Title	A performance model for a local VoD system
Author(s)	Zheng, J; Chan, SC; Ho, KL
Citation	The 1st International Conference on Information, Communications and Signal Processing (ICICS'97), Singapore, 9-12 September 1997. In Conference Proceedings, 1997, v. 3, p. 1638-1641
Issued Date	1997
URL	http://hdl.handle.net/10722/54052
Rights	©1997 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

A Performance Model for a Local VoD System

J. Zheng, S. C. Chan and K. L. Ho

Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
Email: jzheng@eee.hku.hk, scchan@eee.hku.hk
Tel: (852)28578408   Fax: (852)25598738

Abstract

A multimedia information system provides an effective means to convey information to users. This paper studies the problem of carrying out video on demand (VoD) application over a high speed LAN to support Computer Supported Collaborative Working (CSCW) for people working in a local collaborative environment. A multimedia information system using HP 100VG-AnyLAN is proposed and setup for this purpose. Based on this system, a performance model is developed, which can be used to determine the requirement of network bandwidth and evaluate the system performance.

1 Introduction

Nowadays, information technology has become indispensable in the daily operation of business firms as well as manufacturing companies. Due to the effectiveness of multimedia objects (like images, video and audio) in conveying information, there is a great demand for handling these objects both in information systems and over communication networks. Recent advances in computing and networking technology have made on-line access to multimedia information both possible and cost effective. And with the advent of high speed LANs like 100 Mbps Ethernet and 100VG-AnyLAN, it becomes possible to implement such multimedia information systems in a local environment. A local multimedia information system provides an effective means to convey information to users, which can support Computer Supported Collaborative working (CSCW) and raise the working efficiency for people working in a local collaborative environment.

There are many technical issues on the design and implementation of such multimedia information systems, such as system architectures, bandwidth and storage requirements, communications protocols, etc. This paper studies the problem of carrying out VoD application over a high speed LAN to support CSCW. A multimedia information system is proposed and setup for this purpose. We choose Windows NT Server as the video server, HP 100VG-AnyLAN as the networking platform and

Microsoft ActiveMovie as the MPEG playback engine. Based on this system, a performance model is developed, which can be used to determine the requirement of network bandwidth and evaluate the system performance. In Section 2, we briefly describe the architecture of our multimedia information system. In Section 3, we address the networking platform in more details. Section 4 is devoted to the performance model of the system. And in Section 5, we give the conclusion.

2 System Architecture

The architecture of our Multimedia Information system is shown in Figure 1, which consists of three main elements: the video server, the high speed network and the users. This system can actually be considered as a small scale video on demand system over a high speed LAN as we are only concerned with the problem of carrying out the real time playback of videos in this paper. A centralized video server architecture is used for its relatively simple management, which is suitable for a local environment. The video server stores compressed MPEG-1 video data that can be requested by users connected to it via the network. All the user requests are sent to and served at one site. In order to support the real time playback of the videos, the video server must guarantee the necessary storage bandwidth as well as the timely delivery of the video data to the network. On the other hand, the network must provide the required transmission bandwidth and control the end-to-end delay and delay jitter. The client program of the user must guarantee the real time display of the compressed MPEG video data. For these purposes, we choose Windows NT Server as the video server for its multitask operating system, 100VG-AnyLAN as the networking platform for its high performance on time-critical applications and ActiveMovie as the MPEG playback engine[1]. The user operating system may be Windows NT, Windows 95, Windows 3.11 workstations or personal computers.

Windows NT [1] is a 32-bit multitask operating system which can support a wide rage of network operating systems such as Novel NetWare, MS-DOS LAN Manager, Windows 3.11 and Windows 95.

ActiveMovie [1] is the next generation of cross-platform digital video technology for the desktops. With Active-Movie, one can create and deliver high quality synchronized audio and video on multiple platforms such as Windows 95 and Windows NT.
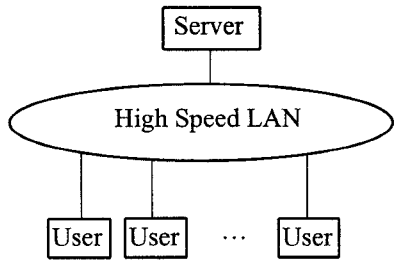

Fig. 1 Architecture of the VoD system

3 Networking Platform

In this section, we will address some considerations in the networking platform and give a brief description of 100VG-AnyLAN.

3.1 Some considerations

A time-critical multimedia application such as VoD service requires a high speed network to guarantee the timely delivery of video data. In recent years, many new networking technologies have been developed for such time-critical multimedia applications, such as ATM switches, FDDI and 100 Mbps LANs. Due to the relatively high cost of ATM switches, they are mainly used as a backbone to connect remote sites. And FDDI is another alternative for large diameter networks. For this reason, 100 Mbps Ethernet and 100VG-AnyLAN (IEEE 802.12 draft standard) become very attractive for our purpose because of their high performances, low costs and compatibility to existing 10Base-T networks. Although 100 Mbps Ethernet has a transmission rate of 100 Mbps on the physical layer and the lowest cost, it still suffers from the disadvantages of collision-based protocols with unpredictable access delay. Therefore it may not be able to deliver multimedia objects to users when the traffic is high. In fact, performances start to degrade significantly when the network traffic approaches 70 Mbps [2].

100VG-AnyLAN uses a demand priority protocol and a star topology. The two priority levels and round robin arbitration mechanism make it possible to guarantee bandwidth to applications and to keep the access delay within bounds. Bandwidth and access delay depend on the size of the network and the way it is configured. With this high speed network, it is now possible to perform time-critical applications such as VoD service and video conferencing. For this reason, 100VG-AnyLAN is selected as our networking platform.

3.2 100VG-AnyLAN
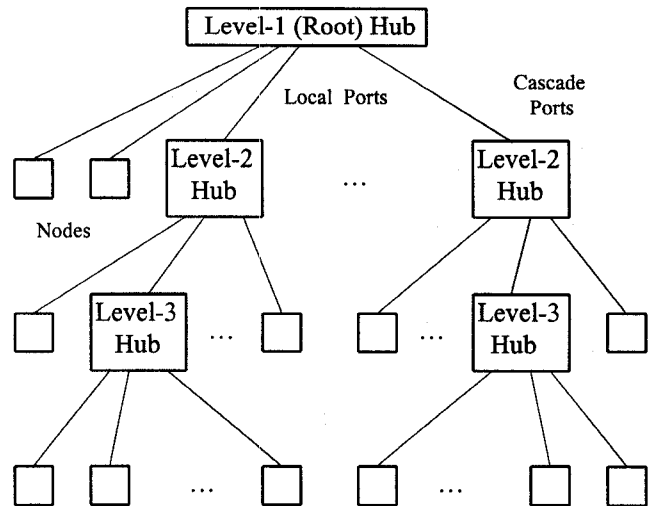
3.2.1 Network Topology

Fig. 2 A cascaded star network

The basic topology used by 100VG-AnyLAN networks is the star configuration [3]. A network can contain several levels of hubs interconnected in a cascade with the top hub being the level 1 (root) hub, as shown in Figure 2. Each hub has two or more local ports and can optionally have one cascade port for connection to a higher level hub. Nodes can be user stations, bridges to other networks, LAN analyzers, or lower level hubs. The maximum number of nodes that can be connected to the network is dependent on the traffic each node generates. The maximum topology diameters and the number of cascading levels are limited by the allowable delay between the node and the root hub, and can be calculated for any proposed configuration.

3.2.2 Demand Priority Protocol

Demand priority is the media access control protocol defined in the IEEE 802.12 standard [4]. Various techniques are used to ensure fairness of access for all nodes and to guarantee access time for individual nodes. Round-robin procedures are used to give each node an equal opportunity to transmit data. The purpose is to ensure that no node can send two successive normal-priority frames until all other nodes have the opportunity to send a frame at normal priority. Two priority levels are provided so that time-critical traffic such as interactive video, audio and multimedia can be given higher priority with guaranteed low delay. A bandwidth allocator can be introduced to control the amount of bandwidth each application can use. These techniques make demand priority able to accommodate multimedia applications with guaranteed performance levels while allowing normal traffic to use the remaining bandwidth in a fair manner.

3.2.3 Guaranteed Performance

Bandwidth and access delay depend on the size of the network and the way it is configured [4]. Nodes using high-priority traffic may need to be configured so that the amount of bandwidth they can use is restricted. Without any other form of control, nodes wishing to send high priority traffic will automatically be allowed to send an equal number of high priority frames. Any bandwidth not used by this high priority traffic is then automatically shared among all nodes wishing to send normal priority traffic. When there is excessive high priority traffic, the bandwidth allowed for high priority requests needs to be restricted in some way so that normal priority traffic is never completely stopped [5]. The deterministic behavior of the round-robin selection procedure makes it easy to establish a reasonably accurate estimate of the maximum access delay that a high-priority packet will experience. However, access delay for a normal-priority request is more difficult to calculate.

4 Performance Model

In this section, we will develop a queuing model for our VoD system and discuss the system performance with a focus on the blocking probability and service response delay at the video server as well as the evaluation of the required bandwidth.

4.1 Model

Our VoD system uses a centralized video server architecture for its relatively simple management. The video server stores compressed MPEG videos that can be requested by the users. All the user requests are sent to and served at one site. As different users may request for different videos or even different portions of the same video simultaneously, the video server must be able to serve multiple concurrent user requests. However, since the video server has a limited I/O transfer rate, it can only serve a limited number of user requests. Therefore when a request for a video is received, the server will first decide whether this request can be accepted. If the video server cannot accept additional requests, the request will have to queue in the system or even be blocked. If the request is accepted, the server will allocate resources to handle this request.

From the viewpoint of queuing theory, a video server can be considered as a serving node with a queue. The capacity of the queue can either be finite or infinite, depending on the system design. Since the video server can serve a maximum of c user requests simultaneously, it can be considered as c virtual servers, each of which is held by a user during his connection [6]. Since the number of users in a local collaborative working environment is generally not very large and these users generate service requests independently, it is more exact to model the arrival

process to the video server as a quasi-random input with finite sources than as a Poisson process. Furthermore, considering the effect of user interaction and varying lengths of different videos, we can further assume that the service time of each virtual server is exponentially distributed. Therefore the video server can be modeled as an M/M/c/K queuing system with n sources, where n is the number of users in the VoD system, c is the number of the virtual servers in the video server and K is the capacity of the queuing system.

4.2 Performance

Now we discuss the $M / M / c / K$ queuing model with n sources [7]. The video server can be designed to be a loss system or a delay system. Since $K \geq n$ corresponds to a $M / M / c / \infty$ model, we assume $c \leq K < n$ with $K = c$ being the loss system and $K > c$ being the delay system. Letting γ be the arrival rate from an idle source and μ^{-1} the mean service time, we have the birth rate and death rate, respectively, as

$$\lambda_k = \begin{cases} \gamma(n-k) & (0 \leq k \leq K-1) \\ 0 & otherwise \end{cases}$$

$$\mu_k = \begin{cases} k\mu & (0 \leq k < c) \\ c\mu & (k \geq c) \end{cases}$$

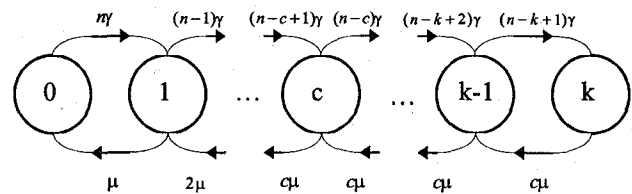Figure 3 shows the state-transition diagram for an M/M/c/K model with n sources.

Fig. 3 State transition of M/M/c/K with n sources

4.2.1 Blocking Probability and Service Response Delay

We can obtain the steady state probability as follows

$$p_k = \begin{cases} \binom{n}{k} \hat{a}^k p_0 & (0 \leq k \leq c-1) \\ \binom{n}{k} \dfrac{k!}{c!} c^{c-k} \hat{a}^k p_0 & (c \leq k \leq K) \end{cases}$$

With the normalization theorem, we have

$$p_0 = [\sum_{k=0}^{c} \binom{n}{k} \hat{a}^k + \sum_{k=c}^{K} \binom{n}{k} \dfrac{k!}{c!} c^{c-k} \hat{a}^k]^{-1}$$

where $\hat{a} = \gamma / \mu$ is the traffic load offered by each idle source. Let π_k be the probability that k requests exist in

the system just prior to a request arrival in the steady state. It can be proved that [7]

$$\pi_k[n] = p_k[n-1] \quad (k = 0,1,\ldots,K)$$

With this equality and the above stationary distribution of system states, we can find the blocking probability P_B and the average service response delay D.

For a loss system, $c = K$. When a request is received and the server cannot take in additional requests, the request will be blocked. In this case, we have

$$P_B = \pi_c[n] = p_c[n-1] = \binom{n-1}{c}\hat{a}^c \,/\, \sum_{j=0}^{c}\binom{n-1}{j}\hat{a}^j$$

For a delay system, $c < K < n$. When a request is received and the server cannot take in additional requests, the request may wait in the queue. If there are insufficient waiting positions in the queue, it will be blocked. In this case, we have

$$P_B = \pi_K[n] = p_K[n-1] = \binom{n-1}{K}\frac{K!}{c!}c^{c-K}\hat{a}^K p_0$$

$$D = \sum_{k=0}^{K-c}\frac{k+1}{c\mu}\binom{n-1}{c+k}\frac{(c+k)!}{c!}c^{-k}\hat{a}^{c+k}p_0$$

From the above results, we can see that all n, c and K have an impact on the blocking probability and service response delay. Since the total I/O transfer rate of the video server is limited, the number c of virtual servers will vary with the average data delivery rate for each video connection. If we assume that the total I/O transfer rate available for the video services is R and the average data delivery rate for each video connection is r, we can obtain c=R/r. This means that the average data delivery rate r also has an impact on the system performance. With different data delivery rates, the blocking probability and service response delay will vary significantly for the same user request arrival rate.

4.2.2 Evaluation of Required Bandwidth

Since we use 100VG-AnyLAN as our networking platform, the bandwidth management is simple to implement in our system. The demand priority protocol and round robin arbitration mechanism allows time-critical applications to have higher priority and deterministic behavior. Therefore, if we assume that all the video services have a fixed average data delivery rate r, it is possible to evaluate the total bandwidth required for these video services. This is the main reason that we choose MPEG for the video services in our study. With the results above, we can obtain the distribution of the bandwidth required, in terms of units of r, which is given by the distribution of the number of busy virtual servers, $\{p_k, k = 0,1,\ldots,c\}$. The average bandwidth required is thus

$$B = \sum_{k=0}^{c} kp_k + \sum_{k=c+1}^{K} cp_k$$

in units of r, where c is the number of virtual servers and K is the capacity of the queuing system.

5 Conclusion

In this paper, we have presented a system to study the problem of carrying out VoD application over a high speed LAN to support CSCW. A performance model is proposed for this system, which can be used to determine the requirement of network bandwidth and evaluate the system performances. The results obtained are useful to the design of the system. Further studies will mainly include the communication protocols to support synchronization of multimedia objects and the traffic control algorithms to support VoD and video conferencing services.

References

[1] K. H. Lee, "Transmission of multimedia data over a high speed computer network", Report on the final year project, Dept. of EEE, The University of Hong Kong

[2] David Strom et al., "Network upgrade systems: Finding a new set of pipes", INFORWORD, June 6, 1996

[3] Alan R. Albrecht and Patricia A. Thaler, "Introduction to 100VG-AnyLAN and the IEEE 802.12 Local Area Network Standard", Hewlett-Packard Journal, pp. 6-12, August 1995.

[4] Alan R. Albrecht and Michael P. Spratt et al., "Demand Priority Protocol", Hewlett-Packard Journal, pp.13-16, August 1995.

[5] John R. Grinham and Michael P. Spratt, "Multimedia Applications and 100VG-AnyLAN", Hewlett-Packard Journal, pp.33-38, August 1995.

[6] O.K.Li et al., "Performance Model of Interactive Video-on Demand Systems", IEEE Journal on Selected Areas in Communications, Vol. 14, No. 16, pp.1099-1109, August 1996

[7] Haruo Akimaru and Konosuke Kawashima, "Teletraffic: Theory and Applications", pp. 13-35, Springer-Verlag, 1993