GUARANTEEING QUALITY OF SERVICE IN INTERACTIVE VIDEO-ON-DEMAND SERVERS

Chih-Yuan Cheng and Yen-Jen Oyang*
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.

and

Meng-Huang Lee
Department of Information Management
Shih Chien University, Taipei, Taiwan, R.O.C.

Abstract

One of the main challenges in disk system design for interactive VOD servers is how to achieve a good quality-of-service (QoS) guarantee. This paper proposes a queueing model for analyzing the I/O bandwidth required in order to achieve a certain level of QoS for a disk system design scheme that provides VCR emulation. The disk system design scheme employs a practice that requests no extra bandwidth to support interactive operations such as fast forward search and fast backward search. The proposed queueing model provides a complete analysis tool for pinning down the resource requirement issue and is verified through simulation for its validity.

1 Introduction

Guaranteeing quality of service (QoS) is one of the main challenges in the design of video-on-demand (VOD) systems that support interactive features such as fast forward search and fast backward search. In recent years, there have been a number of studies focusing on this issue [1, 2, 3, 4, 5]. The study done by Jayanta K. Dey-Sircar and et al. assumed a fast-mode stream plays frames at a rate higher than a normal-mode stream and proposed a queueing model for analyzing the system capacity required in order to achieve a certain level of QoS[1]. The major disadvantage with the operation scheme that Jayanta K. Dey-Sircar and et al. assumed is that a stream will claim higher bandwidth when it switches from the normal playback mode to a fast search mode. As a result, a certain amounts of I/O bandwidth and network bandwidth must be reserved for meeting the additional bandwidth required by streams switching to the fast search modes. This leads to less efficient utilization of system resources.

To avoid this problem, several schemes that support interactive VOD operations based on sampling MPEG[6, 7] GOPs(Group of Pictures) were proposed [2, 3, 4]. These schemes require no extra bandwidth to support interactive operations but suffer the deficiency of unfavorable visual impact, especially in the fast backward search mode.

Shenoy and Vin then proposed a scheme that also requires no extra bandwidth to support interactive operations and provides a good VCR emulation[5]. The problem with the scheme proposed by Shenoy and Vin is that if the system needs to support both fast forward search and fast backward search or to support more than one rates of fast search modes, then multiple video files of about the same size must be created for each video program in the system. Since each video file created in accordance with the scheme is of size no less than the original MPEG file of the program, the disk capacity required are increased by several folds.

This paper proposes a queueing model to analyze the resource requirement for a disk system design scheme that alleviates the deficiencies mentioned above. The disk system scheme employs a practice that provides VCR emulation without requiring extra disk bandwidth to support interactive operations [8, 9]. The practice employed in this paper to support the fast forward and backward searches plays separate MPEG files derived from sampling video frames at a certain rates and compressing the sampled video frames in accordance to the MPEG standard [6, 7]. Figure 1 illustrates the process of creating separate MPEG files to be played in fast search modes. The major advantages of this practice are two folds. First, it facilitates runtime resource allocation. Second, from the aspect of disk

* Corresponding author: Yen-Jen Oyang

Contributed Paper
Manuscript received March 15, 1999

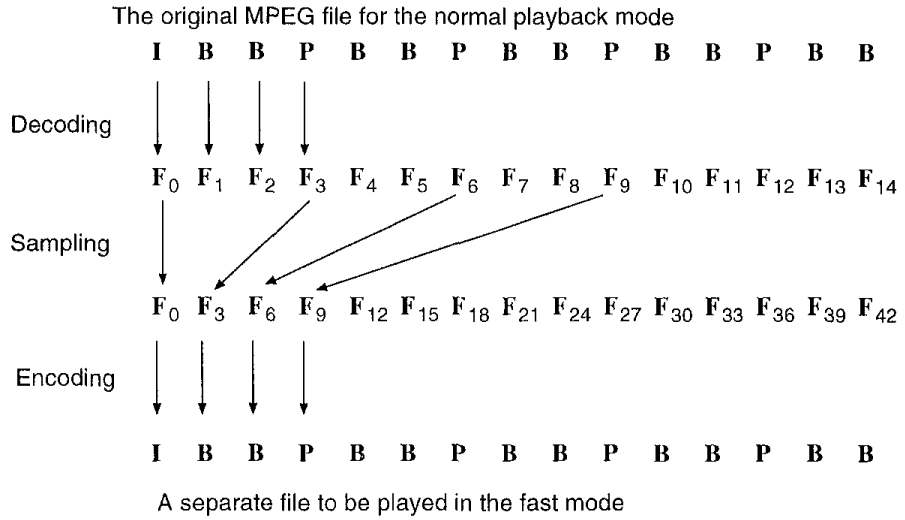


Figure 1: The creation of special MPEG files to be played in fast search modes.

capacity usage, it only introduces a small amount of overhead for supporting fast search modes. When generating a separate MPEG file to be played in a fast search mode, one can select a bit rate no higher than the bit rate of the original MPEG file. As a result, a video stream entering a fast search mode does not need to claim more system resources from the VOD server and the network. This alleviates a lot of run-time resource management problems for supporting interactive VOD operations. Since the quality of individual frames is less sensitive in fast search modes, this practice provides a good tradeoff between system resource management and visual impact. As far as disk capacity is concerned, it is obvious that an additional MPEG file created to support the N-time fast forward or backward search occupies only $1/N$ the amount of disk space required to store the original video program.

With the practice described above employed, the next problem to address is how to figure out the disk system bandwidth required in order to achieve a certain level of QoS. This paper presents a queueing model to tackle this problem and uses simulation to verify the validity of the queueing model. The simulation results reveal that the queueing model provides good estimates of the I/O bandwidth required.

The remaining part of this paper is organized as follows. Section 2 elaborates the disk system design scheme and the proposed queueing model to determine the system resources required for achieving a certain level of QoS. Section 3 verifies the validity of the queueing model through simulation. Finally, the discussion of this paper is concluded in Section 4.

2 Disk system design for guaranteeing QoS

The discussion in this section first focuses on the proposed disk system design scheme that supports interactive VOD features without requiring higher data bandwidth. Then, in 2.2, a queueing model is proposed for determining the required I/O bandwidth of the disk system.

2.1 Disk system architecture

Figure 2 depicts the general disk system architecture. The entire disk system consists of two disk arrays. One disk array stores MPEG files for normal playback. Another disk array stores specially generated MPEG files to be played in fast search modes. The disk array that stores normal playback files provides the disk bandwidth for servicing streams in the normal playback mode, while the disk array that stores fast mode files provides the disk bandwidth for servicing streams in fast search modes. Though there are two disk arrays, a stream entering a fast search mode does not relinquish disk bandwidth that it claims from the disk array storing the normal mode files. The retained disk bandwidth can be used by the same stream when it returns from the fast search mode.

With the disk system configuration described above and with the bit rate of a fast mode file being carefully selected to be no higher than the bit rate of the corresponding normal playback file, a stream switching from the normal playback mode to a fast search mode requires no extra run-time system resources from the VOD server and the network. This alleviates a lot of run-time resource manage-

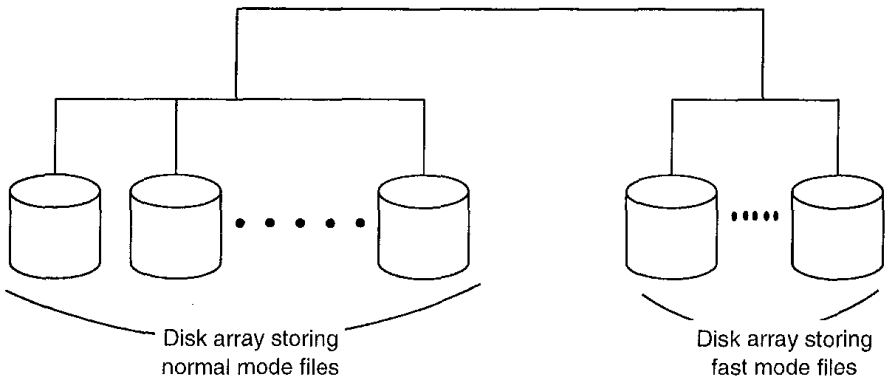


Figure 2: The disk system architecture.

ment problems in interactive VOD systems design. In particular, the VOD server does not need to pump more data per unit of time and the network does not need to allocate more data bandwidth to service a stream that switches from the normal playback mode to a fast search mode.

2.2 The queueing model

In order to guarantee quality of service (QoS), a queueing model must be developed to determine the amounts of system resources required. Figure 3 shows a 2-dimensional queueing model that emulates the behavior of the system. However, the 2-dimensional queueing model is too complicated to be analyzed mathematically. Therefore, an approximate approach is adopted. Figure 4 shows the two Markov processes used to model the behavior of the system. The first Markov process, which has only one state as shown in Figure 4(a), models arrival of new streams and termination of existing streams in the VOD system. The second Markov process, which has two states as shown in Figure 4(b), models streams switching back and forth between the normal playback mode and either one of the two fast search modes, the fast forward search mode and the fast backward search mode. In the first Markov process, the arrival behavior of new streams is a Poisson process with parameter λ_0 and the duration of a stream staying in the system is exponentially distributed with mean $\frac{1}{\mu_0}$. The first Markov process is used to figure out the amount of disk bandwidth that the VOD system must provide in order to reduce the average waiting time of an incoming request for creating a new stream to the level given in the QoS specification. In the second Markov process, the interval during which a stream stays in the normal playback mode before switching to a fast search mode is exponentially distributed with mean $\frac{1}{\lambda_1}$ and the interval during which a stream stays in a fast search mode is exponentially distributed with mean $\frac{1}{\mu_1}$. The second Markov process is used to figure out the amount of bandwidth that the disk array storing fast mode MPEG files must provide.

According to the queueing theory developed for the Markov process shown in Figure 4(a) [10], if the average waiting time is the primary QoS criterion, then the number of streams that the disk array storing normal playback files is able to service at one time, denoted by m, must satisfy the following inequality in order to reduce the average waiting time of a new stream to a certain level:

$$\frac{\left(\frac{(m\rho_0)^m}{\lambda_0 m!}\right)\left(\frac{\rho_0}{(1-\rho_0)^2}\right)}{\sum_{k=0}^{m-1}\frac{(m\rho_0)^k}{k!}+\left(\frac{(m\rho_0)^m}{m!}\right)\left(\frac{1}{1-\rho_0}\right)} \le Q_n, \quad (1)$$

where

$$\rho_0 = \frac{\lambda_0}{m\mu_0},$$

and Q_n is the allowed average waiting time given in the QoS specification. The remaining problem after applying the inequality (1) above is that the video streams in the system may require different amounts of bandwidth to play. Therefore, just calculating the number of streams that the system needs to support does not yield the amount of bandwidth required. This paper uses the product of the weighted average bandwidth and the number of streams that the disk array must be able to service at one time, which is denoted by m, to determine the total amount of bandwidth required. That is,

$$\text{total amount of disk bandwidth required} = m\sum_{i=1}^{l} p_i b_i,$$

where l is the number of video programs stored in the VOD system, p_i the probability that program i is selected by a

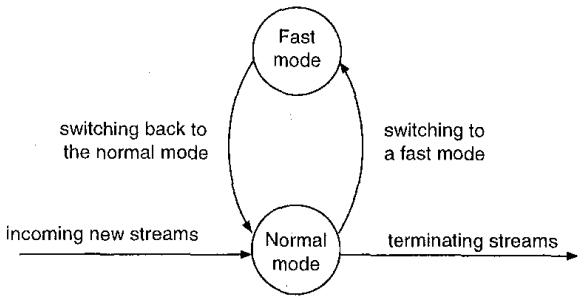
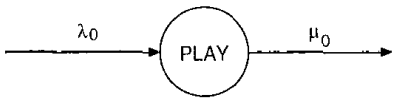
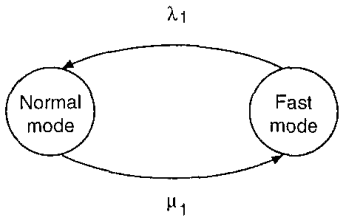


Figure 3: A 2-dimensional queueing model that emulates the system behavior.



(a) The Markov process that models arrival of new streams and termination of existing streams in the VOD system.



(b) The Markov process that models streams switching back and forth between the normal playback mode and the fast search modes.

Figure 4: The proposed queueing model.

newly-created stream, and b_i is the bandwidth required to play program i. Nevertheless, using the weighted average bandwidth introduces a source of discrepancy between the required amount of bandwidth derived from the queueing model and the real world requirement and must be verified through simulation.

With the amount of bandwidth that the disk array storing normal mode files must provide determined, the next issue is to determine the amount of bandwidth that the disk array storing fast mode files must provide in order to reduce the average waiting time observed by a stream switching from the normal playback mode to a fast mode to less than the value given in the QoS specification, denoted by Q_f, in the following discussion. According to the queueing theory developed for the Markov process shown in Figure 4(b)[10], if there are s streams present in the VOD system and the disk array storing fast mode files can service t streams at one time, then the expected waiting time observed by a stream switching from the normal mode to a fast mode is equal to

$$\begin{cases} \sum_{j=t}^{s} q_{sj} \dfrac{j-t+1}{t\mu_1} & \text{if } i > t \\ 0 & \text{if } i \le t, \end{cases}$$

where q_{sj} is the probability that j streams are either in the fast search mode or are in the queue waiting to switch to a fast search mode under the condition that there are totally s streams present in the system and

$$q_{sj} = \frac{\rho_1^j \binom{s}{j} \frac{j!}{t!} t^{t-j}}{\sum_{k=0}^{t-1} \rho_1^k \binom{s}{k} + \sum_{k=t}^{s} \rho_1^k \binom{s}{k} \frac{k!}{t!} t^{t-k}},$$

where

$$\rho_1 = \frac{\lambda_1}{\mu_1}.$$

Because the number of streams in the VOD system does change from time to time, this paper uses the following inequality to figure out the number of streams that the disk array storing fast search files must be able to service at one time, denoted by n, in order to reduce the the average waiting time observed by a stream switching from the normal mode to a fast search mode to less than Q_f.

$$\sum_{s=n+1}^{m} p_s \left(\sum_{j=n}^{s} q_{sj} \frac{j-n+1}{n\mu_1} \right) \le Q_f,$$

where m is the maximum number of streams that the disk array storing the normal mode files is able to service at one time and p_s is the probability that s streams are in the system based on the Markov process model in Figure 4(a). According to the queueing theory,

$$p_s = \begin{cases} p_0 \dfrac{(m\rho_0)^s}{s!}, & \text{if } s < m, \\ p_0 \left(\dfrac{(m\rho_0)^m}{m!} \right) \left(\dfrac{1}{1-\rho_0} \right), & \text{if } s = m, \end{cases}$$

and

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho_0)^k}{k!} + \left(\frac{(m\rho_0)^m}{m!} \right) \left(\frac{1}{1-\rho_0} \right) \right]^{-1}$$

Once the number of streams that the disk array storing fast-mode files must provide is determined, denoted by n, n is multiplied by the weighted average bandwidth to figure out the amount of bandwidth required.

Obviously, applying the mechanism described above in determining the number of streams that the disk array storing the fast mode files must be able to support at one time is not in fully accordance with the condition on which the Markov process shown in Figure 4(b) is based. The Markov process shown in Figure 4(b) assumes the number of streams in the system remains a constant. Therefore, the validity of this mechanism is subject to verification. The verification of the queueing model presented in this section is presented in next section.

2.3 Verification of the queueing model

In this paper, the validity of the queueing model is verified through simulation. Figures 5-9 depict the simulation results based on different sets of parameters and compare them with the results derived from applying the queueing model shown in Figure 4. In all cases, it is assumed that the length of a video program is 90 minutes, which is about the length of a typical movie. Parts (a) of the figures show the parameters used in simulation runs. Figures 5 shows a case in which all the programs require the same amount of bandwidth for playing back. Figures 6 shows a case in which the bandwidth required to play back the programs is uniformly distributed. Figures 7-9 show three cases in which the bandwidth required to play back the programs is of the form defined as follows:

$$Probability[x = k] =$$

$$\begin{cases} P_{n(\mu,\sigma)}[X \le k + 0.5] & \text{if } k = 2 \\ P_{n(\mu,\sigma)}[X \le k + 0.5] - & \\ P_{n(\mu,\sigma)}[X \le k - 0.5] & \text{if } 3 \le k \le 7 \\ 1 - P_{n(\mu,\sigma)}[X \le k - 0.5] & \text{if } k = 8 \\ 0 & \text{otherwise,} \end{cases}$$

where $P_{n(\mu,\sigma)}$ is the probability function of the normal distribution with mean μ and variance σ^2. In Figures 7, 8,

and 9, the means of the normal distributions are all equal to 5 and the variances are 1, 2, and 3, respectively.

Parts (c) of the figures show how the average waiting time observed by a client changes with respect to the bandwidth of the disk array. Here, the bandwidth of the disk array is measured in units of the weighted average of the playback bandwidth of the video programs in the system. The simulation results presented in the figures are derived from averaging 100 independent simulation runs with the same set of parameters but different random number seeds. Parts (d) of the figures show the number of disk bandwidth units that the disk array storing the normal-mode files needs to provide in order to reduce the average waiting time for a new stream to less than 5 seconds, and parts (e) show the number of disk bandwidth units that the disk array storing the fast-mode files needs to provide in order to reduce the average waiting time of a stream switching to a fast mode to less than 1 second. As the figures reveal, in all 5 cases, the values determined by applying the proposed queueing model consistently fall within 5% range from the values derived from simulation runs.

3 Conclusions

This paper discusses guaranteeing quality of service in interactive video-on-demand servers. This paper proposes a queueing model for analyzing the I/O bandwidth required in order to achieve a certain level of QoS for a disk system design scheme that provides VCR emulation without requiring extra disk bandwidth to support interactive operations. The major advantages of the proposed disk system design scheme include:

1. providing VCR emulation with similar visualization effects;

2. requesting no extra bandwidth to support a stream switching from the normal playback mode to a fast search mode;

3. introducing little overhead in respect to disk space utilization.

The proposed queueing model provides a complete analysis tool for pinning down the resource requirement issue. Since the proposed queueing model employs several approximations, its validity must be verified. This paper uses simulations to check the accuracy of the queueing model. The simulation results show that the queueing model is quite accurate in determining the disk bandwidth required for achieving a certain level of QoS.

References

[1] Jayanta K. Dey-Sircar, James D. Salehi, James F. Kurose, and Don Towsley. "Providing VCR capabilities in large scale video server," In *Proceesings of the Second ACM International Conference on Multimedia*, pp. 25–32, October 1994.

[2] Ming-Syan Chen, Dilip D. Kandlur, and Philip S. Yu. "Support for fully interactive playout in a disk-array-based video server," *ACM Multimedia*, Vol. 3, No. 3, pp. 126–135, April 1995.

[3] Chih-Yuan Cheng, Chun-Hung Wen, Meng-Huang Lee, and Yen-Jen Oyang. "Effective utilization of disk bandwidth for supporting interactive video-on-demand," *IEEE Transaction on Consumer Electronics*, Vol. 42, No. 1, pp. 71–79, February 1996.

[4] Ming-Syan Chen and Dilip D. Kandlur. "Stream conversion to support interactive video playout," *IEEE Multimedia Magazine*, Vol. 3, No. 2, pp. 51–58, Summer 1996.

[5] Prashant J. Shenoy and Marrick M. Vin. "Efficient support for scan operations in video servers," In *Proceedings of the Third ACM Conference on Multimedia*, November 1995.

[6] "International standard iso/iec 11172-1,". The International Organization for Standardization and the International Electronical Commision, August 1993.

[7] "International standard iso/iec dis 13818-1,". The International Organization for Standardization and the International Electronical Commision, October 1995.

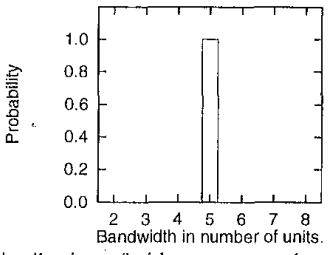
[8] Chih-Yuan Cheng, Meng-Huang Lee, and Yen-Jen Oyang. "Disk system design for guaranteeing quality of service in interactive video-on-demand systems," In *Proceedings of IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, January 1998.

[9] George Apostolopoulos, Marwan Krunz, and Satish Tripathi. "Supporting interactive scanning operations in vod systems," In *Proceedings of IS&T/SPIE Symposium on Electronic Imaging: Science and Technology, Conference on Multimedia Computing and Networking*, January 1998.

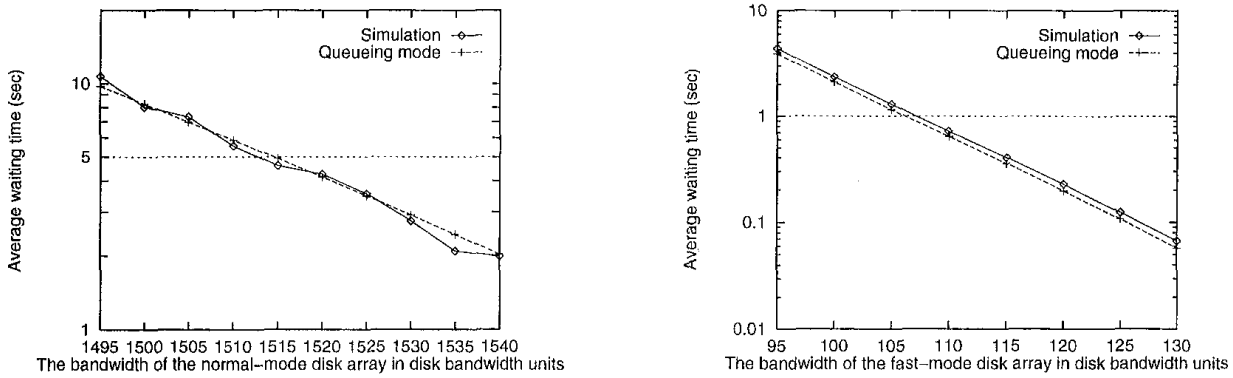
[10] Leonard Kleinrock. *Queueing Systems Vol.1*. A Wiley-Interscience Publication, 1975.

λ_1	μ_1	λ_2	μ_2
$1/20\ sec^{-1}$	$1/5400\ sec^{-1}$	$1/300\ sec^{-1}$	$1/20\ sec^{-1}$

(a) Simulation parameters.



(b) Distribution of video programs bandwidth



(c) Average waiting time v.s. disk array bandwidth

	By queueing model	By simulation
Required bandwidth in disk bandwidth units	1515	1515

(d) The bandwidth required by the normal-mode disk array for reducing the average waiting time of a newly-created stream to less than 5 seconds.

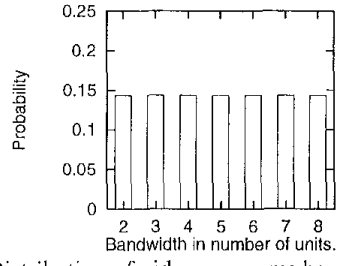
	By queueing model	By simulation
Required bandwidth in disk bandwidth units	110	110

(e) The bandwidth required by the fast-mode disk array for reducing the average waiting time of a stream switching from the normal mode to a fast search mode to less than 1 second.

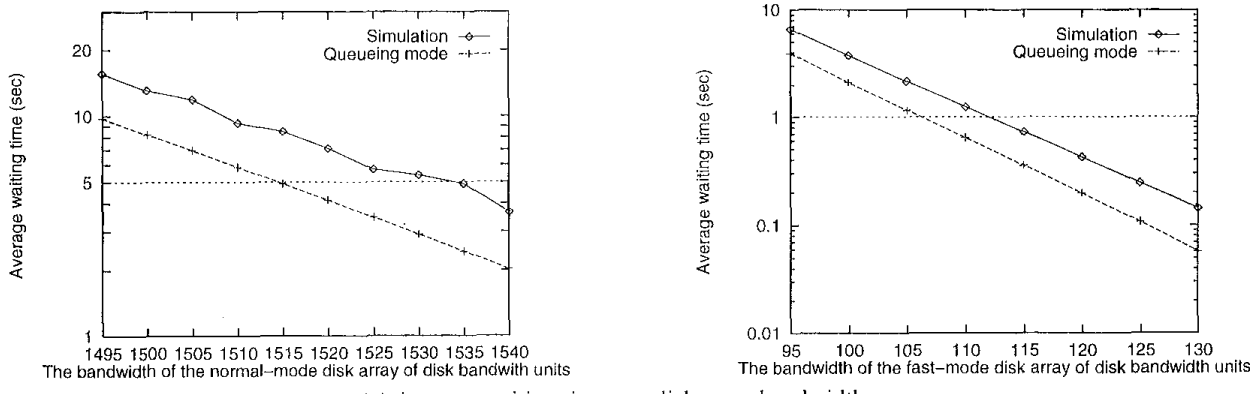
Figure 5: Simulation case 1

λ_1	μ_1	λ_2	μ_2
$1/20\ sec^{-1}$	$1/5400\ sec^{-1}$	$1/300\ sec^{-1}$	$1/20\ sec^{-1}$

(a) Simulation parameters.



(b) Distribution of video programs bandwidth.



(c) Average waiting time v.s. disk array bandwidth

	By queueing model	By simulation
Required bandwidth in disk bandwidth units	1535	1515

(d) The bandwidth required by the normal-mode disk array for reducing the average waiting time of a newly-created stream to less than 5 seconds.

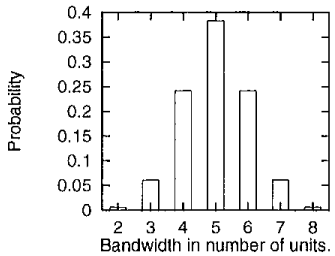
	By queueing model	By simulation
Required bandwidth in disk bandwidth units	115	110

(e) The bandwidth required by the fast-mode disk array for reducing the average waiting time of a stream switching to a fast search mode to less than 1 second.

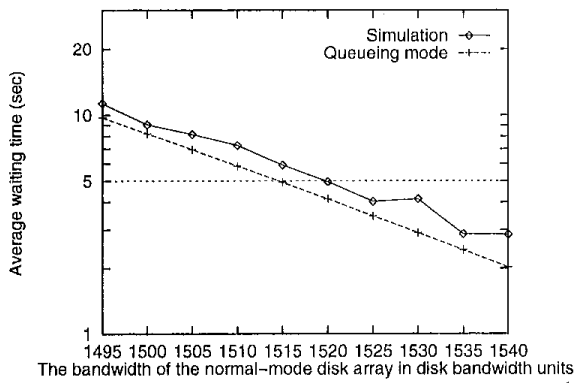
Figure 6: Simulation case 2

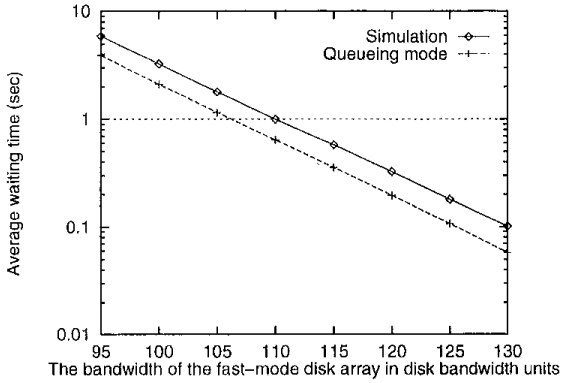
λ_1	μ_1	λ_2	μ_2
$1/20\ sec^{-1}$	$1/5400\ sec^{-1}$	$1/300\ sec^{-1}$	$1/20\ sec^{-1}$

(a) Simulation parameters



(b) Distribution od video programs bandwidth





(c) Average waiting time v.s. disk array bandwidth

	By queueing model	By simulation
Required bandwidth in disk bandwidth units	1520	1515

(d) The bandwidth required by the normal-mode disk array for reducing the average waiting time of a newly-created stream to less than 5 seconds.

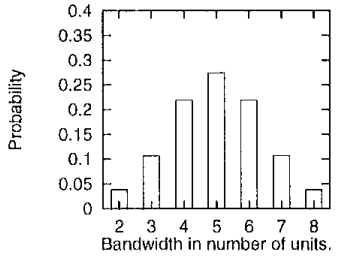
	By queueing model	By simulation
Required bandwidth in disk bandwidth units	110	110

(e) The bandwidth required by the fast-mode disk array for reducing the average waiting time of a stream switching to a fast search mode to less than 1 second.

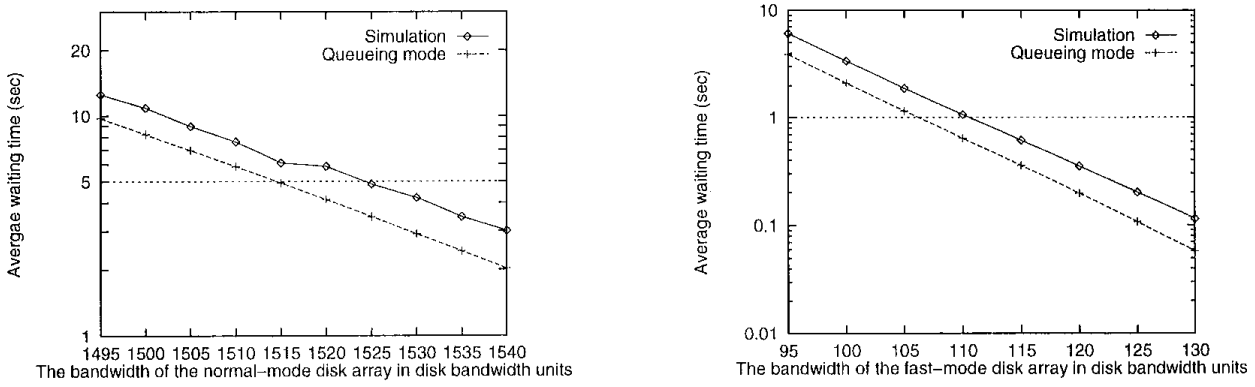
Figure 7: Simulation case 3

λ_1	μ_1	λ_2	μ_2
$1/20\ sec^{-1}$	$1/5400\ sec^{-1}$	$1/300\ sec^{-1}$	$1/20\ sec^{-1}$

(a) Simulation parameters



(b) Distribution of video programs bandwidth



(c) Average waiting time v.s. disk array bandwidth

	By queueing model	By simulation
Required bandwidth in disk bandwidth units	1525	1515

(d) The bandwidth required by the normal-mode disk array for reducing the average waiting time of a newly-created stream to less than 5 seconds.

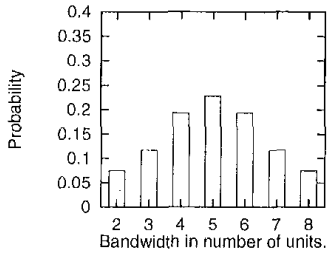
	By queueing model	By simulation
Required bandwidth in disk bandwidth units	115	110

(e) The bandwidth required by the fast-mode disk array for reducing the average waiting time of a stream switching to a fast search mode to less than 1 second.

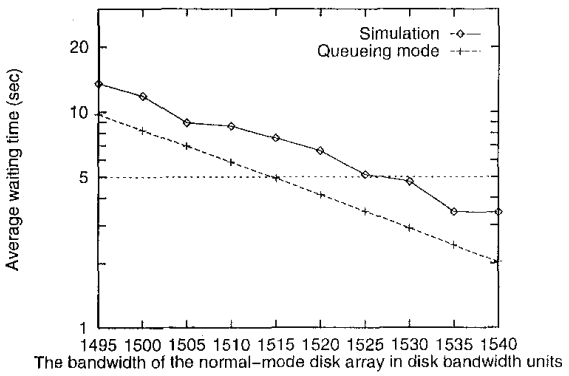
Figure 8: Simulation case 4

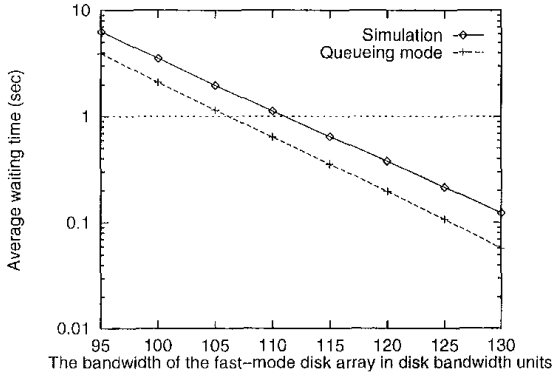
λ_1	μ_1	λ_2	μ_2
$1/20\ sec^{-1}$	$1/5400\ sec^{-1}$	$1/300\ sec^{-1}$	$1/20\ sec^{-1}$

(a) Simulation parameters



(b) Distribution of video programs bandwidth





(c) Average waiting time v.s. disk array bandwidth

	By queueing model	By simulation
Required bandwidth in disk bandwidth units	1530	1515

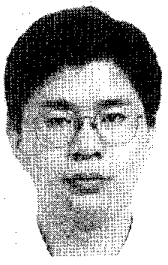
(d) The bandwidth required by the normal-mode disk array for reducing the average waiting time of a newly-created stream to less than 5 seconds.

	By queueing model	By simulation
Required bandwidth in of disk bandwidth units	115	110

(e) The bandwidth required by the fast-mode disk array for reducing the average waiting time of a stream switching to a fast search mode to less than 1 second.

Figure 9: Simulation case 5

Chih-Yuan Cheng received the B.S. degree in Computer Science and Information Engineering from National Taiwan University in 1991, the M.S. degree in 1993, and Ph.D. degree in 1998. He is currently a project manager in Bridgewell Inc. His research interests include multimedia storage systems, and operating systems.



Meng-Huang Lee received the B.S. degree in Electrical Engineering from National Cheng Kung University in 1987, the M.S. degree in 1989, and the Ph.D. degree in Computer Science and Information Engineering from National Taiwan University in 1996. He is currently an Associate Professor in the Department of Information Management, Shih Chien University. His research interests include multimedia storage systems and operating systems.

Yen-Jen Oyang received the B.S. degree in Information Engineering from National Taiwan University in 1982, the M.S. degree in Computer Science from the California Institute of Technology in 1984, and the Ph.D. degree in Electrical Engineering from Stanford University in 1988. He is currently a professor in the Department of Computer Science and Information Engineering, National Taiwan University. From 1989 to 1996, he was an associate professor in the same department. His research interests include design of video server systems and digital libraries. He can be reached at yjoyang@csie.ntu.edu.tw.