and consumption processes overlap. Hence for this period, $\mathbf{b}_i(t, \pi_i)$ increases at a rate equal to difference in object arrival and consumption rates. For the rest of the presentation duration, the contents of the object already stored in the buffer are consumed at rate μ_i.

From now on we will drop the superscript for continuous media object.

3 Buffering Requirements at the Destination

In this section, we find the minimum and the maximum buffer required at the destination for an uninterrupted presentation of the object O_i. A major conclusion of this analysis is that the transmission rate and control time dictate the buffering requirement at the destination. We consider the case of channel sufficient system [15], hence throughout our discussion we assume that $(C_i \leq A; \leq \mu_i)$.

3.1 Minimum Buffering Requirement

We first consider the case of CM, for which buffering is required before the start of its presentation in order to reduce the jittering effect. Additional buffering is also needed if there is any asynchrony in arrival and consumption processes. The minimum buffer requirement largely depends on channel utilization. An increase in this utilization results in a need for more buffering. This dependence is described in the following theorem.

Theorem 3.1 *For an uninterrupted presentation of a CM object O_i, the minimum buffer space required, b_i^{min}, is given* by

$$b_i^{min} = (1 - \rho_i)s_i = (1 - \frac{1}{\eta_i})s_i \quad ; \eta_i \geq 1$$

Where ρ_i is the ratio of the consumption rate to the transmission rate for O_i and $\eta_i = \frac{1}{\rho_i}$ is the channel utilization factor.

Proof: For successful presentation of the CM object O;, minimum buffer space is required for the case when $\omega_i = 0$, i.e. when no buffering is required for jitter compensation. Under this ideal situation, maximum overlap equal to the object transmission duration (τ_i^s) occur between object arrival and its consumption processes. Hence, a part of the: object, $\mu_i \tau_i^s$ can be presented without buffering. Thus, the minimum buffer space required at the destination is given by

$$\begin{aligned} b_i^{min} &= s_i - \mu_i \tau_i^s \\ &= (1 - \frac{\mu_i}{\lambda_i})s_i \\ &= (1 - \rho_i)s_i \\ &= (1 - \frac{1}{\eta_i})s_i \end{aligned}$$

∎

This theorem provides a lower bound on the buffer size needed to compensate for the difference in the arrival and the display rates. Later in the report (Section (5)), we provide a better bound on the buffer size as we consider the QOP parameters.

For a DM object, we need to buffer the whole object before it can be presented. Therefore, for this case $b_i^{min,DM} = s_i$.

3.2 Maximum Buffering Requirement

For DM objects, the maximum and the minimum buffering requirements remain the same and is given by $b_i^{min,DM}$, which is equal to the size of the object. However, for the case of CM object, buffering of the complete object is neither required nor desirable as these objects typically involve huge amount of data which can be continuously consumed. For this data the expected size of the maximum buffer depends on the selection of the control time, T_i^c. In this section, we show that a *proper* selection of T_i^c can result in a significant reduction in the destination buffering requirement.

The maximum buffer size consists of b_i^{min} and an additional space to reduce the jittering effect. This is stated in the following theorem.

Theorem *3.2* *The buffer space requirement at the destination for the CM object is a random variable, with expected value given by*

$$\begin{aligned} E[\mathbf{b}_i^{max}] &= (1-\rho_i)s_i + \bar{\gamma}_i F_{\mathbf{D}_i}(T_i^c)s_i \\ &= (1-\rho_i)s_i + \bar{\gamma}_i(1-\mathcal{P}_i^d)s_i \end{aligned} \tag{2}$$

Where γ_i denotes the random fraction of the time for which the object has to wait at the destination buffer before its presentation can begin.

$$\gamma_i \triangleq \frac{\textit{Random wait time}}{\textit{Total life of the CM object}} = \frac{\omega_i}{\tau_i}$$

This can be interpreted as the jitter compensator *to reduce the jittering effect over the presentation. $\bar{\gamma}_i$ denotes the expected value of the jitter compensator.*

$$\bar{\gamma}_i \triangleq E[\gamma_i] = \frac{T_i^c - \bar{D}_i}{\tau_i}$$

$\mathcal{P}_i^d$ represents the tolerable probability of buffer overflow. and $F_{\mathbf{D}_i}(\cdot)$ is the end-to-end delay distribution function.

Proof: Consider the following cases:

Case 1. $0 \leq \omega_i \leq \tau_i^s$

The buffer occupancy function for this case is plotted in Fig. **3,** and is given by

$$b_i(t,\pi_i) = \begin{cases} \lambda_i(t - \pi_i + w;) & \text{if } \pi_i - \omega_i \leq t \leq \pi_i \\ (\lambda_i - \mu_i)(t - \pi_i) + \lambda_i\omega_i & \text{if } \pi_i \leq t \leq \pi_i - \omega_i + \tau_i^s \\ s_i - \mu_i(t - \pi_i) & \text{if } \pi_i - \omega_i + \tau_i^s \leq t \leq \pi_i + \tau_i \\ 0 & \text{otherwise} \end{cases}$$

As $\lambda_i > 0$, $\mathbf{b}_i(t,\pi_i)$ is strictly monotonically increasing in $[\pi_i - \omega_i, \pi_i]$, and also in $[\pi_i, \pi_i - \omega_i + \tau_i^s]$. As $\mu_i > 0$, the buffer accumulation function is strictly monotonically decreasing over $[\pi_i - \omega_i + \tau_i^s, \pi_i + \tau_i]$. Furthermore, $\mathbf{b}_i(t,\pi_i)$ is continuous at $t = \pi_i - w; + \tau_i^s$ as both left and right hand side limits exists and are equal to value of function at this point. Hence, it has a unique maximum value, $\mathbf{b}_i^{max}$ in $[\pi_i - \omega_i, \pi_i + \tau_i]$ where

$$\begin{aligned} by^{a2} &= \mathbf{b}_i(\pi_i - \omega_i + \tau_i^s, \pi_i) \\ &= (\lambda_i - \mu_i)(\tau_i^s - \omega_i) + \lambda_i\omega_i) \\ &- \lambda_i\tau_i^s - \mu_i(\tau_i^s - \omega_i) \\ &= s_i - \mu_i(\tau_i^s - \omega_i) \\ &- s_i - \$5; \ + \mu_i\omega_i \\ &= (1 - \rho_i)s_i + \mu_i\omega_i \end{aligned}$$

If γ_i denotes the fraction of the time for which the object waits in destination buffers,

$$\gamma_i \triangleq \frac{\omega_i}{\tau_i}$$

then

$$\mathbf{b}_i^{max} = (1 - \rho_i)s_i + \gamma_i s_i$$

Case 2. $\omega_i \geq \tau_i^s > 0$

This is the case when we receive the whole object prior to its deadline and thus, $\mathbf{b}_i(t,\pi_i)$ is given by

$$\mathbf{b}_i(t,\pi_i) = \begin{cases} \lambda_i(t - \pi_i + \omega_i) & \text{if } \pi_i - w; \leq t \leq \pi_i - w; + \tau_i^s \\ s_i & \text{if } \pi_i - \omega_i + \tau_i^s \leq t \leq \pi_i \\ s_i - \mu_i(t - \pi_i) & \text{if } \pi_i \leq t \leq \pi_i + \tau_i \\ 0 & \text{otherwise} \end{cases}$$

Using similar arguments, as in case 1, it can be shown that the maxima of $\mathbf{b}_i(t, \pi_i)$ occurs at $\pi_i - \omega_i + \tau_i^s$, and

$$\begin{aligned} \mathbf{b}_i^{max} &= \mathbf{b}_i(\pi_i - \omega_i + \tau_i^s, \pi_i) \\ &= s_i \end{aligned}$$

Case 3. $-(\tau_i - \tau_i^s) \leq \omega_i \leq 0$

In this case, the arrival of the object starts after the expiration of its deadline but the whole object is received during the playout of the object. Without loss of generality, we consider the case when the presentation starts as soon as the object is available and continue at the same rate till its expected end-of-play. Under this assumption, the buffer accumulation is given by

$$\mathbf{b}_i(t, \pi_i) = \begin{cases} (\lambda_i - \mu_i)(t - \pi_i + \omega_i) & \text{if } \pi_i - \omega_i \leq t \leq \pi_i - \omega_i + \tau_i^s \\ s_i - \mu_i(t - (\pi_i - \omega_i)) & \text{if } \pi_i - \omega_i + \tau_i^s \leq t \leq \tau_i - \omega_i \\ 0 & \text{otherwise} \end{cases}$$

We can show that the maxima of $\mathbf{b}_i(t, a;)$ occurs at $\pi_i - \omega_i + \tau_i^s$, *and*

$$\begin{aligned} \mathbf{b}_i^{max} &= \mathbf{b}_i(\pi_i - \omega_i + \tau_i^s, \pi_i) \\ &= (\lambda_i - \mu_i)\tau_i^s \\ &= s_i - s_i\frac{\mu_i}{\lambda_i} \\ &= (1 - \rho_i)s_i \end{aligned}$$

Case 4. $-(\tau_i - \tau_i^s) \leq \omega_i \leq -\tau_i$

Under the same assumption as in case 3, the buffer accumulation function increases at the rate $(A; -\mu_i)$ for the duration $\tau_i - |\omega_i|$. Hence,

$$\begin{aligned} \mathbf{b}_i^{max} &= (\lambda_i - \mu_i)(\tau_i - \omega_i) \\ &= (1 - \gamma_i)\frac{1-\rho_i}{\rho_i}s_i \end{aligned}$$

Case 5. $w; \leq -\tau_i$

In this case, the whole object is missed, therefore, $\mathbf{b}_i^{max} = 0$.

Using the expression for the $\mathbf{b}_i^{max}$ found above, the expected value of maximum buffer required can be found.

$$\begin{aligned} E[\mathbf{b}_i^{max}] &= s_i\mathcal{P}\{\omega_i \geq \tau_i^s\} + [(1 - \rho_i)s_i + \bar{\gamma}_i s_i]\mathcal{P}\{0 \leq \omega_i \leq \tau_i^s\} + \\ &\quad (1 - \rho_i)s_i\mathcal{P}\{-(\tau_i - \tau_i^s) \leq \omega_i \leq 0\} + (1 - \bar{\gamma}_i)\frac{1-\rho_i}{\rho_i}s_i\mathcal{P}\{-(\tau_i - \tau_i^s) \leq \omega_i \leq -\tau_i\} \end{aligned}$$

Manupulating the above expression, neglecting terms like $\mathcal{P}\{\omega_i \leq -\tau_i\}$ and substituting the required probabilities in terms of end-to-end delay distribution, we can gel;

$$E[\mathrm{b}_i^{max}] \;=\; (1-\rho_i)s_i + \bar{\gamma}_i F_{\mathbf{D}i}(T_i^c)s_i$$

∎

In equation (2), the first term represents b_i^{min} while the second term denotes the buffering required for jitter compensation. It can be noticed that the buffering required to reduce the jittering effect due to non-deterministic delays in the network on the presentation process depends on the end-to-end delay distribution and can be controlled by selecting an appropriate value for T_i^c. For a properly selected T_i^c, the tail of the jitter distribution function $(F_{\mathbf{D}_i^j}(\cdot))$ determines the buffering requirement (distribution function with sharp tail implies less buffering). Hence, networks with low jitter variance are desirable.

If jitter delays are not properly compensated for, the deadline misses for the objects can occur. On the other hand, over compensation requires more buffering. Therefore, there exists a trade-off between buffer utilization and the QOP. This trade-off is discussed in the following section.

3.3 Performance Trade-offs

In this section we elaborate the effect of T_i^c on one of the QOP parameters, $\mathcal{P}_i^d$, and the required buffering at the destination. We consider an example of multimedia presentation of a video clip consisting of 18 Gbit, which is retrieved over the network, operating at 300 Mbit/sec. We assume that the end-to-end delay jitter has a Gaussian distribution, an assumption supported in [12], [13], [10]. In order to estimate the average jitter delays, we assume that each switching node over the virtual channel is an $M/M/1$ queue with utilization factor, $\hat{\rho} = 0.50$ [7]. With transmission speed λ_i = 300 Mbits/sec, one unit of consumable information (one video frame of size s_i^{min} = 10 Mbit) experiences a jitter delay of the order of

$\frac{s_i^{min}}{\lambda_i(1-\rho)} \cong 60.0$ milliseconds. Assuming that on the average 5 switching nodes are encountered on this channel and all have identical traffic environment, the expected end-to-end delay can be of the order of 300 milliseconds. Such low end-to-end delays are realistic in ATM networks as they are expected to provide high throughput with low end-to-end delay [5]. We further assume that the variance of the delay distribution is 10 milliseconds Assuming an end-to-end distance of 3000 miles, the propagation of data at the speed of light yields a propagation delay of 5 milliseconds. Note that these assumptions are made for numerical results only. Our analysis is not restricted to any particular network environment.

For this example, the effect of T_i^c on the buffering requirement at the destination is shown in Fig. 4(a). For the purpose of illustration, this curve is divided into three regions (for the case when $\rho_i = 0.5$). It can be seen that selecting an arbitrary large control time (T_i^c) increases the buffering requirement at the destination, i.e., as $T_i^c \rightarrow \infty$, $E[b_i^{max}] \rightarrow s_i$, which is the size of the whole object. As can be noticed, for a smaller control time: ($T_i^c < D$;, region R_1) the required buffering is $(1 - \rho_i)s_i$ and is only needed to compensate for the difference between the transmission and the consumption rates. This is because, in this case, the object does not have to wait in the destination buffer for its playout to start. Therefore, both the arrival and the consumption processes overlap during the complete transmission duration (τ_i^s) of the object. If we choose small T_i^c, the "buffering effect" of the network helps. However, if the network transmission rate increases , this buffering effect diminishes, as we can notice from Fig. 4(a) that an increase in channel utilization increases the demand for destination buffer space. For the case when $\rho_i \neq 1$, the part of object given as $(1-\rho_i)s_i$ needs to be buffered to compensate for the rate differences between the arrival and consumption processes.

For a given transmission rate, if T_i^c is increased beyond $\bar{D}_i$ (region R_2), then we observe an almost linear increase in buffering requirement. This observation is consistent with our intuition as larger control time means early scheduling of transmission of O_i and hence an

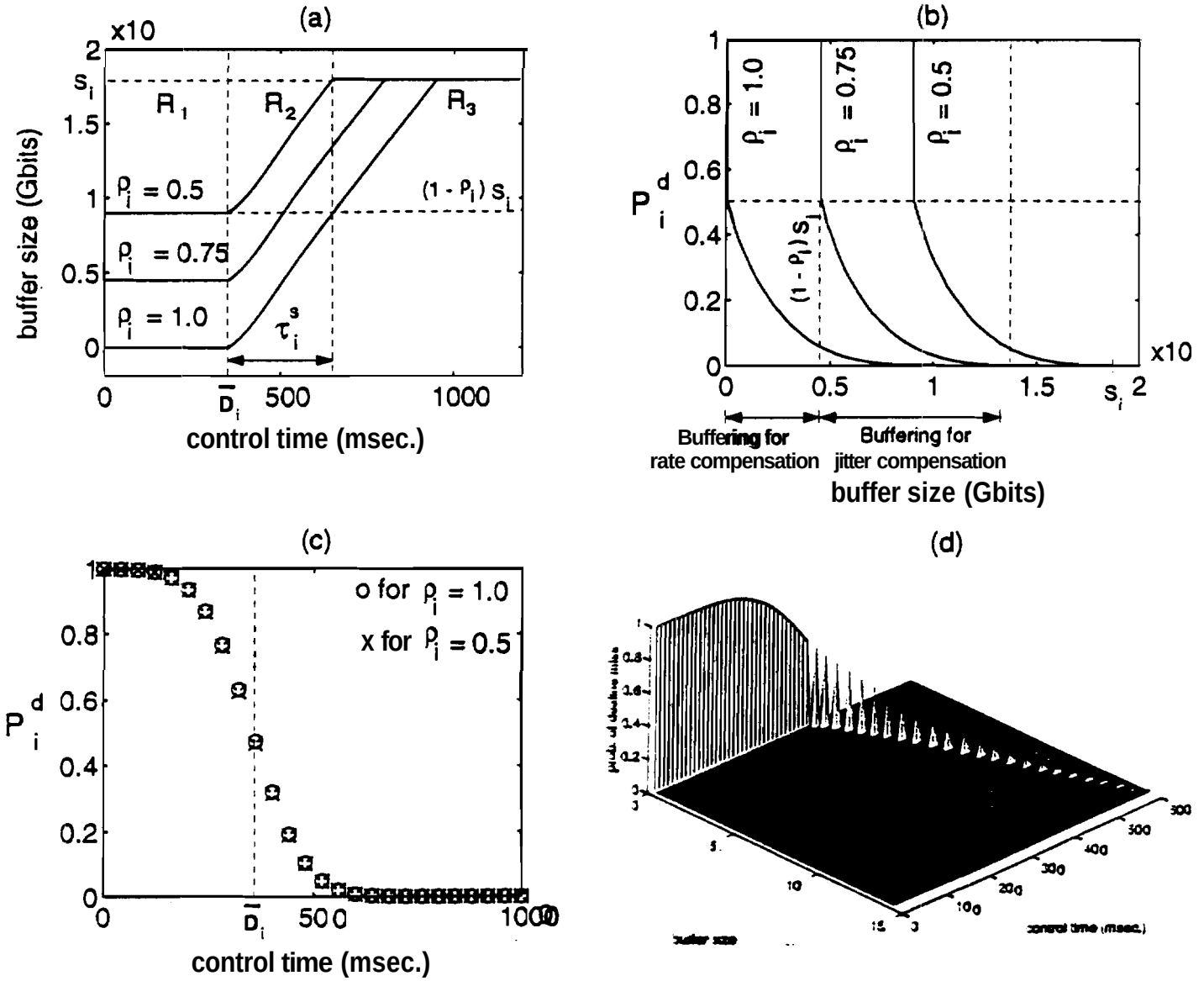


Figure 4: Trade-offs between destination buffer requirements and the QOP

early arrival of the object. Therefore, the part of the object arrived prior to its deadline needs to be buffered. In region R_3, when $T_i^c \geq \tilde{D}_i + \tau_i^s$, excessive pre-fetching upto the size of the whole object, necessitates a buffer size which is equal to the size of the object. Thus scheduling the object with such T_i^c does not take any advantage of the network buffering effect. Note that the transmission duration may be fairly large as $\tau_i^s = \rho_i \tau_i$, hence by a "proper" selection of T_i^c we can reduce the buffering requirements by a huge amount especially for objects with longer duration.

Fig. (4 (b)) indicates that in order to reduce $\mathcal{P}_i^d$, we must pre-fetch the data earlier enough which in turn increases our buffering needs. However, it is important to note that even for arbitrary low values of $\mathcal{P}_i^d$, we are not required to buffer the whole object. This is because of the fact that increasing control time beyond some value (this value depends on the mean and the variance of the end-to-end delay) does not result in significant decrease in $\mathcal{P}_i^d$ (see Fig. 4 (c)). Hence, with a proper selection of T_i^c it is possible to reduce the destination buffering requirements by a significant amount. Fig. 4 (c) provides the desired optimal value for the control time for a given value of $\mathcal{P}_i^d$. We notice that for a given T_i^c, an increase in channel utilization does not have any significant effect on the $\mathcal{P}_i^d$. Fig. (4(d)) summarizes these trade-offs.

4 Object Scheduling: QOP Considerations

In this section we consider the case when resources including channel utilization and buffer size at the destination are fixed and determine the best value of T_i^c such that the specified values of QOP parameters $(\mathcal{P}_i^d, \mathcal{P}_i^b)$ can be guaranteed. Through this analysis, we establish an upper bound on T_i^c for a given value of $\mathcal{P}_i^b$ and fixed buffering capabilities at the destination. We also provide a lower bound on T_i^c which guarantees that the resulting miss probability is less than the specified $\mathcal{P}_i^d$. Subsequently, we identify a condition for schedulability of objects

in terms of network resources $(K;, C_i)$ and the **QOP.**

4.1 Destination Buffering Consideration

For a fixed buffer size at the destination, we need to avoid excessive pre-fetching of the object prior to its deadline, otherwise buffer overflow can occur. This in turn bounds T_i^c as stated in the following theorem.

Theorem 4.1 *In order to bound the probability of buffer overflow at the destination within a specified limit $(\mathcal{P}_i^b)$, T_i^c should satisfy the following inequality*

$$T_i^c \leq F_{\mathbf{D}_i}{}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1-\rho_i)\tau_i \tag{3}$$

where $k_i = K_i/s_i$ denotes the fraction of the object that can be buffered at the destination.

Proof: In theorem (3.2), we found

$$E[\mathbf{b}_i^{max}] = (1-\rho_i)s_i + \bar{\gamma}_i(1-\mathcal{P}_i^d)s_i$$

To avoid buffer overflow, we would like to bound $E[\mathbf{b}_i^{max}]$ by some fraction of the maximum available buffer space, $K;$, i.e.,

$$\begin{aligned} & (1-\rho_i)s_i + \gamma_i(1-\mathcal{P}_i^d)s_i \leq r_i K_i \\ \Rightarrow\ & (1-\rho_i) + \tfrac{T_i^c - \bar{D}_i}{\tau_i}(1-\mathcal{P}_i^d) \leq r_i k_i \\ \Rightarrow\ & T_i^c \leq \bar{D}_i + \tfrac{1}{1-\mathcal{P}_i^c}[r_i k_i - (1-\rho_i)]\tau_i \end{aligned}$$

Where, $0 \leq r_i \leq 1$ is the design factor whose value must be dictated by the maximum tolerable probability of buffer overflow, $\mathcal{P}_i^b$. For an arbitrary low $\mathcal{P}_i^d$, the upper bound on control time can be found from the following equation

$$T_i^c \leq \bar{D}_i + [r_i k_i - (1-\rho_i)]\tau_i \tag{4}$$

Next, we need to find some suitable value for the design factor, r_i for a given probability of buffer overflow. Note that

$$\begin{aligned} \mathcal{P}_i^b &= \mathcal{P}\{\mathbf{b}_i^{max} \geq K_i\} \\ \Rightarrow\ \mathcal{P}_i^b &= \mathcal{P}\{(1-\rho_i)s_i + \tfrac{\omega_i}{\tau_i} si \geq K_i\} \\ &= \mathcal{P}\{\mathbf{D}_i \leq T_i^c - [k;-(1-\rho_i)]\tau_i\} \end{aligned}$$

For largest value of control time from (4), we can write

$$\begin{aligned} \mathcal{P}_i^b &= F_{\mathbf{D}_i}(\bar{D}_i - (1 - r_i)k_i\tau_i) \\ \Rightarrow \quad r_i &= 1 - \frac{\bar{D}_i - F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b)}{k_i\tau_i} \end{aligned} \tag{5}$$

If $F_{\mathbf{D}_i}^{-1}$ denotes the inverse distribution function for the end-to-end delay then

$$r_i = 1 - \frac{\bar{D}_i - F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b)}{k_i\tau_i}$$

Now from (4) we found the following bound on the control time T_i^c

$$T_i^c \leq F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \rho_i)\tau_i$$

■

In Equation (3), the term $k_i\tau_i$ represents the portion of the total playout duration of the object for which the object is already stored in the buffer. $F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b)$ denotes the relaxation on control time for tolerating buffer overflow. The term $(1 - \rho_i)\tau_i$ represents the effect of the channel utilization on T_i^c. Note that for a given buffer size, higher channel utilization requires small T_i^c in order to maintain the desired $\mathcal{P}_i^b$.

For a DM object, as mentioned earlier, we need to buffer the entire object. Hence for a given buffer with size equal to the size of the object, there is no upper bound on the control time. This is true for the case of CM object also if such buffering capabilities at the destination are available.

4.2 Missing Deadline Consideration

As mentioned above, in order to reduce the destination buffering requirements, transmission of the object with an arbitrary large control time is not practical. Hence, we would like to find the earliest possible schedulable time which satisfies the given $\mathcal{P}_i^d$. Such T_i^c is given by the following theorem.

Theorem 4.2 *In order to ensure that the $\mathcal{P}_i^d$ remains bounded within the specified limit, the control time must satisfy the following inequality*

$$T_i^c \geq F_{\mathbf{D}_i}^{-1}(1 - \mathcal{P}_i^d) \tag{6}$$

Proof: If $\lambda_i \geq \mu_i$, then the probability of deadline misses is equivalent to probability that object arrival time (θ_i) is greater than the deadline of presentation of object. In other words,

$$\begin{aligned} \mathcal{P}_i^d &= Pr\{\theta_i > \pi_i\} \\ &= Pr\{\omega_i < 0\} \\ &= Pr\{T_i^c - \mathbf{D}_i < 0\} \\ &= Pr\{\mathbf{D}_i > T_i^c\} \\ &= 1 - F_{\mathbf{D}_i}(T_i^c) \end{aligned}$$

∎

In other words, more strict requirements on deadline misses force control time to be larger than some minimum value given in Equation (6). This restriction on T_i^c need to be observed for both CM and DM objects. Note that the channel utilization does not have any significant influence (expect a small change in transmission delays) on this lower bound.

4.3 Schedulability Condition and Schedulable Region

Since, a ,given $\mathcal{P}_i^d$ forces a lower bound on T_i^c, while a constrain on $\mathcal{P}_i^b$ does not allow T_i^c to exceed some value, there exits a range of T_i^c which can guarantee the desired QOP. This section establishes a condition for the schedulability and discusses how different parameters including, $\mathcal{P}_i^d$, $\mathcal{P}_i^b$, k_i, and the transmission link utilization (η_i) effects this condition. This condition is given by the following theorem.

Theorem 4.3 *Given a transmission rate λ_i , a finite buffer space $K_i \leq s$; at the receiver, and the desired values of $\mathcal{P}_i^b$ and $\mathcal{P}_i^b$, the object O_i is schedulable iff*

$$F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \tfrac{\mu_i}{\lambda_i})\tau_i \geq F_{\mathbf{D}_i}^{-1}(1 - \mathcal{P}_i^d) \tag{7}$$

Proof: $\Rightarrow$: We define that an object is schedulable if we can find somt: control time T_i^c such that transmission of object at time $\pi_i - T_i^c$ guarantees that presentation at the receiver will satisfy the specified values of QOP parameters. Such T_i^c needs to satisfy conditions (3) and (6), i.e.,

$$F_{\mathbf{D}_i}^{-1}(1 - \mathcal{P}_i^d) \leq T_i^c \leq F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \tfrac{\mu_i}{\lambda_i})\tau_i \quad \forall \quad \mu_i \leq \lambda_i \leq C_i$$

$\Leftarrow$: If condition in Equation (7) is satisfied then we can find some control time, T_i^c such that

$$F_{\mathbf{D}_i}^{-1}(1 - \mathcal{P}_i^d) + \leq T_i^c \leq F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \tfrac{\mu_i}{\lambda_i})\tau_i$$

The control time found can be used to generate a schedule which can meet the desired QOP. ∎

The inequality in this theorem can be used to identify the **schedulability** region, in a three dimensional space spanned by the tuple $(\mathcal{P}_i^b, \mathcal{P}_i^d, T_i^c)$ for which the transmission of O_i is feasible. These regions are shown in Fig. (5), for the example used in Section (3.3). Each plot shows two surfaces, an upper surface $\mathcal{S}_1$, defined by the equation $T_i^c = F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \frac{\mu_i}{\lambda_i})\tau_i$ and a lower surface $\mathcal{S}_2$, defined by $T_i^c = F_{\mathbf{D}_i}^{-1}(1 - \mathcal{P}_i^d)$. The values for tuple $(\mathcal{P}_i^b, \mathcal{P}_i^d, T_i^c)$ that lie in region defined by the intersection of the $\mathcal{S}_1$ and the $\mathcal{S}_2$ (above $\mathcal{S}_2$ and below $\mathcal{S}_1$) correspond to a feasible transmission schedule. The values of the QOP parameters in an unschedulable region cannot be guaranteed (for the given set of resources). The size of the schedulable region represents the flexibility in choosing arbitrary values for QOP parameters. This size increases with an increase in the destination buffer size and decreases with an increase in the channel utilization. For example, with an increase in destination buffer size [Fig. 5 (b)], an object with a given QOP which was unschedulnble in [Fig. 5 (a)] is becomes schedulable. For a given buffer size, the size of a schedulable region decreases,

due to an increase in transmission link utilization [Fig.5 (c)]. This provide a trade-off in terms of the QOP, the destination buffering and the channel utilization.

For the case of discrete media the whole region above the surface S_2 [Fig. 5 (d)] is schedulable because, as mentioned earlier, such an object needs to be buffered prior to its presentation. This is also true for a CM object with $k_i = 1$.

In Section (5) we discuss how to choose "optimal" control time which can not only guarantee the desired QOP but also minimizes the resource requirements.

5 Optimal Control Time for Minimum Resource Requirements to Guarantee QOP

In order to maintain the desired quality of presentation at the destination, it is important that some minimum amount of system resources (buffer at the destination and channel capacity) should be dedicated. In this section, we find bounds on the resources needed to satisfy the desired QOP. We also find the crossponding control time to guarantee the desired QOP as well as to minimize resource requirements.

5.1 Minimum Buffer Requirement (Fixed Channel Utilization)

Theorem (3.1), provides a loose bound on the buffer space required to support real-time presentation of an object at the destination. This bound only takes into account the buffering required to compensate for asynchrony in the transmission and consumption processes. Some additional buffering is needed in order to satisfy the QOP parameters. The following theorem states a tighter bound on the minimum buffer requirement under new considerations.

Theorem *5.1 For a given transmission rate λ_i and a desired QOP, the minimum buffer required at the destination is*

$$\begin{aligned} K_i^{min} &= (1-\rho_i)s_i + \frac{F_{D_i}^{-1}(1-\mathcal{P}_i^d) - F_{D_i}^{-1}(\mathcal{P}_i^b)}{\tau_i} s_i \\ or \quad k_i^{min} &= (1-\rho_i) + \frac{F_{D_i}^{-1}(1-\mathcal{P}_i^d) - F_{D_i}^{-1}(\mathcal{P}_i^b)}{\tau_i} \end{aligned} \tag{8}$$

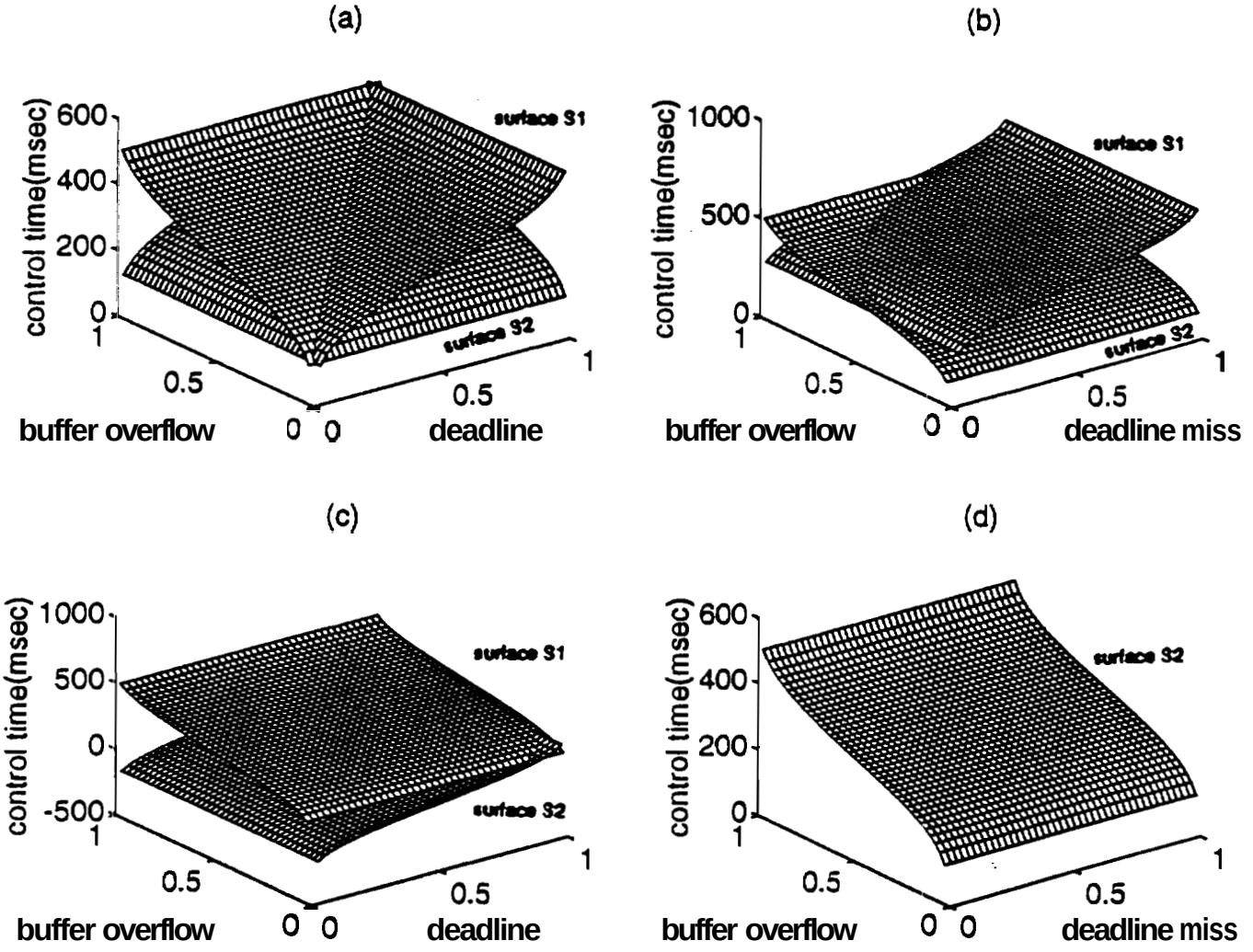


Figure 5: Schedulable regions (a). $\rho_i = 1; k_i = 0$, (b). $\rho_i = 1; k_i = 0.25$, (c). $\rho_i = 0.5; k_i = 0$, (d). $\rho_i = 0.5; k_i = 1$.

The control time that can achieve this value and the desired QOP is

$$T_i^{c,opt} = F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i^{min}\tau_i - (1-\rho_i)\tau_i \tag{9}$$

Proof: For a given transmission rate (λ_i) and desired QOP, theorem (4.3) states the following condition for schedulability

$$\begin{aligned} & F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1-\tfrac{\mu_i}{\lambda_i})\tau_i \;\geq\; F_{\mathbf{D}_i}^{-1}(1-\mathcal{P}_i^d) \\ \Rightarrow \quad & k_i \;\geq\; (1-\rho_i) + \tfrac{F_{\mathbf{D}_i}^{-1}(1-\mathcal{P}_i^d) - F_{\mathbf{D}_i}^{-1}(\mathcal{P}_i^b)}{\tau_i} \end{aligned}$$

∎

According to this theorem, if the available buffer space at the destination is at least K_i^{min}, and the object transmission rate is λ_i, then scheduling the object with T_i^{opt} guarantees the desired QOP. Fig.(6(a)) and Fig.(6(b)) shows this minimum buffer for various values of QOP parameters for the case of $\rho_i = 1$ and $\rho_i = 0.5$, respectively. The values of parameters selected are the same as used in Section (3.3). The surface shown in these figure represents the minimum resource requirement surface for a feasible schedule, i.e., each point on or above this surface, represents a feasible schedulable that satisfy the schedulability condition (Equation (9)) and the QOP. Therefore, for a given QOP the crossponding point on the surface denotes the minimum size of the buffer needed to attain this QOP. By comparing height of the surfaces in Fig.(6(a)) and Fig.(6(b)), we can deduce that an increase in channel utilization results in more buffering needs for the same QOP. This is intuitively obvious since for increase in channel utilization, an additional amount ($(1-p_i)$) of buffer is need for rate compensation.

5.1.1 Further Buffer Minimization at the Cost of Channel Utilization

By adjusting the transmission rate to a suitable value, one can further reduce the required buffer space at the destination. The following theorem describes the optimal transmission rate which minimizes buffering requirement at the receiver.

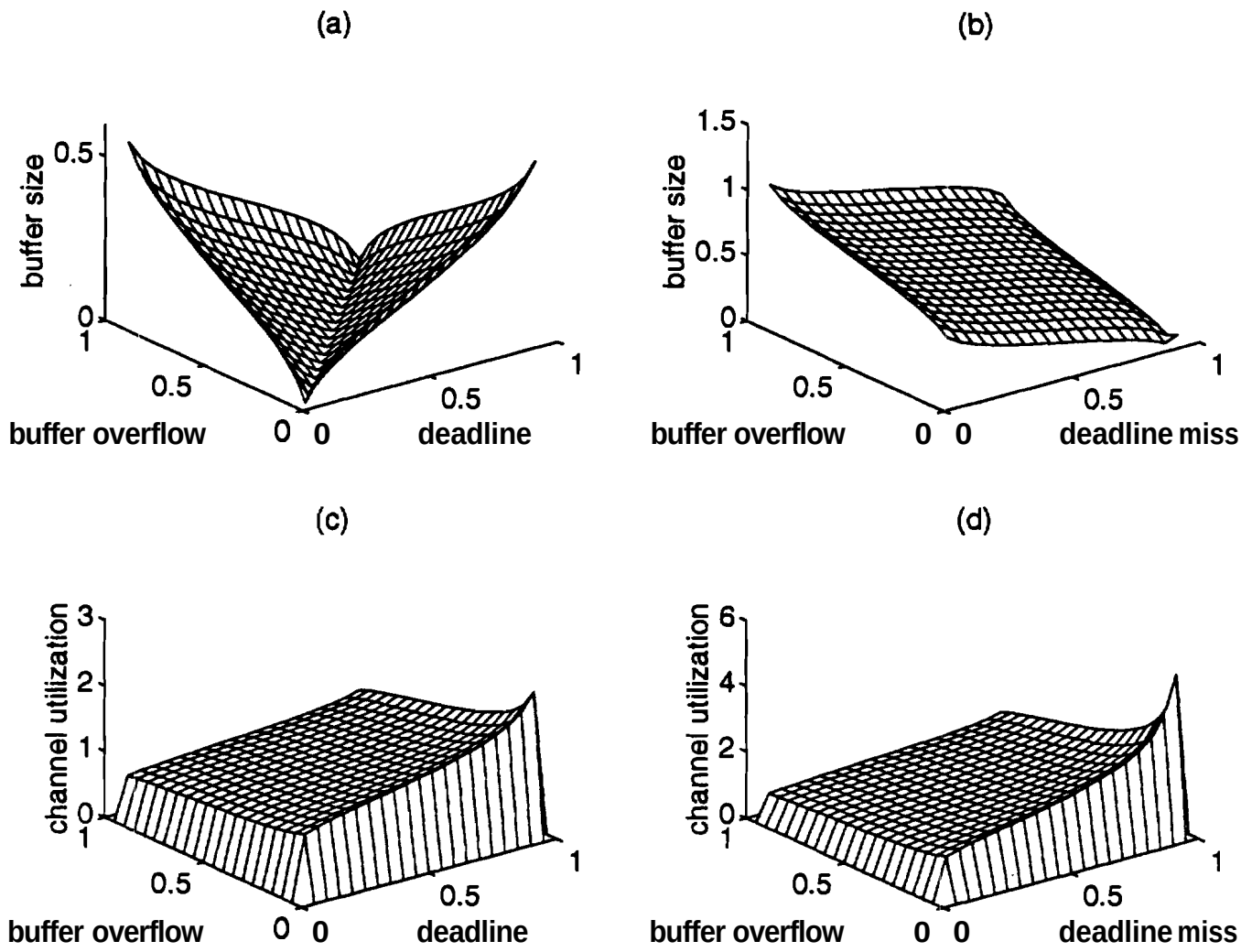


Figure 6: Resource utilization (a) k_i^{min} when $rho_i = 1$, (b) k_i^{min} when $rho; = 0.5$, (c) η_i^{max} when $k_i = 0$, (d) η_i^{max} when $k; = 0.25$.

Theorem **5.2** *For a given set of QOP parameters the receiver buffer requirement is minimized when the object transmission rate matches the object consumption rate, i.e,*

$$\lambda_i^{\text{min buffer}} = \mu_i$$

The size of the minimum buffer is given by

$$\hat{k}_i^{min} = \frac{F_{\mathbf{D}i}^{-1}(1 - \mathcal{P}_i^d) + F_{\mathbf{D}i}^{-1}(\mathcal{P}_i^b)}{\tau_i}$$

Proof: In theorem (5.1), we found that for a given transmission speed, minimum fraction of object that need to be buffered at the receiver in order to satisfy desired quality of presentation is given by

$$k_i^{min}(\lambda_i) \geq (1 - \frac{\mu_i}{\lambda_i}) + \frac{F_{\mathbf{D}i}^{-1}(1-\mathcal{P}_i^d) - F_{\mathbf{D}i}^{-1}(\mathcal{P}_i^b)}{\tau_i}$$

Recall that both probability of buffer overflow (see Equation 5, 6) and probability of deadline misses (see Equation 6) are independent of the transmission rate. Hence, function $k_i^{min}(\lambda_i)$ has attains its minimum value when $\lambda_i = \mu_i$. ∎

The size of minimum buffer for the case when there is no asynchrony in object transmission rate and its consumtion rate is given by the following corollary.

Corollary **5.2.1** *For the case when the transmission rate is equal to consumption rate, some buffering at the receiver is still needed in order to satisfy the QOP parameters. The size of the minimum buffer is given by*

$$\hat{k}_i^{min} = \frac{F_{\mathbf{D}i}^{-1}(1 - \mathcal{P}_i^d) + F_{\mathbf{D}i}^{-1}(\mathcal{P}_i^b)}{\tau_i}$$

5.2 Maximum Channel Utilization (Fixed Buffer)

If we assume that enough capacity is available on the channel to support transmission of the object, then for a given buffer size, we can find the maximum rate of transmission for

best channel utilization. For a given buffer size, channel utilization is limited mainly by $\mathcal{P}_i^b$. Therefore, there exists an upper bound on channel utilization given by the following theorem.

Theorem 5.3 *If* $\exists\ \lambda_i^{sch} \leq C_i$ *such that*

$$1 \leq \eta_i^{max} \leq \frac{1}{1-k_i+\frac{F_{\mathbf{D}_i}{}^{-1}(1-\mathcal{P}_i^d)-F_{\mathbf{D}_i}{}^{-1}(\mathcal{P}_i^b)}{\tau_i}} \tag{10}$$

where $\eta_i^{max} = \frac{1}{\mu_i}$, *then object* O_i *can be transmitted wing the following optimal value of* T_i^c

$$T_i^{c,opt} = F_{\mathbf{D}_i}{}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \frac{1}{\eta_i^{max}})\tau_i \tag{11}$$

This control time guarantees the desired QOP and mazimizes the channel utilization.

Proof: From theorem (4.3), it is obvious that for a given buffer size $K_i = k_i s_i \leq s_i$ and the QOP parameters, object O_i is schedulable iff we can find $\mu_i \leq \lambda_i^{sch} \leq C_i$ such that schedulability condition (7) is satisfied. Therefore we need to satisfy the following condition

$$\begin{aligned} & F_{\mathbf{D}_i}{}^{-1}(\mathcal{P}_i^b) + k_i\tau_i - (1 - \tfrac{\mu_i}{\lambda_i^{sch}})\tau_i \geq F_{\mathbf{D}_i}{}^{-1}(1-\mathcal{P}_i^d) \\ \Rightarrow \quad & \lambda_i^{sch} \leq \frac{\mu_i}{1-k_i+\frac{F_{\mathbf{D}_i}{}^{-1}(1-\mathcal{P}_i^d)-F_{\mathbf{D}_i}{}^{-1}(\mathcal{P}_i^b)}{\tau_i}} \end{aligned}$$

The above condition coupled with $\mu_i \leq \lambda_i^{sch} \leq C_i$ completes the proof. ∎

For the same set of parameters used in Section (3.3), Fig.6(c) and Fig.6(d) show channel utilization verses QOP parameters for the case $k_i = 0$ and $k_i = 0.25$, respectively. The surfaces shown in these figures represent the maximum channel utilization for a feasible schedule, i.e. for each point on or below this surface, we can find a schedulable control time from Equation (11) such that the desired QOP parameters are guaranteed. By comparing height of the surfaces in Fig. 6(c) and Fig. 6(d), we can deduce that an increase in available buffer size at the destination allows us to operate the available channel to its maximum utilization.

6 Conclusion

In this report, we have presented a framework for evaluating performance of scheduling pre-orchestrated multimedia information over broadband integrated networks. We have proposed a set of Quality Of Presentation (QOP) parameters which quantify the quality of multimedia presentation process from user's point of view. We have presented, trade-offs between proposed QOP parameters and the system resources which include channel utilization and buffering at the destination. Based on these trade-offs, one can design an optimal transmission schedule for multimedia information both at the object and ROCPN level.

References

[1] Anick, D., Mitra, D., and Sondhi, M.M., "Stochastic Theory of a Data-Handling System with Multiple Sources," Bell System Technical Journal, vol. 61, no. 8, October 1982, pp. 1871-1894.

[2] CCITT Draft Recommendation I.121, "Broadband Aspects of ISDN," Study Group COM XVIII-R 34-E, Geneva, Switzerland, June 1990, pp. 20-22.

[3] Chowdhury, S., "Distribution of the Total Delay of Packets in Virtual Circuits," in Proc. *of* IEEE INFOCOM, vol. 2, pp. 911-918, 1991.

[4] Daigal, J.N., and Langford, J. D., "Models for the Analysis of Packet Voice Communications Systems," IEEE J. Select. Areas Communic., vol. SAC-4, no. 6, September 1986, pp. 847-855.

[5] Ferrari, D., "Client Requirements for Real-Time Communication Services," IEEE Communication Magazine, vol. 1, no. 3, 1991, pp. 203-226.

[6] Kroner, H., Elberspacher, M., Theimer, T.H., Kuhn, P. J., Briem, U., "Approximate Analysis for the End-to-End delay in ATM Networks," in Proc. of IEEE INFOCOM, vol. 3, 1992, pp. 978-986.

[7] Kleinrock, L., "The Latency/Bandwidth Tradeoff in Gigabit Networks," IEEE Communication Magazine, vol. 30, no. 4, April 1992, pp. 36-40.

[8] Kleinrock, L., "Queueing Systems, Vol. 2 : Computer Applications," Newyork: Wiley, 1976.

[9] Lazar, A. A., and Pacifici, G., "Control of Resources in Broadband Networks with Quality of Service Guarantees," IEEE Communication Magazine, October 1991, pp. 66-73.

[10] Little, T.D.C., and Ghafoor, A., "Multimedia Synchronization Protocols for Broadband Integrated Services," IEEE Journal on Selected Areas in *Communications*, vol. 9, no. 9, December 1991, pp. 1368-1382.

[11] Little, T.D.C., and Ghafoor, A., "Synchronization and Storage Models for Multimedia Objects," IEEE Journal on Selected Areas in Communications, vol. 8, no. 3, April 1990, pp. 413-427.

[12] Montgomery, W. A., "Techniques for Packet Voice Synchronization", IEEE Journal on Selected Areas in Communications, vol. SAC-1, no. 6, December 1983, pp. 1022-1028.

[13] Ramanathan, S., and Rangan, P.V., "Adaptive Feedback Techniques for Synchronized Multimedia Retrieval over Integrated Networks," *IEEE/ACM* Transactions on Networking, vol. 1, no. 2, April 1993, pp. 246-259.

[14] Steinmetz, R., "Synchronization Properties in Multimedia Systems":,IEEE Journal on Selected Areas in Communications, vol. 8, No. 3, April 1990, pp. 401-412.

[15] Woo, M., Qazi, N.U., Ghafoor, A., "A Synchronization Framework for Communication of Pre-orchestrated Multimedia Information over Broadband Networks," To appear in the *IEEE* Network Magazine, Jan. 1994.

[16] Ohba, Y., Murata, M., Miyahara, H., "Analysis of the Interdeparture Process for Bursty Traffic in ATM Networks," *IEEE* Journal on Selected *Areas* in Communications, vol. 9, No. 3, April 1991, pp. 468-476.