한국시뮬레이션학회 논문지
Vol. 15, No. 2, pp. 23-29 (2006. 6)

시뮬레이션을 통한 멀티미디어 스트림 서버의 성능 민감도 분석

박진원[1†] · 박종원[2]

Sensitivity Analysis on the Performance of Multimedia Stream Server using Simulation

Jin-Won Park · Chong-Won Park

ABSTRACT

This paper studies the Internet stream server which serves multimedia stream services such as VoD and MoD. The Internet multimedia stream server is developed to provide 50 users with continuous multimedia stream contents simultaneously. This paper focuses on introducing design concepts, operation details and on evaluating system performance and doing sensitivity analysis with respect to the change of the system components. The performance evaluation is performed using simulation in order to identify the system component that mostly affects the system performance. The sensitivity analysis shows the service delay rates of the multimedia stream server depending on the user's service request pattern. Also, the analysis is made on the utilization rates of the core system components such as the PCI bus and the Ethernet card, and on the system performance change by adding more system components.

Key words : Simulation, Stream server, Delay rate, Utilization rate, Sensitivity analysis

요 약

본 논문은 인터넷을 통해 제공되는 VoD나 MoD 등의 멀티미디어 스트림 서비스를 위한 인터넷 스트림 서버의 사용자 요구사항을 성능 측면에서 만족시키는 방안에 대해 연구한 것이다. 본 연구의 대상인 인터넷 멀티미디어 스트림 서버는 동시에 50명의 사용자에게 화면이 끊어지지 않고 연속적으로 서비스를 제공하는 것이 목표이다. 본 논문은 이러한 요구사항을 만족시키기 위한 설계 개념 및 동작 방식에 대해 소개하고, 설계된 멀티미디어 스트림 서버를 대상으로 성능을 시뮬레이션을 통하여 분석하며 시스템 요소 변화에 따른 성능 민감도를 분석하였다. 이는 멀티미디어 스트림 서버를 개발하는 과정에서 시스템 성능에 큰 영향을 미치는 시스템 구성 요소를 분석하여 사용자 요구사항을 만족시키는 최적의 시스템을 설계 개발하기 위해 필요한 작업이다. 그리고, 성능 민감도 분석을 통해 사용자의 서비스 요구 패턴 변화에 대해 멀티미디어 스트림 서버가 보여주는 서비스 지연 확률을 분석하였고, PCI bus와 같은 중요 시스템 요소의 사용률도 분석하였으며, 시스템 자원을 추가로 투입할 때 나타나는 시스템 성능 변화도 분석하였다.

주요어 : 한글 키워드 기재해주세요

1. Introduction

The multimedia stream server is defined as an Internet server for various stream services such as MoD(Music on Demand) and VoD(Video on Demand) services. A typical multimedia stream server is designed by employing processors, a memory unit, a PCI bus, Gigabit ethernet devices, a TOE(TCP/IP Off-load Engine) and disks for providing multimedia data stream services efficiently.

The optimal design for multimedia stream servers may include price performance evaluation using mathematical analysis, statistical study, simulation experiments and some optimization techniques. This study focuses on evaluating the system performance

* 본 연구는 2003년 한국과학재단 지역대학 우수과학자 지원연구사업의 일환으로 수행되었음.
(R05-2003-000-10212-0)
2006년 1월 27일 접수, 2006년 5월 31일 채택
[1] 홍익대학교 과학기술대학 컴퓨터정보통신공학과
[2] (주) 다코스 E&I
주 저 자 : 박진원
교신저자 : 박진원
E-mail; jinon@hongik.ac.kr

and doing sensitivity analysis based on the previous works on the design and analysis on a typical Internet stream server by Park[1] and Park[2]. The performance analysis on the multimedia stream server is vital issue in designing the server, since the identification and the optimal allocation of the system components in the design stage is efficient and cost saving if it is done before the actual system is built.

The multimedia stream services have the characteristics of supplying users with stream data without discontinuation. Thus, the performance evaluation study is concentrated on calculating the delay rate of pumping stream data, where the delay rate is the portion of stream data being transmitted after the predetermined time limit. The utilization rates of the system components, especially the PCI bus and the Gigabit ethernet, are also the target of the performance study, where we expect the balance of the utilization rates between the components.

Early studies on the performance of Internet servers dealt with the performance of multimedia server allowing some faults[3], and with the disk scheduling algorithms affecting the VoD server performance[4]. Also, Lee. et. al.[5] did research on the performance of RAID system that greatly affects the overall performance of an Internet server. Wu, et. al.,[6] introduced the performance issues for the design of stream servers.

Recently, Yu, et. al.[7] proposed a model for selecting and connecting the replicated server that provides optimal services. Son, et. al,[8] suggested a new transmission technique to overcome the discontinuity problem for stored variable bit rate video, and Kwon, et.al.[9] proposed efficient techniques to solve the storage or network I/O bottleneck problem of the VoD system.

However, the multimedia stream server has evolved to include new components such as a TOE(TCP/IP Offload Engine), SAN(Storage Area Network), NAS (Network Attached Storage), and has been changed to handle specialized services. Thus, the multimedia stream server has been designed for serving limited scope of services such as stream services, and has the form of hierarchical structure in some cases.

The performance issue in an Internet stream server is important since the stream services include the additional function to process high speed multimedia stream services compared to typical Internet servers. Some analysis on the operational characteristics in a multimedia stream server is described in [10-12]. However, no major research has been done on this change for the recent multimedia stream server architecture.

Park[2] recently studied the operational characteristics on the Internet stream server consisting of processors, a TOE, a PCI bus, Gigabit ethernet devices and a memory buffer unit, found the right way of the operation of giving the priority to user requests of stream services. However, the focus of the research was on the operational scheme for avoiding the deadlock phenomena, but lacked in the study on the performance issue.

This paper shortly introduces the architecture of the multimedia stream server, which is designed for performing stream services in section 2. In section 3, a discrete event simulation model using AweSim is developed for testing whether the multimedia stream server can serve stream services as designed. The simulation model is used for examining the performance of the multimedia stream server in terms of the delay rate, and the utilization rates of the system components. Based on the analysis on the simulation experimental results, we give more simulation results on the sensitivity on the performance of the multimedia stream server in section 4. Final conclusion is made in section 5.

2. Design Concepts of the multimedia stream server

The design of the multimedia stream server(the server in short) is shown in Figure 1. The design concepts have been described in Park[1]. Summarizing the design concepts of the server, we may notice that the server has a dedicated processor for handling TCP/IP protocols in a offload fashion[10,13-14], and has a simple data path using buffer memory on the I/O bus

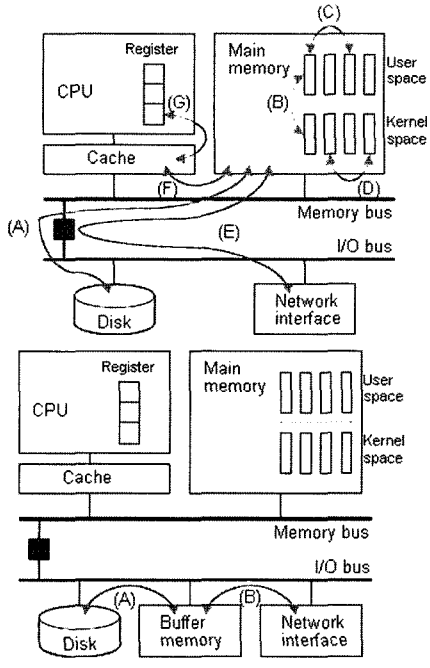


Fig. 1. Data transfer and Copy operations of the multimedia stream server

as shown in Figure 1.

The upper part in Figure 1 is the traditional computer server systems, showing the data transfer from the disk to the memory(A) and eventually to the final data manipulation from the cache to CPU(G). The detail of the operation is described in Park[1].

The lower part in Figure 1 is the data path of the proposed stream server, showing the data transfer from the disk to the buffer memory(A) to the network interface(B). Clearly, the data path in the proposed stream server is simplified compared to that of the traditional server, resulting in performance enhancement.

The server also adopts a multimedia file system that can consider the characteristics of continuous form of stream data such as several megabyte data blocks. The server also called on to use striping technique such as RAID(Redundant Array of Inexpensive Disks) for preventing the degradation of disk reading speed[14].

The multimedia stream server considered in this paper is assumed to have the user requirements that can

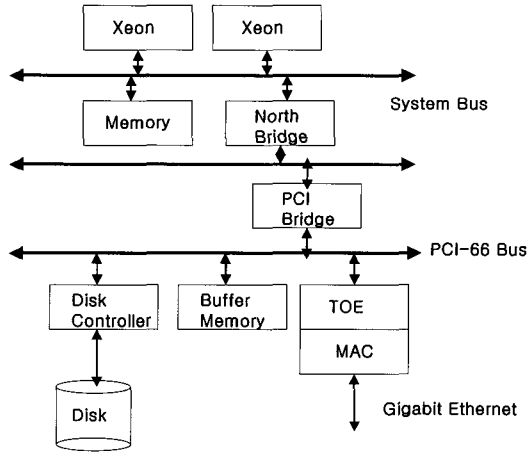


Fig. 2. Schematic diagram of the multimedia stream server

serve upto 50 users simultaneously with 20 Mbps (2.5MB/s) multimedia stream data. Also, the server should have a storage system that can store a large volume of multimedia data, and should maintain scalability, high availability and the system resource management function to maximize the performance. The schematic diagram of the server is depicted in Figure 2.

3. Performance Model of the multimedia stream server

The AweSim simulation model for the architecture and the operations of the server is developed based on the schematic diagram of Figure 2. The model defines two types of stream services, one with the services for supplying multimedia stream data to general users(user service), and the other with the loading of multimedia stream data from the external global server to the stream server(uploading). These two stream services are assumed to be processed simultaneously.

The user services are assumed to be initiated by the processors giving the command to the disk controller for supplying stream data to a user. Then the corresponding stream data is transferred from the disk to the disk controller(buffer) and to the memory through the PCI bus. The PCI bus and 1 unit of memory space

Table 1. System Parameters in the model

Parameters	Base	Note
Data block size	2MB	1MB, 500KB
PCI-66 Bus	66MHz, 64 bit	1, 2 units
Buffer Memory	256MB (128 units of 2MB)	constant
TOE Buffer	16MB (8 units of 2MB)	constant
Gigabit ethernet device	20 Mbps/user (2.5MB/s)	1, 2 units
Size of 1 set of total stream data	10GB of average 1 movie (5000 units of 2MB)	constant
Number of users	50	20, 30, 40

(2MB) are required to perform the operation.

The stream data stored in the memory is moved to the TOE buffer through the PCI bus, then is transferred to the user by the Gigabit ethernet device. Operation of moving the data to the TOE buffer uses 1 unit of the memory, the PCI bus and 1 unit of the TOE buffer.

Meanwhile, the downloading services are initiated with transferring stream data from the Gigabit ethernet device to the TOE buffer, then are moved to the memory through the PCI bus and are finally done with storing the data to the disks.

The system parameters employed in our simulation model are presented in Table 1. The numbers presented in Table 1 are from the real implementation of the system and from the expert's opinion.

Since the stream services are performed in the unit of 2MB, the transmission time of a block of data(2MB) is set as 3.788ms for the PCI bus and 16ms for the Gigabit ethernet device.

The first stage of the simulation experiments is focused on testing whether the Internet stream server designed as in Figure 1 operates correctly and 50 users can be served simultaneously. In every experimentation, the uploading service is assumed to be concurrently conveyed with 5 movies uploaded from outside to the Internet stream server.

Extensive simulation study performed in Park[2] revealed that the correct order of operations should be

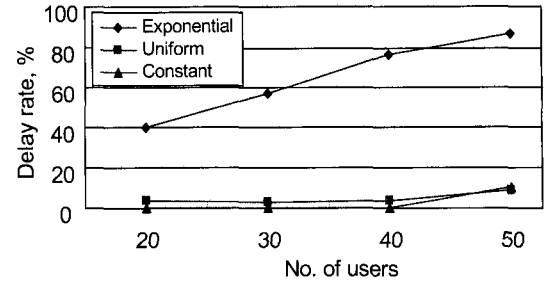


Fig. 3. Delay rate

made in order to avoid the deadlock phenomena of the server. Park[2] showed that the priority to use the PCI bus should be given to the data transfer operations from the buffer memory to the TOE buffer(user service direction) and from the buffer memory to the disks (uploading direction) equivalently. Also, the TOE buffer should be taken by the data transfer operation from the buffer memory to the TOE buffer first(user service direction) and then to the operation from outside to the TOE buffer(uploading direction).

4. Sensitivity Analysis on the performance of a multimedia stream server

The performance experiments were set up to measure the delay rate for the user's service requests, where the delay rate is defined as the portion of services that is not peformed within the predetermined time limit. Also, we assumed 3 different service request patterns, constant, uniform and exponential inter-arrival distributions.

The preliminary experiments showed that if the user's service request pattern is constant, that is, if the inter-arrival time between the user requests is constant, the server can serve upto 40 users simultaneously without any delay if the correct order of the operations is kept. However, if the number of users increases or if the user's service request pattern becomes random, the delay ratio becomes large as in Figure 3. The worst case happens when the service request time is

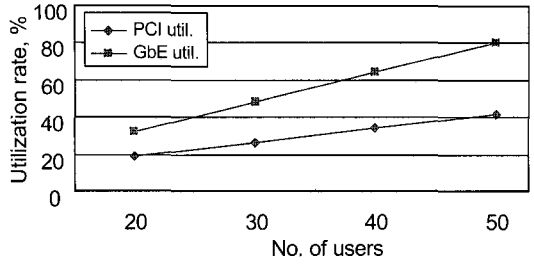


Fig. 4. Utilization rates with constant user request

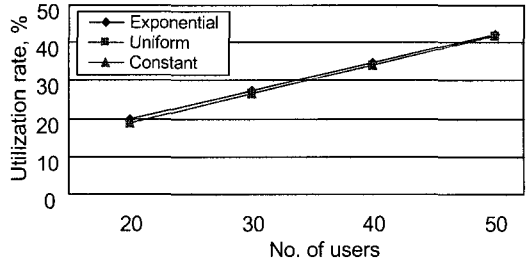


Fig. 5. Utilization rates of the PCI bus

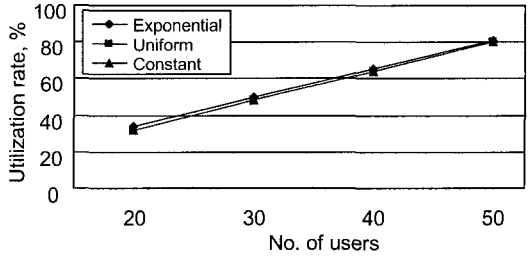


Fig. 6. Utilization rates of the Gigabit ethernet device

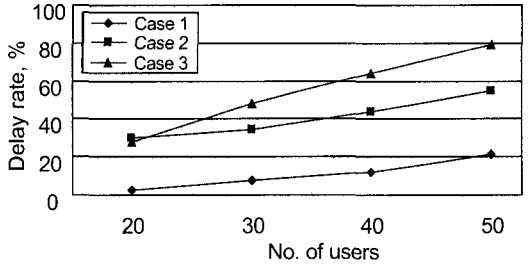


Fig. 7. Sensitivity analysis on increasing the number of components

distributed as exponential. Even with 30 users requesting the stream services, more than 50% of the users experience the delay of data transmission.

The utilization rate, defined as the portion of time that a component is being used, of the system components is another interesting performance factor in the server. Figure 4 shows the utilization rates with respect to the number of users being served simultaneously. We only show the utilization rates of the PCI bus and the Gigabit ethernet, since the memory unit and the TOE buffer have enough capacity to handle the services, resulting in the low level of the utilization rates. Figure 4 is the results for the constant service request pattern.

We have experimented the effect of random service requests on the utilization rates of the multimedia stream server. As we see in Figures 5 and 6, the utilization rates do not change significantly as the inter-arrival time of the user's service request changes from constant to uniform and to exponential distribution.

More analysis has been performed for decreasing the delay rate. Figure 7 shows the delay rate with the exponential inter-request time when we change the capacities of the critical components of the server, the PCI bus and the Gigabit ethernet device.

Case 1 is when we install 2 Gigabit ethernet device and dual PCI buses. As we see in Figure 7, the delay ratio decreases dramatically. With exponential inter-request time and 40 users requesting the stream services, only 12% experiences the delay compared to 76% in Figure 3. In the case of increasing the capacity of Gigabit ethernet device to 2Gbps whereas maintaining single PCI bus in Case 2, the delay rate again decreases a little with 20, 30 user cases, but more with 40, 50 use cases.

When we put dual PCI buses only without increasing the Gigabit ethernet capacity in case 3, the delay rate decreases a little as compared to Figure 3, but remains high with 40, 50 user cases.

Final experiments have been implemented with changing the data block size from 2MB to 1MB and 500KB to see the effect of making small data blocks on the delay rate and the utilization rates of the system components. However, no significant change has been observed, thus the results are not presented here.

5. Conclusion

We presented a design and performance evaluation results for a multimedia stream server. From the results of the performance evaluation study, we come up with the conclusion that we have a number of users may experience the high level of the delay in transmitting the stream data when the user's service request pattern is random. Also, we noticed that the utilization rate among the system components varies with the number of users, but remains relatively unchanged with respect to the user's service request pattern.

Critical components in the stream server in terms of the system performance are PCI bus and Ethernet device. Among them, Ethernet device affects more on the system performance.

When we consider the practical situation when the user's service request pattern is quite random, we may have to install more network devices in order to maintain the low level of the delay rate.

In this study, a perfectly reliable stream server was assumed but an unreliable server may be a target to analyze in the future study. Also, some probabilistic service mechanism can be considered for more realistic situation.

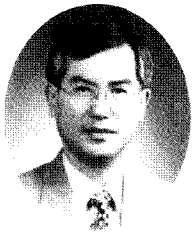
References

1. Park Chong-Won, S.W. Kim, J-W Park, "System Architecture of a Multimedia Streaming Server for the Next Generation Internet," LNCS 3207, Embedded and Ubiquitous Computing, EUC 2004, Aizu-Wakamatsu City, Japan, pp662-671, 2004.8.
2. Park Jin-Won, "Simulation Study on the Stream Server for Deciding the Priority for Using Resources," Journal of Korea Society for Simulation, Vol. 12, No. 4, pp95-102, 2003.12. (in Korean)
3. Park Kiejin, Sungsoo Kim, "A Performance Evaluation of Multimedia-on-Demand Server Using Simulation Method," Journal of Korea Society for Simulation, Vol.7, No.2, pp33-43, 1998.12. (in Korean)
4. Chung Ji Yung, S.S. Kim, "A Design of Real-time VOD Server Simulator," Journal of Korea Society for Simulation, Vol. 9, No. 3, pp65-75, 2000.9. (in Korean)
5. Lee Chan-Su, Y.R. Seong, H.R. Ha, "Modeling and Simulation of a RAID System," Journal of Korea Society for Simulation, Vol. 11, No. 1, pp11-22, 2002. 3. (in Korean)
6. Wu Dapeng, Yiwei Thomas Hou, Wenwu Zhu, Ya-Qin Zhang and Joe M. Peha, "Streaming Video over the Internet, Approaches and Directions," IEEE Trans. on Circuits and Systems for Video Technology, Vol.11, No.3, 2001.
7. Yu Ki-Sung, W.H. Lee, S.J. Ahn, J-W Chung, "Server selection system model and algorithm for resolving replicated server using downstream measurement on server-side," Journal of Korea Society for Simulation, Vol. 14, No. 2, pp1-13, 2005.6. (in Korean)
8. Son Sung-Hoon, Y.C. Baek, "An Adaptive Transmission Scheme for Variable Bit Rate Streaming Video over Internet," The KIPS Transactions: Part A, Vol. 12-A, No.3, pp197-204, 2005.6.
9. Kwon Chun Ja, H.K. Choi, "An Efficient P2P Based Proxy Patching Scheme for Large Scale VOD Systems," The KIPS Transactions: Part A, Vol.12-A, No.5, pp341-354, 2005.10.
10. Adaptec Company: Advantages of a TCP/IP Offload ASIC, White paper.
11. Waldvogel Marcel, Deng W. and Janakiraman R., "Efficient Buffer Management for Scalable Media-on- Demand," IBM Research Report, 2002.
12. Clark D.D. and D.L. Tennenhouse, "Architectural considerations for a new generation of protocols," Proc. of ACM SIGCOMM'90, pp200-208, 1990.
13. Xiran Company: http://www.xi ran.com/solutions -stream.php
14. Thomas Plagemann, Vera Goebel, Pal Halvorsen and Otto Anshus: Operating System Support for Multimedia Systems. The Computer Communication Journal, Elsevier, Vol. 23, No. 3, pp.267-289, 2000.

박 진 원 (jinon@hongik.ac.kr)

1975 Seoul National Univeristy (B.S.)
1982 The Ohio State Univeristy (M.S. and Ph.D. in Industrial and Systems Engineering)
1987 University of Southern Colorado (Asst. Professor)
1988 ETRI (Principal Researcher)
1999 Youngsan University(Asst. Professor)
2000~present, Hongik University (Associate Professor)

Interested Area : System Simulation, Computer Architecture, Performance Evaluation.



박 종 원 (cwpark@dreamwiz.com)

1981 Hanyang Univeristy, Electronics & Telecommunications Eng. (B.S.)
1983 Hanyang Univeristy, Electronics & Telecommunications Eng. (M.S.)
2002 Hanyang Univeristy, Electronics & Telecommunications Eng. (Ph.D.)
1984 ETRI (Principal Researcher)
2005~present, DAKOS E&I Co. Ltd., R&D Center (CTO)

Interested Area : Computer Architecture, Storage System, Computer Communications, Ubiquitous Computing