Quality of Service Guarantees for Multimedia Digital Libraries and Beyond

Gerhard Weikum
University of the Saarland, Saarbrücken, Germany
E-mail: weikum@cs.uni-sb.de, WWW: http://www-dbs.cs.uni-sb.de/

Extended Abstract

Servers for multimedia digital libraries have to manage huge amounts of data and pose challenging performance requirements. Most notably, for smooth playback of video and audio on client machines the server has to guarantee *continuous data streams* with just-in-time delivery of the underlying data fragments. This problem is referred to in the multimedia information and networking community as the need for guaranteed *quality of service*, in short *QoS*, where the details of such guarantees depend on the data and playback specifications (e.g., video formats) [1, 4, 8, 9].

Servers aim to maximize their throughput in terms of the concurrently sustained number of continuous data streams, but at the same time have to ensure that each active stream meets the QoS demands of the clients. Therefore servers (and potentially also network components) need to employ an *admission control* that limits the number of simultaneously active streams. For given server resources such as number of disks, memory size, etc. a new stream is admitted only if it can be safely determined that *all* data fragments of *all* active streams (including the new one) will meet their deadlines throughout the duration of these streams. This kind of QoS is known as a *deterministic* or *worst-case* guarantee.

As continuous multimedia data like video and audio are usually encoded with variable bit rate and the detailed quantitative behavior of the server can be characterized only stochastically, worst-case guarantees will typically result in significant underutilization of the server's performance capacity. In virtually all digital library applications, clients would, however, tolerate infrequent "glitches" that result from not meeting a delivery deadline or dropping a data fragment. Often such glitches would even be unnoticed by the user provided they occur infrequently enough. This consideration leads to the notion of *stochastic QoS* where smooth playback of video or audio is guaranteed with a certain, specifiable, probability close to one. To this end the admission control needs a stochastic model that quantifies the glitch probability and other derived metrics for given data characteristics and server resources [10]. In networking this approach is known as *statistical multiplexing* [18]. The first part of the talk will elaborate on such stochatic models and will show how they can be leveraged for a practically viable multimedia server.

The second part of the talk generalizes this notion of QoS to other types of data and user requests. Digital libraries maintain large volumes of "discrete" data objects such as text, XML, or image documents in addition to the above mentioned continuous data. QoS for discrete data requests has largely been ignored by the multimedia community but is equally important from an application viewpoint. The server should guarantee that it can provide a requested discrete data object within a specified, user-acceptable, response time with probability close to one (e.g., within 2 seconds in 95 percent of all requests). Multimedia digital libraries should generally be treated as a "soft" *real-time application* [5, 15], where guarantees would be "merely" stochastic as worst-case guarantees are practically infeasible with a huge number of potentially active clients.

For a given application with specific data and workload characteristics as well as QoS requirements, the server needs to be properly configured so as to provide guarantees for both continuous data streams as well as discrete data requests. The talk will present how stochastic models can be used for this purpose. With advanced models it is even feasible to configure a mixed workload server with dynamic resource sharing among both of these workload classes (i.e., without a fixed partitioning of disks or memory) and appropriate scheduling [11, 12].

A particularly important mechanism for improving the response time of discrete data requests to digital libraries is caching and prefetching, along the entire storage hierarchy of client, proxy, and server caches (see, e.g., [2, 7, 14]). The key for a good caching and prefetching policy is the (speculative) prediction of near-future data requests based on statistical profiling (see, e.g., [3, 13]). Advanced policies along these lines take into account several dimensions: the estimated probability of accessing a data item in the near future, the current context of user sessions, the variable size of data objects, and the different performance or availability characteristics of different servers that a client or proxy interacts with. The third and last part of the talk will discuss such advanced caching and prefetching policies (e.g., based on continuous-time Markov chain models) [6, 17] and their

ADVANCES IN DATA MANAGEMENT 2000
Krithi Ramamritham, T.M. Vijayaraman (Editors)
Tata McGraw-Hill Publishing Company Ltd.

benefit with regard to QoS guarantees.

The presented work is part of a strategic research direction that aims at a comprehensive understanding of general *quality guarantees* for information services [16]. In addition to the performance-oriented metrics that are in the focus of this talk, such a generalized notion of QoS should include service availability and failure masking, behavioral properties such as guaranteed termination with certain results, and the accuracy, completeness, timeliness, credibility, and cost-effectivity of search results. Understanding the tradeoffs, interdependencies, and composability among these broader aspects of QoS poses major research challenges that are of crucial importance for next-generation multimedia digital libraries.

1. REFERENCES

[1] S. Christodoulakis, P. Triantafillou: Research and Development Issues for Large-Scale Multimedia Information Systems, ACM Computing Surveys Vol.27 No.4, 1995.

[2] L. Fan, P. Cao, W. Lin, Q. Jacobson: Web Prefetching Between Low-Bandwidth Clients and Proxies: Potential and Performance, ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Atlanta, 1999.

[3] M. Friedrich, S. Hollfelder, K. Aberer: Stochastic Resource Prediction and Admission for Interactive Sessions on Multimedia Servers, ACM International Conference on Multimedia, Los Angeles, 2000.

[4] W.I. Grosky, R. Jain, R. Mehrotra: The Handbook of Multimedia Information Management, Prentice Hall, 1997.

[5] J. Haritsa, K. Ramamritham: Real-Time Database Systems in the New Millenium (Guest Editorial), International Journal of Real-Time Systems, Special Issue on Real-Time Databases, Vol.19 No.3, Kluwer, 2000.

[6] A. Kraiss and G. Weikum: Integrated Document Caching and Prefetching in Storage Hierarchies Based On Markov-Chain Predictions, The VLDB Journal Vol.7 No.3, Springer, 1998.

[7] B. Krishnamurthy, C.E. Wills: Proxy Cache Coherency and Replacement – Towards a More Complete Picture, IEEE CS International Conference on Distributed Computing Systems, Austin, 1999.

[8] J.F. Kurose, K.W. Ross: Computer Networking: A Top-Down Approach Featuring the Internet, Addison-Wesley, 2001.

[9] F. Kuo, W. Effelsberg, J.J. Garcia-Luna-Aceves: Multimedia Communications: Protocols and Applications, Prentice Hall, 1998.

[10] G. Nerjes, P. Muth, and G. Weikum: Stochastic Service Guarantees for Continuous Data on Multi-Zone Disks, Proceedings of the ACM International Symposium on Principles of Database Systems (PODS), Tucson, Arizona, 1997.

[11] G. Nerjes, P. Muth, G. Weikum: A Performance Model of Mixed-Workload Multimedia Information Servers, 10th GI/NTG Conference on Performance Evaluation of Computer and Communication Systems, Trier, Germany, 1999.

[12] G. Nerjes, Y. Romboyannakis, P. Muth, M. Paterakis, P. Triantafillou, G. Weikum: Incremental Scheduling of Mixed Workloads in Multimedia Information Servers, International Journal on Multimedia Tools and Applications, Vol.11 No.1, Kluwer, 2000.

[13] R.R. Sarukkai: Link Prediction and Path Analysis Using Markov Chains, International World Wide Web Conference, Amsterdam, 2000.

[14] J. Shim, P. Scheuermann, R. Vingralek: Proxy Cache Algorithms: Design, Implementation, and Performance, IEEE Transactions on Knowledge and Data Engineering, Vol.11 No.4, 1999.

[15] J.A. Stankovic et al.: Strategic Directions in Real-Time and Embedded Systems, ACM Computing Surveys Vol.28 No.4, 1996.

[16] G. Weikum: Towards Guaranteed Quality and Dependability of Information Services) (Invited Keynote), 8th German Conference on Databases in Office, Engineering, and Scientific Applications, Freiburg, 1999, Springer.

[17] G. Weikum, A.C. König, A. Kraiss, M. Sinnwell: Towards Self-Tuning Memory Management for Data Servers, IEEE Data Engineering Bulletin Vol.22 No.2, 1999.

[18] Z.-L. Zhang, J. Kurose, J. Salehi, D. Towsley: Smoothing, Statistical Multiplexing and Call Admission Control for Stored Video, IEEE Journal on Selected Areas in Communications, 1997.