Resource Prediction for Admission Control of Interactive Multimedia Sessions

Silvia Hollfelder and Karl Aberer
GMD - IPSI, Dolivostr. 15, 64293 Darmstadt, Germany
E-mail: hollfelder,aberer@darmstadt.gmd.de

Abstract

Interactive multimedia sessions have uncertain, varying consumption rates due to users' interactive behavior. In this paper, we propose two approaches for prediction of consumption rates required for admission control. They differ in the assumptions of the available knowledge: The observation-based approach predicts the future consumption from the past system behavior. The stochastical approach requires a user behavior model from which the prediction is deduced. We discuss the pros and cons of the two approaches with respect to available information sources, accuracy of prediction, and suitable scenarios.

Keywords: *interactive multimedia applications, user modeling, predictive service*

1 Motivation

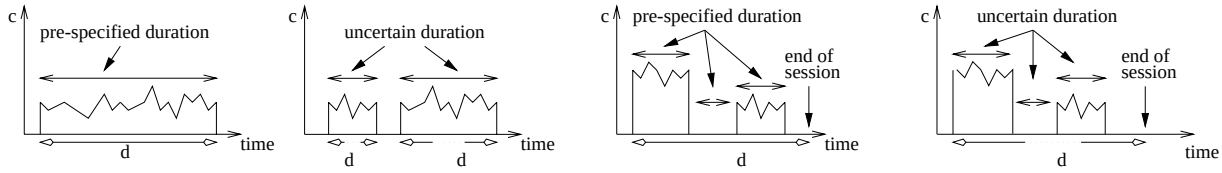
In highly interactive applications a user controls the course of a running session in terms of which media(-combinations) will be presented at what time. Such applications play a key role in domains like home entertainment (e.g., news-on-demand, interactive VoD, action games), home shopping (e.g., product catalogs), education and training (e.g., interactive computer-based training - CBT), tourist information (e.g., virtual travel guides), and multimedia production (e.g., video editing).

Multimedia systems have to provide mechanisms that are able to deal with the special characteristics of such applications to enforce QoS parameters like continuous playout of media data and low start-up latency between two subsequent media presentations. To achieve the required QoS, the clients compete for limited resources on the server. Resource management for multimedia applications is a classical problem of research in the area of distributed multimedia systems. An admission control mechanism usually checks at the server if enough resources are available for the adequate delivery of data to a new request, given the current set of admitted clients.

Figure 1 classifies related work on admission control by the reservation duration d of an admission and the consumption rate of a client. In multimedia systems, the admission can be given on single media streams, like an audio or video stream, both for constant bit rate (CBR) and variable bit rate (VBR) media. Most of those single-stream based approaches ([ZK97], [MNH97], [ND96], [ORSN96]) target at applications with less interactive behavior, such as video-on-demand. In non-interactive applications, resources are reserved for the presentation duration of a medium (see Figure 1 (a)), and in the interactive case (see Figure 1 (b)) the admitted media streams are usually served until a user interaction, like a pause, occurs. In order to continue the delivery of data after an interaction, the system has to re-admit the stream for the service. To reduce start-up delay after VCR-interactions, such as pause or fast forward, priority streams are introduced in [Red97]. But this granularity of admission is not suitable for multimedia applications that combine media, like a preorchestrated multimedia document that starts with an audio, continues with a text, and ends with synchronized audio and video. Re-admissions due to varying resource demands may lead to unacceptable start-up delays between the subsequent media presentations in case of high workload. Furthermore, specified synchronization requirements of composite presentations have to be considered.

Thus, to reduce start-up latency between subsequent media presentations, admission for entire multimedia sessions can be given. A session is a sequence of typically short media presentations within a common application context. Sessions can be distinguished into non-interactive and interactive ones. In contrast to single streams, the consumption rate of sessions varies much higher due to the combination of various media within one session (e.g., combination of MPEG-1, MPEG-2, M-JPEG video, MP3 audio streams). Figure 1 (c) displays the data rate variations for media with variable bit rates for the non-interactive case.

The main target here is to find an admission criterion for those sessions that use the available resources economically. Economical resource utilization means that multiple sessions share the available resources in such



(a) Single Stream (non-interactive) (b) Single Stream (interactive) (c) Sessions (non-interactive) (d) Sessions (interactive)

Figure 1: Classification of Related Work on Admission Control.

a way that resource requirements can be interlinked so that a maximum number of sessions can be served. For the non-interactive case, the exact temporal course and, thus, the consumption rate of a session is known in advance. For example, in [ZT98] and [BO98] admission control mechanisms are proposed for this case.

From our point of view, admission for highly *interactive* applications has to be granted on session granularity, too, aiming to reduce start-up latency after frequently occuring user interactions. Highly interactive applications offer the user VCR-functionality, enable a user to control playback duration and to control the course of a session by means of selecting alternative media (e.g., video browsing) or whole presentation paths (e.g., preorchestrated training units in computer-based training). High start-up delays are intolerable for these applications for the following reasons: typically, the presentation time of requested media clips is short, compared to media streams presented in applications such as video-on-demand, and media switches occur frequently. One problem here is that both, the presentation duration of the media and their choice are *uncertain*. The arrows in Figure 1 (d) display the ways a user may influence the presentation progress and thus the consumption rate of an interactive session.

To the best of our knowledge, there exists no suitable mechanism that enables session-based admission control for *interactive* applications that are characterized by highly varying resource requirements caused by variable bit rate media, media combinations within a session *and* user interactions. Since the users' interactive behavior determines the resources required, the consumption rate of a multimedia session is, compared to single stream-based and session-based, non-interactive approaches, more difficult to specify.

The objective of our work is the development of session-based admission control mechanisms, which are based on predicted consumption rates, aiming at support for highly interactive multimedia applications. The admission control criteria need to be flexible and adaptable and fulfill the target metrics high server utilization and good Quality of Service in terms of served requests within their deadlines. By flexibility, we mean that an admission criterion needs to be able to handle various types of interactive multimedia sessions, such as more uniform ones with similar consumption rates as well as more bursty ones, and those with short as well as long session durations. Thus, a quite simple strategy, like ”admit a specific number of sessions” is not suitable. Furthermore, the criterion should be able to adapt to worse admission decisions and, thus, enable the system to recover from underloaded or overloaded situations.

In earlier work, we proposed a generic, observation-based approach [HA98], and a stochastical approach with Markov-chains prediction, that is based on an application specific user model [FHA99]. For both approaches, we specify corresponding admission control criteria to give predictive guarantees and to achieve efficient resource utilization. In this position paper, we compare the both approaches with respect to the available information sources, the preciseness of predition, the computational overhead and suitable scenarios.

2 Prediction of Consumption Rates of Interactive Multimedia Sessions

Generally, a pending client c^p will be admitted, when sufficient system resources are available. Sufficient system resources means in this context that the predicted system utilization is not higher than the available one. Thus, as well for the admitted clients c^a as for a pending client session the future consumption rates have to be estimated in advance.

The resource prediction might be static or dynamic. A static prediction holds for the whole future. The dynamic predictions will be frequently actualized and adapted to the present available knowledge such as the current presentation states of the clients, or the current request behavior.

2.1 Granularity of Consumption Rate Prediction

The prediction of the future consumption rates can be given in a fine-grained or coarse-grained way. Fine-grained analysis means that resource requirements are more precisely specified, coarse-grained prediction results in smoothed future values.

We first introduce a time window w that determines how far the prediction holds for the future. This window w, starting from present time t_0, is divided into single segments s, with n as number of segments of w (see Figure 2). For simplification, all segments have the same length. A segment is the smallest time unit for which the consumption rates will be predicted. The higher the number of segments n for a given window is, the more fine-grained is the prediction. In the other extreme, in the most coarse-grained analysis, the window w consists of only one segment ($n = 1$). In the following, we denote by $cr(c_t)$ the consumption rate of client c in segment t.

The optimal length of w and the number of n results from the available information since a fine-grained prediction is not always possible due to lack of available knowledge of the running and pending sessions.

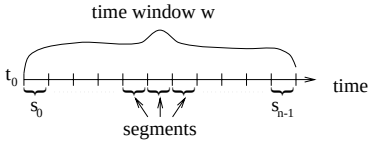


Figure 2: Time Window and Segments for the Resource Prediction.

In the following we identify information sources from which a more or less accurate prediction on the consumption rates can be given and discuss two approaches that differ in the assumptions of available information.

2.2 History Observation

When a multimedia server has neither knowledge about the kind of application it has to serve nor about the user behavior, the only way is to evaluate the clients past request behavior. The main idea behind this observation-based approach is to predict the consumption rate from the history, assuming that the past behavior is an indicator for the future. Therefore, the recording of historical information is required to calculate access statistics. For history observation, the critical parameters are the length of the observed history, the granularity of observation, and evaluated bookkeeping parameters.

2.2.1 Granularity of Monitoring

Accordingly to the time window w, we specify a time window h for the observed history which we call *history window*. We divide the window into segments, called *history units* u with m as the number of history units of that window (see Figure 3 with t_0 as current time). The history units are the time intervals that are individually inspected with respect to the request behavior. Thus, the larger the number m for a given window h, the more detailed the request behavior will be analyzed, but the more expensive is the computation, and vice versa.

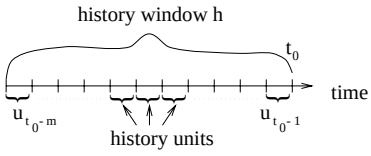


Figure 3: History Window and History Units that are Observed.

2.2.2 Bookkeeping and Admission Control

Monitoring can be employed for single clients or for the entire set of admitted clients. One question here is, whether the individual clients' behavior of the past is representative for its individual future consumption rate.

Having a typical interactive multimedia session in mind that consists of any combinations of time-dependent and discrete media, the individual behavior is not necessarily representative. We are not able to make any predictions for the media combinations, the average presentation duration of a media, or the probabilities for VCR-interactions, without having application-specific data or user profiles. But, on the other hand, the past behavior over all admitted clients might be an indicator for their overall future behavior, assuming that for a large number of clients useful statistics on future access patterns can be generated. Thus, in our approach, average access patterns for a large number of clients are used for the prediction.

We define the bookkeeping parameters 'average request rate' $\overline{r}$ and 'average served rate' $\overline{s}$ for a history h in [HA98]. These parameters are generated as follows: first, for each history unit u the average request rate and the average served requests for all clients are monitored. At admission time, the average values $\overline{r_h}$ or $\overline{s_h}$ for the whole history h are calculated. Then we employ the heuristic rules that the overall request behavior for the whole group of admitted clients will not change and deduce the consumption rate of a client by $cr(c) = \overline{r_h}$ or $cr(c) = \overline{s_h}$, both for admitted and pending clients.

Furthermore, we introduce a safety margin $\tau \in [0, 1]$. Let k be the number of currently admitted clients C^a, and s_{max} the maximal amount of resources that are available. A pending client c^p is admitted at time t_0, if

$$(k + 1) * cr(c) < \tau * s_{max} \tag{1}$$

Due to the fact that we use average access statistics, one drawback is that we are not able to make fine-grained predictions. Thus, the admission criterion refers to a coarse-grained window w with $n = 1$. Another drawback is that the bookkeeping of a long, fine-grained history is expensive in terms of monitoring overhead.

To enable a more cautios prediction, access statistics like minimum and maximum values, deviations in between subsequent time intervals, or distribution of access patterns need to be considered. In this case, effects of outliers can be recorded, too. The question is how these effects can be represented in the admission control criterion, aiming to achieve high server utilization, and whether the more expensive monitoring is worth it.

In our approach, the prediction is dynamic, since at each admission time updated access patterns are used. Thus, we are able to react flexible on the current system behavior. The approach and first evaluations on the size of window, bookkeeping parameters, and admission criterion are described in detail in [HA98]. Further simulation studies will show in chapter 3 that it is applicable when the overall client behavior is uniform and rare outliers of clients can be smoothed by the (typically) large number of uniform clients.

Another observation-based approach is proposed in [ACS98] to make short-term predictions about network delay to improve receiver playback buffer management. An observation-based approach for disk scheduling is introduced by [VGGG94]. The benefit of the observation-based approach is that it does not require high-level understanding of the semantics of the application and, therefore, it is totally application independent. On the other side, the implicit heuristic assumption in the observation-based approach that the overall user behavior does not change may be inappropriate if opposite knowledge exists. We will now discuss situations in which information on the future access behavior is available.

2.3 Modeling of User Behavior by Interpretation of Application Semantics

When a server has the information about the type of application it serves, it can employ application specific knowledge to make a more accurate prediction of resource needs. This is realistic for special purpose servers, such as news-on-demand system, and for general-purpose systems, assuming that multimedia sessions have to identify their application type to the server.

The idea here is to extract useful information from the specific characteristica of the applications. Employing application semantics might limit the subset of media a user requests within a multimedia session. For example, users request access to a multimedia database for video browsing by sending a query with respect to a specific content. Thus, the result list of the retrieval request represents a subset that may bear certain characteristics which allow much more precise estimations of future resource usage. When the data rates of the media are available as meta data, the required consumption rates of such applications can be restricted.

2.3.1 User Modeling

Another important aspect is that assumptions on specific user behavior can be made by employing application semantics. In the context of this paper, a user behavior model is a model for representing media presentations and

possible user interactions within a multimedia session that fulfills the following requirements:

- Representation of uncertainness. Since the user behavior is not deterministic, methods for specifying uncertainness with respect to user interactions are required. Typically, probabilistic models are employed.
- Representation of temporal aspect. To specify required resources, the presentation time of continuous media is important. Furthermore, fine-grained temporal prediction on the consumption rate should be enabled.
- Response on current behavior. Due to the high dynamics of interactive sessions, the model must be able to consider the presentation process within a session (i.e., the actual presented media) and/or the current system behavior with respect to admitted sessions (e.g., underloaded or overloaded periods).

We represent user behavior by means of Continuous Time Markov Chains (CTMC). A CTMC is a stochastic process in which changes of the state may occur at each point of time. In contrast to Discrete Time Markov Chains, times between successive transitions, i.e., the holding times of the states, are exponentially distributed. CTMC are proposed, for example, to model user behavior and to predict access on multimedia documents stored in tertiary and secondary storage system in [KW97].

We model user behavior in interactive multimedia sessions as follows: the states of a CTMC represent the media presentations and the transition probabilities stand for the user interactions. CTMC fulfill our requirements on user models as follows: The uncertain user behavior is adequately represented by means of exponential distribution of holding times and transition probabilites, the holding times of states truly consider the temporal dimension of media presentations. CTMCs provide mathematical methods for prediction: For closed CTMC, i.e., multimedia sessions where each presentation state can be reached from any other states, the long-run probabilities for consumption rate of a session can be predicted by means of equilibrium analysis. Fine-grained and coarse-grained prediction is possible, both for open and closed chains, by specifying time windows for the more complex transient analysis [Tij94]. Furthermore, the response on current behavior of sessions is possible since for the transient analysis the current state of a session can be employed.

Other proposed models for representing dynamic behavior are, for example, labeled (stochastic) trees, where probabilities of user selections are modeled to generate materialized views for interactive multimedia presentations [CLS98], and Coupled Hidden Markov Chains that model interactive processes in perceptual computing [BOP97]. Both models cannot be employed for our purpose since the temporal aspect of media presentation is not considered.

2.3.2 Resource Prediction and Admission Control

Since the user behavior can be formally specified within that mathematical model, the probabilities for each media presentation at each time during a running multimedia session and from that a fine-grained prediction of the consumption rate $cr(c_t)$ can be employed.

We calculate for each segment *t* of the time window w an upper bound $ub(t)$ for the overload probability. An overload occurs when the time T^t_{serve} a server needs in a segment for serving the data requested by all admitted clients is larger than the length of the segment. For $ub(t)$ the following holds:

$$ub(t) \geq P(T^t_{serve} > segment) \tag{2}$$

This upper bound is calculated in two steps: First, a matrix is determined to predict the number of clients in each of the presentation states by using the time dependent transition probabilities $p_{i,j}(t)$, representing that a client moves from state i to state j within time t. The $p_{i,j}(t)$ can be calculated by the uniformization method [Tij94]. Secondly, to calculate $ub(t)$, we use the *Chernov Inequality* [Kle75] which has the form

$$p(Y \geq x) \leq \inf_{\theta \geq 0} e^{(-\theta x)} G_Y(\theta) \tag{3}$$

In this inequality, $G_Y(\theta)$ is the so-called *Moment Generating Function* of Y. Between the Moment Generating Function G_Y and the so-called *Laplace Transformation* $F^*_Y(\theta)$ the relationship $G_Y(\theta) = F^*_Y(-\theta)$ holds [Kle75].

Thus, to compute $ub(t)$, we first have to solve the problem of how to determine $G_{T^t_{serve}}(\theta)$, i.e., the Moment Generating Function of T^t_{serve}. To achieve this, we proceed as follows: First, we express T^t_{serve} in terms of known

random variables. Then, we use this expression to calculate the Laplace Transformation of T^t_{serve} and derive from this Laplace Transformation the desired Moment Generating Function using the equation above. For more details on the stochastical model, we refer to [FHA99]. A pending client c^p is admitted at time t_0, if

$$ub(t) \leq \tau \; \forall \; t \; in \; w \tag{4}$$

Our admission criterion is very restrictive in the fine-grained analysis since a single bottleneck within a segment leads to the rejection of the pending client, independent on the utilization of the other segments of w. Due to the fact that the current presentation states of the admitted clients are considered, the prediction is dynamic, too. Note that for the calculation of the matrix both admitted clients and pending clients c^p are taken into account.

One drawback of the stochastic approach is that it usually results in a much higher computational complexity, which is dependent on the chosen user behavior model (i.e., the number of states and transition probabilities), the number of segments to be considered at admission time and thus the preciseness of the prediction, and the mathematical analysis (transient analysis versus equilibrium analysis). Another problem is that the heuristic assumptions on the user behavior may not adequatly reflect the real world. Hence, additional evaluations of real scenarios and adaptations on observed behavior are needed.

2.3.3 Heuristic Parameter Setting

One question with respect to the stochastic aproach is how to develop a proper user model. The structure of a user model can be extracted, for example, from application specific knowledge, when users present preorchestrated multimedia documents with a restricted subset of involved media and prespecified temporal presentation order. We analyse user behavior for presenting preorchestrated multimedia documents in [FHA99]. In [AH99] we give an example how application semantics can be extracted in browsing scenarios, assuming a rationale user whose behavior is triggered by relevance values related to the ranked results of a previous query. From the result set, consisting of video shots with corresponding relevance values, the holding times of the presentation states and the transition probabilites are heuristically deduced [HET99].

2.4 The Role of User Profiles

User profiles enable to improve the service to individual users based on their personal preferences. The data stored in a user profile, such as average session duration, preferred media quality, available equipment (e.g., network connection), and the content a user is interested in can also be employed for our purposes to predict consumption rates. For example, users with low network bandwidth may always drop high resolution video presentations.

User profiles may be used to precise the parameters of a user model (e.g., the mean presentation time of a media) and to reduce the complexity of the model (e.g., drop all presentations that relate to sport video clips, knowing that a user is not interested in that topic). One drawback is that user profiles typically contain information from which only fuzzy knowledge can be deduced. But user profiles can be combined with history-based approaches to adapt the profile by studying the real user behavior. Another drawback is that user profiles are primary not developed for resource prediction purpose, and thus the required data for specification of consumption rates are not necessarily contained in a profile by default. This results in additional administration overhead to keep such data.

3 Simulation Results

In the following, we evaluate our approaches with respect to the requirements of flexibility and adaptability. We study the QoS metrics server utilization and number of requests served within their deadlines.

One question that refers to the observation-based approach is, whether a coarse-grained prediction can be given for a small number of clients to be served. Thus, we study the system behavior by reducing the available system resources both for uniform clients and clients with drastically varying request behavior, so-called non-uniform clients.

We use two metrics for our evaluation: utilization is computed from the number of requests served by the server within a history unit, and server load measures all open requests (see [HA98]). Table 1 shows the average utilization $\overline{u}$, the variance of utilization (var. of $\overline{u}$), the average server load $\overline{l}$, the variance in server load (var. of $\overline{l}$), and QoS in percentage of requests served in time.

The results show that the available resources are a critical parameter for the coarse-grained prediction. The lower the resources, the worse the QoS. Table 1 shows that for non-uniform clients the QoS and the utilization is lower (left side) than for uniform clients with the same available resources (right side).

	non-uniform clients					*uniform clients*				
s_{max}	$\overline{u}$	var. of $\overline{u}$	$\overline{l}$	var. of $\overline{l}$	QoS	$\overline{u}$	var. of $\overline{u}$	$\overline{l}$	var. of $\overline{l}$	QoS
5000	0.53	0.35	14.96	43.28	0.60	0.61	0.26	1.29	1.42	0.81
6000	0.65	0.25	0.95	0.79	0.85	0.69	0.24	0.93	0.69	0.87
7000	0.67	0.23	0.98	0.92	0.90	0.74	0.23	0.93	0.55	0.90
8000	0.66	0.25	0.96	0.82	0.91	0.75	0.20	0.90	0.50	0.92
9000	0.73	0.20	0.91	0.46	0.92	0.72	0.19	0.97	0.77	0.96
10000	0.71	0.22	0.93	0.61	0.94	0.75	0.19	0.94	0.51	0.94
25000	0.75	0.21	0.96	0.64	0.95	0.82	0.17	0.93	0.38	0.97

Table 1: System Behavior with Varying Available Resources

Further, we observed in this experiment a much better behavior by using the 'average request rate' $\overline{r_h}$ than the metric 'average served rate' $\overline{s_h}$ for the prediction (see Chapter 2.2.2). Even for uniform clients the system behavior became instable for low s_{max}-values when the parameter $\overline{s_h}$ was employed for the prediction. For example, for s_{max} = 5000 we obtained oscillating behavior with QoS-values lower than 0.10 for 1000 simulation periods. The values in Table 1 are computed with the parameter $\overline{r_h}$.

We employed the stochastic approach to evaluate situations in which the server has to deal with excessively high variations in data rates. For sure, the utilization decreases, since probable bottlenecks for a fine-grained window are considered. In our experiments, the system behavior was stable and the QoS values were high.

A critical point of the stochastic approach is the correctness of the user model used for prediction. Thus, we studied deviations of the assumed user behavior to the real one. Simulations show, that even with unexpected behavior good results can be achieved. Furthermore, we evaluated changes in length of segments and length of the time window for the stochastic approach. We observed that with increasing size of window w similar QoS results were achieved. The reason for this is that the prediction converges into an equilibrium state for a large time window. For more details on the experimental testbed and the results we refer to [FHA99].

To sum up, the observation-based approach with a coarse-grained prediction is suitable for a large number of sessions, even with non-uniform access patterns, since users with unexpected behavior can be neglected for statistical reasons. If, on the other side, only few users with extremely differing access behavior will be served, a more precise prediction should be performed to avoid overload situations.

4 Conclusions

In this position paper, we compared two approaches for resource prediction of interactive multimedia sessions with respect to various degrees of knowledge that can be applied for the prediction.

Within the observation-based approach, the decision, whether a new client will be admitted, is based on the past system utilization. The main advantages of this approach are that it is fully application-independent and it truly reflects the actual system usage. So, the mechanism is able to recover from phases of overloaded and underloaded periods without a need for validation of a user model. The main drawback is that it enables only coarse-grained prediction and, thus, it is sensible on unexpected behavior.

The stochastic prediction from user models enables fine-grained analysis of future time units and thus results in a more accurate prediction. The drawbacks are, that many assumptions and heuristics might be involved in the modeling of user behavior. Furthermore, a fine-grained analysis may lead to extremely complex models with expensive computational overhead.

On the other hand, a precise prediction is not always needed. The design decision stands in direct relation to the ratio of available system resources, the number of sessions that can be served, and the deviations of access patterns. The rules of thumb are: (1) the larger the number of sessions the less important is the individual behavior. (2) the more extremely the user behavior varies in scenarios, the more important is a detailed modeling approach.

We are currently working on the observation of user traces for video browsing to evaluate the heuristics for the browsing scenario and on further simulations to compare both approaches under various scenarios and workload.

References

[ACS98] Prathima Agrawal, Jyh-Cheng Chen, and Cormac J. Sreenan. Use of statistical methods to reduce delays for media playback buffering. In *IEEE Multimedia Systems*, pages 259–263, June/July 1998.

[AH99] Karl Aberer and Silvia Hollfelder. Resource prediction and admission control for interactive video browsing scenarios using application semantics. In *Proc. of Int. Conf. on Data Semantics - 8 (DS-8), Semantic Issues in Multimedia Systems, IFIP TC-2 Working Conference*, pages 27–46, January 1999.

[BO98] Nevzat Hurkan Balkir and Gultekin Ozsoyoglu. Delivering presentations from multimedia servers. *VLDB Journal. Special Issue on Multimedia Databases*, pages 297–307, December 1998.

[BOP97] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden Markov models for complex action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, June 1997.

[CLS98] K. Selçuk Candan, Eric Lemar, and V.S. Subrahmanian. Managing and rendering multimedia views. In Sushil Jajodia, M. Tamer Özsu, and Asuman Dogac, editors, *Proc. of 4th Int. Workshop Multimedia Information Systems (MIS)*, pages 45–56. Springer Lecture Notes in Computer Science, September 1998.

[FHA99] Matthias Friedrich, Silvia Hollfelder, and Karl Aberer. Stochastic resource prediction and admission for interactive sessions on multimedia servers. GMD Technical Report 50, GMD, Sankt Augustin, Germany, March 1999. http://www.darmstadt.gmd.de/oasys/reports/index.html, submitted for publication.

[HA98] Silvia Hollfelder and Karl Aberer. An admission control framework for applications with variable consumption rates in client-pull architectures. In Sushil Jajodia, M. Tamer Özsu, and Asuman Dogac, editors, *Proc. of 4th Int. Workshop Multimedia Information Systems (MIS)*, pages 82–97. Springer Lecture Notes in Computer Science, September 1998.

[HET99] Silvia Hollfelder, Andre Everts, and Ulrich Thiel. Concept-based browsing in video libraries. In *IEEE Forum on Research and Technology Advances in Digital Libraries (IEEE ADL 99)*, pages 105–115. IEEE Computer Society, Los Alamitos, May 1999.

[Kle75] Leonard Kleinrock. *Queueing Systems, Volume I: Theory*. Wiley, 1975.

[KW97] Achim Kraiss and Gerhard Weikum. Vertical data migration in large near-line document archives based on markov-chain predictions. In *VLDB*, pages 246–255, 1997.

[MNH97] Dwight Makaroff, Gerald Neufeld, and Norman Hutchinson. An evaluation of VBR disk admission algorithms for continuous media file servers. In *Proc. of ACM Multimedia*, pages 143–153, 1997.

[ND96] Raymond T. Ng and Rita Dilek. Statistical modelling and buffer allocation for MPEG streams. In S. M. Chung, editor, *Multimedia Information Storage and Management*, pages 147–162. Kluwer Academic Publishers, 1996.

[ORSN96] Banu Özden, Rajeev Rastogi, Avi Silberschatz, and P. S. Narayanan. The Fellini multimedia storage server. In S. M. Chung, editor, *Multimedia Information Storage and Management*. Kluwer Academic Publishers, 1996.

[Red97] Narasimha Reddy. Improving latency in interactive video server. In *Proc. of SPIE Multimedia Computing and Networking Conference*, pages 108–112, Feb 1997.

[Tij94] Henk C. Tijms. *Stochastic Models. An Algorithmic Approach*. Wiley series in probability and mathematical statistics. Wiley, 1994.

[VGGG94] Harrick M. Vin, Alok Goyal, Anshuman Goyal, and Pawan Goyal. An observation-based admission control algorithm for multimedia servers. In *Proc. of the First IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, pages 234–243, May 1994.

[ZK97] Hui Zhang and Edward W. Knightly. RED-VBR: a renegotiation-based approach to support delay-sensitive VBR video. *Multimedia Systems*, pages 167–176, May 1997.

[ZT98] Wei Zhao and Satish K. Tripathi. A resource reservation scheme for synchronized distributed multimedia sessions. *Multimedia Tools and Applications*, 7(1/2):133–146, July 1998.