Slice-and-Patch – An Algorithm to Support VBR Video Streaming in a Multicast-based Video-on-Demand System

C. W. Kong
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
cwkong1@ie.cuhk.edu.hk

Jack Y. B. Lee
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
yblee@ie.cuhk.edu.hk

Abstract

In recent years, a number of sophisticated architectures have been proposed to provide VoD service using multicast transmissions. Compared to their unicast counterparts, these multicast VoD systems are highly scalable and can potentially serve millions of concurrent users. Nevertheless, these systems are designed for streaming constant-bit-rate (CBR) encoded videos and thus cannot benefit from the improved visual quality obtainable from variable-bit-rate (VBR) encoding techniques. To tackle this challenge, this paper presents a novel Slice-and-Patch (S&P) algorithm to support VBR video streaming in a multicast VoD system. Extensive trace-driven simulations are conducted to compare performance of the S&P algorithm with two other algorithms based on priority scheduling. Results show that the S&P algorithm outperforms the other two priority scheduling algorithms for most videos. Compared to the CBR counterpart serving videos of the same average bitrate, the S&P algorithm is able to support VBR video streaming with only 50% increase in latency. Given that VBR-encoded video can achieve visual quality comparable to CBR-encoded video at half the bitrate, this S&P algorithm can potentially achieve perform comparable to CBR-based systems when combined with VBR encoding techniques.

1. Introduction

In a true-video-on-demand (TVoD) system, the video server has to reserve a dedicated video channel for each user for the duration of the session (e.g. two hours for a movie). Consequently, the server and network resources required increase linearly with the number of concurrent users to be supported. Although current PC servers are already very powerful and capable to serve up to hundreds of concurrent video streams, scaling up a system to beyond the thousands and even millions of concurrent video streams is still prohibitively expensive.

One promising solution to this scalability challenge is through the intelligent use of network multicast. Network multicast enables a server to send a few streams of video data for reception by a large number of clients, thereby significantly reducing the amount of resources required. A number of pioneering studies have investigated such architectures, such as batching [1-3], patching [4-7], and periodic broadcasting [8-11].

A common assumption with these multicast VoD architectures is that the videos are constant-bit-rate (CBR) encoded. This significantly simplifies system design and analysis, and enables one to study the system performance independent from video encoding variations. Nevertheless, the visual quality of CBR video is not constant and tends to vary according to the video content. For example, complex video scenes with a lot of motions will typically result in lower visual quality than simple video scenes with little movement.

By contrast, videos encoded with constant-quality encoding algorithms will have consistent visual quality, at the expense of bitrate variations. A study by Tan *et al*. [12] had shown that VBR-encoded video can achieve visual quality similar to CBR-encoded video using only half the bitrate. This result suggests that VBR encoding is desirable for providing high-quality VoD services. The challenge is the complex resource allocation and scheduling problems resulting from the video bitrate variations.

This study addresses this challenge and presents a slice-and-patch (S&P) algorithm for allocating resources and scheduling video data transmissions in a multicast VoD system proposed by Lee and Lee [13]. The original multicast VoD system is designed for CBR videos, and combines techniques of batching, patching, and periodic broadcasting. This multicast VoD system can be scaled up

Proceedings of the Ninth International Conference on Parallel and Distributed Systems (ICPADS'02)
1521-9097/02 $17.00 © 2002 **IEEE**

to an unlimited number of concurrent users and thus is most suitable for serving popular movies in a metropolitan-scale VoD service. We give a brief overview of this VoD architecture in Section III and refer the interested readers to Lee and Lee [13] for details.

The S&P algorithm is designed with two principles. First, video data corresponding to video bitrate peaks are prefetched at startup. This step reduces the worst-case peak rate of the video stream, and thus allows more efficient resource allocations. Second, the video stream, minus the previously-mentioned peaks, is sliced into two sub-streams of lower bitrates and multicast periodically in two static multicast channels. A client, after prefetching the peaks, will initiate patching to begin playback using a dynamically allocated video channel while at the same time, caches video data from the static multicast channels. Eventually, video playback will reach the point where video data are already cached and the client can then release the dynamic channel and continue video playback using data received from the multicast channels.

We use simulations to study and compare the S&P algorithm with another two algorithms based on priority scheduling. Our simulation results show that the S&P algorithm outperforms the priority scheduling algorithm for most of the 50 tested videos. Compared to the CBR version of the system, the S&P algorithm can serve VBR videos of the same average bitrate with an average latency increase of only 50%. As VBR-encoded video requires only half the bitrate to achieve the same quality as CBR-encoded video, this S&P algorithm can potentially support VBR video streaming with resources comparable to CBR-based VoD systems.

The rest of the paper is organized as follows. Section II reviews some related works and compares them with this study; Section III reviews the multicast VoD architecture; Section IV presents the two priority scheduling algorithms; Section V presents the Slice-and-Patch algorithm; Section VI evaluates and compares the three algorithms using simulation results; and Section VII concludes the paper.

2. Background

The problem of VBR video delivery in unicast VoD systems has been studied extensively. We review some of the more relevant previous works in Section A and compare them with this study in Section B.

2.1. Previous Work

One of the most well-known solutions for VBR video delivery is temporal smoothing [14-17]. Smoothing makes use of a client-side buffer to receive data in advance of playback. This work-ahead technique enables the server to transmit video data in a piecewise linear schedule that can be optimized to minimize rate variability [15] or to minimize the number of rate changes [16]. The schedule can be computed offline and with proper resource reservation, deterministic performance can be guaranteed. Interested readers are referred to Feng *et al*. [17] for a thorough comparison of various smoothing algorithms.

In another study by Lee and Yeom [18], a data prefetch technique is proposed to improve video server performance in serving VBR videos. Unlike smoothing, where all video data are retrieved from the disk in sequence, data prefetching preloads video data corresponding to a video's bitrate peaks into the server's memory during system initialization. During operation, the server then only needs to retrieve the remaining video data from the disk to combine with the prefetched data for transmission to the clients. As the remaining video stream has a lower peak bitrate, disk utilization is increased. Their simulation results show that up to 81% more streams can be served using this prefetch technique. The tradeoffs are increased server buffer requirement and additional offline preprocessing of the video data.

A third approach proposed by Saparilla *et al.* [8], schedules video data transmission using a priority scheduler (the Join-the-Shortest Queue). In particular, the server schedules video data transmission according to the demand of data of each channel. A channel with the greatest demand of data (the clients listening to this channel is most likely to run out of data) will have the highest priority in the next round of transmission. However, while server efficiency is improved, this priority scheduler does not guarantee a client can receive all data in time. In particular, a channel will simply be skipped (i.e. not transmitted) if the data cannot be transmitted in time for playback. Their simulation results show that with their Join-the-Shortest Queue priority scheduling and allowing the client to retrieve data from seven channels synchronously, the start-up latency can be limited to around 100 seconds with a loss probability of 10^{-6}.

2.2. Comparison

Compared to the S&P algorithm investigated in this study, both temporal smoothing and the data prefetch techniques discussed previously are orthogonal and complementary. For temporal smoothing, a smoothed VBR video stream can be considered as just another VBR video stream, albeit one requiring additional client buffer for proper playback. For the data prefetch technique, the focus is on improving disk retrieval efficiency by intelligently preloading some video data into the server memory. Obviously, this technique does not affect the transmission schedule at all and thus can be integrated with any transmission scheduling algorithms including S&P.

Compares to the study by Saparilla *et al.* [8], S&P differs in two major ways. First, the S&P algorithm guarantees that no video data will be skipped, thus

ensuring visual quality. Second, S&P is targeted to clients with limited access bandwidth (twice the average bit rate of the video). By contrast, the algorithm proposed by Saparilla *et al*. assumes the client to have sufficient bandwidth to receive data from many channels simultaneously, which currently may not be practical.

This study is a first step in exploring algorithms for supporting VBR video delivery in multicast VoD systems. Designing the S&P algorithm reveals many difficulties and challenges that are not present in conventional unicast VoD systems. Nevertheless, this will be an important area as VBR encoding is necessary to provide good visual quality, plus multicast VoD systems may be the only way to deploy cost-effective metropolitan-scale VoD services in the near future.

3. System Architecture

In this section, we give a brief overview of the multicast VoD architecture – super-scalar VoD (SS-VoD), investigated in this paper. The system comprises a number of service nodes delivering video data over multicast channels to the clients. SS-VoD achieves scalability and bandwidth efficiency by sending video data to a large number of clients using a few multicast channels. However, simple periodic multicast schemes such as those used in a near-video-on-demand (NVoD) system limit the time for which a client may start a new video session. Depending on the number of multicast channels allocated for a video title, this startup delay can range from a few minutes to tens of minutes. To tackle this initial delay problem, SS-VoD employ patching to enable a client to start video playback at any time using a dynamic multicast channel until it can be merged back onto an existing multicast channel. The following sections present these techniques in more detail.

3.1. Transmission Scheduling

Each service node in the system streams video data into multiple multicast channels. Let M be the number of video titles served by each service node and let N be the total number of multicast channels available to a service node. For simplicity, we assume N is divisible by M and hence each video title is served by the same number of multicast channels, denoted by $N_M=N/M$. These multicast channels are then divided into two groups of N_S static multicast channels and $N_D=N_M-N_S$ dynamic multicast channels.

The video title is multicast repeatedly over all N_S static multicast channels in a time-staggered manner as shown in Fig. 1. Specifically, adjacent channels are offset by

$$T_R = L / N_S \qquad (1)$$

seconds, where L is the length of the video in seconds. Transmissions are repeated continuously, i.e. restarted from the beginning of a video title every time

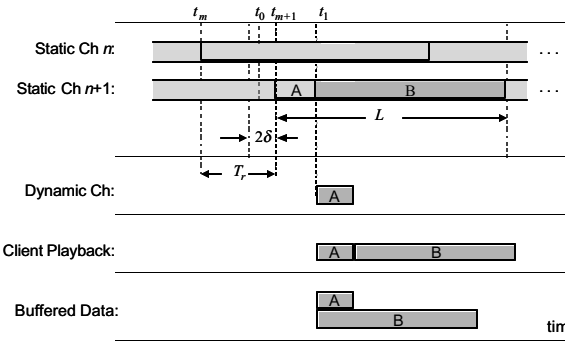

Figure 1. The patching process in the super-scalar video-on-demand system supporting CBR video.

transmission completes, regardless of the load of the server or how many users are active. These static multicast channels are used as the main channels for delivering video data to the clients. A client may start out with a dynamic multicast channel but it will shortly be merged back to one of these static multicast channels as explained in the next section.

3.2. Admission Control

To reduce the response time while still leveraging the bandwidth efficiency of multicast, SS-VoD allocates a portion of the multicast channels and schedules them dynamically according to the request arrival pattern. A new user either waits for the next upcoming multicast transmission from a static multicast channel, or starts playback with a dynamic multicast channel.

Suppose a new request arrives at time t_0, which is between the start time of the previous multicast cycle, denoted by t_m, and the start time of the next multicast cycle, denoted by t_{m+1} (see Fig. 1). The new request will be assigned to wait for the next multicast cycle to start playback if the waiting time, denoted by w_i, is equal to or smaller than a predefined admission threshold 2δ, i.e., $w_i = t_{m+1} - t_0 \le 2\delta$. We call these requests *statically-admitted*. This admission threshold is introduced to reduce the amount of load going to the dynamic multicast channels.

On the other hand, if the waiting time is longer than the threshold, then the client will request a dynamic multicast channel to begin playback (*dynamically admitted*), while at the same time caches video data from the multicast channel with the multicast cycle started at time t_m. Note that the client may need to queue up and wait for a dynamic multicast channel to become available. If additional clients requesting the same video arrives during the wait, they will be batched together and served by the same dynamic multicast channel once it becomes available. Eventually, the client playback will reach the

point where the cached data began and the client can then release the dynamic multicast channel and continue playback using data received from the static multicast channel. This integration of batching with patching significantly increases the system's efficiency at heavy loads.

Compared to TVoD systems, a SS-VoD client must have the capability to receive two multicast channels concurrently and have a local buffer large enough to hold up to T_R seconds of video data. Given a video bitrate of 3Mbps (e.g. high-quality MPEG-4 video), a total of 6Mbps downstream bandwidth is required during the initial patching phase of the video session. For a two-hour movie served using 25 static multicast channels, the buffer requirement is 108MB. This can easily be accommodated today using a small harddisk in the client, and in the near future simply using memory as technology improves.

3.3. Challenges in Supporting VBR-encoded Video

The SS-VoD architecture is originally designed for CBR videos. A problem arises if we want to support VBR videos. Specifically, although a client has the capability to receive twice the video bitrate, it may not be sufficient to support two channels of VBR video of the same average bitrate due to bitrate variations. The use of temporal smoothing can alleviate this problem but cannot solve it completely without adding excessive start-up delay. We investigate in the next section two possible solutions to this problem based on priority scheduling.

4. Priority Scheduling

The primary problem with supporting VBR video in SS-VoD is that dynamically-admitted clients may not have sufficient access bandwidth to accommodate both the dynamic and a static multicast channel. For example, let R_V be the average video bitrate, then the client has an access bandwidth of $2R_V$. However, a VBR video of average rate R_V will likely have bitrate peaks (valleys) higher (lower) than R_V even after smoothing is applied. It is easy to see that the access channel will become congested whenever peaks from both dynamic channel and static channel overlap.

Assuming client access bandwidth is limited, then we will need to prioritize the transmission and reception of video data to stay within the given access bandwidth. The following sections present two such priority-scheduling algorithms.

4.1. Static Channel Priority

In the static channel priority algorithm, we let the static channels transmit at the original video bitrate and adjust the transmission rate of the dynamic channel to fit within the access bandwidth limit. Let $v(t)$ be the video data consumption rate function that defines the rate at which video data are being consumed t seconds after playback has begun. Assume the client arrives at time t_0, and the immediate previous multicast cycle begins at time t_m, then the client will be caching video data starting from a playback point of $t_c=t_0-t_m$ and the amount of access bandwidth left for the dynamic channel at time t is equal to $u(t) = 2R_V - v(t-t_m)$ for $t \geq t_0$.

As the client does not yet have any video data of playback point earlier than t_c, a dynamic channel will be allocated to begin streaming data from the beginning of the video to the playback point t_c. If the bandwidth available to the dynamic channel is sufficient for streaming the video, i.e., $u(t) \geq v(t-t_0-w)$ for $(t_0+w) \leq t \leq (t_0+w+t_c)$:

$$\int_{t_0+w}^{\tau} u(t)dt \geq \int_{t_0+w}^{\tau} v(t-t_0-w)\,dt, \quad \text{for } (t_0+w) \leq \tau (t_0+w+t_c) \tag{2}$$

where w is time waiting for the dynamic channel, then no further action needs to be done. Otherwise, the client cannot begin playback immediately when data are received because playback continuity cannot be sustained when the condition in (1) fails.

To tackle this problem, the client will have to delay the playback by t_s seconds so that the continuity condition is satisfied:

$$\int_{t_0+w}^{\tau} u(t)dt \geq \int_{t_0+w+t_s}^{\tau} v(t-t_0-w-t_s)\,dt, \quad \text{for } (t_0+w+t_s) \leq \tau (t_0+w+t_c) \tag{3}$$

and this is also the tradeoff for this algorithm.

4.2. Dynamic Channel Priority

To avoid the startup delay in the previous static channel priority algorithm, we can give priority to the dynamic channel during admission. Unlike the previous algorithm, we cannot simply transmit video data of the static channel using the left-over access bandwidth because the static channels are periodically multicast in a fixed schedule to a large number clients. Therefore, once a dynamic channel becomes available, the server will transmit video data from the beginning of the video at the maximum rate $2R_V$ until it catches up with the playback point, say s, currently being multicast by the static channel. At that instant, the client can then releases the dynamic channel and continue receiving data from the static channel for the rest of the session.

Similarly, we again invoke the playback continuity condition to find the value of s that satisfies the following condition:

$$2R_V(s-t_0-w) = \int_0^s v(t)\,dt, \quad \text{where } s \geq (t_0+w) \tag{4}$$

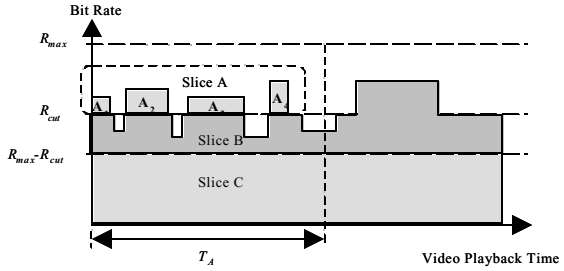

Figure 2. Video slicing in the Slice-and-Patch algorithm.

This algorithm does not incur start-up delay but a dynamic channel will consume more resource than the same in the static priority algorithm for two reasons. First, the dynamic channel itself will be streamed at the maximum access bitrate. Second, the client cannot cache video data from the static channel while the dynamic channel is streaming. This increases the time the dynamic channel takes to catch up with the static channel. Both factors increase a dynamic channel's bandwidth consumption and will result in a longer time to wait for an available dynamic channel.

5. Slice-and-Patch

The two algorithms presented in the previous section both have their tradeoffs. In this section, we present a slice-and-patch (S&P) algorithm that combines the virtues of the static channel priority and the dynamic channel priority algorithms. In S&P, we divide the video stream into three portions (i.e. slicing) and admit clients using a three-phase patching process (i.e. patching). The following sections present the algorithm in detail.

5.1. Video Slicing

Video slicing is an offline process that divides a video data stream into three parts (i.e. slices) for transmission in three separate multicast channels. As shown in Fig. 2, the video data stream is sliced at two bitrates: R_{cut} and R_{max}–R_{cut}. The parameter R_{cut} is configurable from R_V to $(2/3)R_{max}$ and can be optimized for a particular video.

To generate the first part – Slice A, we collect all video data exceeding the bitrate R_{cut} (e.g. A_1, A_2, etc., in Fig. 2), starting from the beginning of the video until the playback point T_A given by

$$T_A = \left(\frac{R_{cut}}{R_{\max} - R_{cut}} \right) T_R \qquad (5)$$

where R_{max} is the maximum access bandwidth of the client, and T_R is the repeating interval for the static multicast channels. We will derive T_A in Section C when we explain the three-phase patching process. The purpose of this slicing is to reduce the peak rate of the video stream to prevent congesting the client's access channel

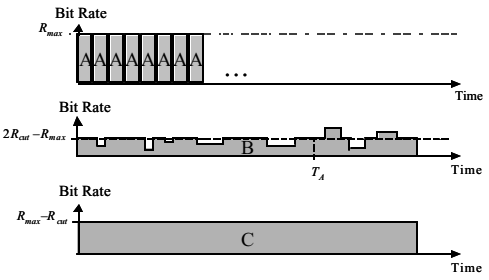

Figure 3. The three types of multicast channels in the Slice-and-Patch algorithm.

during patching. The resultant slice will be repeatedly multicast over a dedicated channel at a constant bitrate R_{max} as shown in Fig. 3. Assume the size of the block is A Mb, then the slice will be multicast repeatedly once every $t_{da}=A/R_{max}$ seconds.

The second part – Slice B in Fig. 2, comprises two portions. The first portion, covering the first T_A seconds of the video, comprises the *remaining* video data that exceeds the bitrate (R_{max}–R_{cut}). The second portion, covering from playback point T_A until the end of the video, comprises *all* video data that exceeds the bitrate (R_{max}–R_{cut}). This slice will be multicast repeatedly over a separate multicast channel following the actual video data rate (as opposed to the constant transmitting rate for Slice A).

Lastly, the third part, – Slice C in Fig. 2, comprises the rest of the video data. This slice will be multicast repeatedly over a separate multicast channel following the actual video data rate, but no higher than (R_{max}–R_{cut}).

5.2. Bandwidth Allocation

Let B_{max} be the total server (or network, whichever is smaller) bandwidth available for a video of average bitrate R_V Bps and length L seconds. First, a bandwidth of R_{max} will be allocated for multicasting Slice A. Then the remaining bandwidth will be equally divided between dynamic multicast channels and static multicast channels. Simulation results have shown that this equal allocation always results in the best performance.

There are two types of static multicast channels, one type transmitting Slice B and the other transmitting Slice C. As the number of these channels are equal, we will refer a pair of such channels as a static multicast channel. Unlike the case of CBR videos, a static multicast channel in S&P does not occupy a fixed bandwidth. Therefore offline numerical procedures are needed to compute the maximum number of static multicast channels that can fit within the bandwidth (B_{max}–R_{max})/2. The remaining bandwidth will be used by the dynamic channels.

5.3. Three-Phase Patching

A new client will go through a three-phase patching process to begin a new video streaming session. Let the

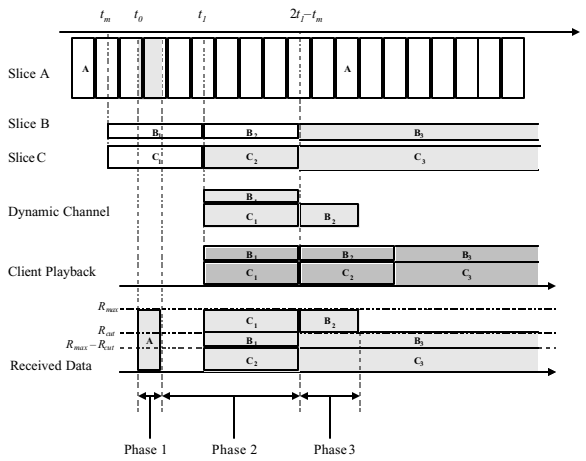

Figure 4. The three-phase patching process in the Slice-and-Patch algorithm.

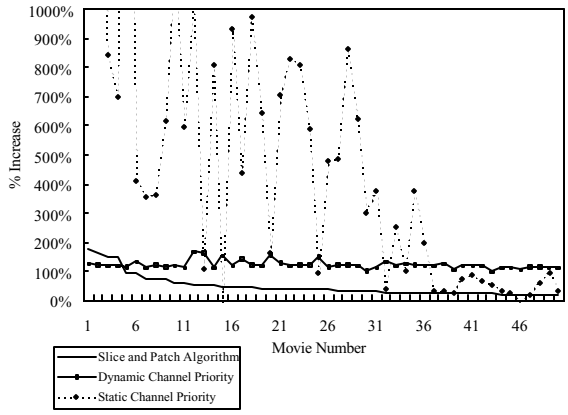

Figure 5. Percentage of latency increases over CBR-based system for 50 different videos.

client arrives at time t_0. It immediately enters Phase 1 by caching Slice A at the maximum rate R_{max} for a duration of t_{da} seconds (see Fig. 4). Next, the client will request a dynamic channel to begin Phase 2. Once a dynamic channel becomes available at time t_1, the client begins receiving and playing back video data blocks $\{B_1, C_1\}$ while simultaneously caching block C_2 into a local buffer. Note that the length of the blocks $\{B_1, C_1\}$ is equal to t_1-t_m seconds and this is also the duration of Phase 2.

At the beginning of Phase 3, the client will have cached block C_2 and completed playback of blocks $\{B_1, C_1\}$. As Fig. 4 shows, to continue playback the client will need block B_2. This is supplied by the dynamic channel at the bitrate $R_{max}-R_{cut}$, as the static channel transmitting blocks $\{B_3, C_3\}$ occupies the remaining bandwidth of R_{cut}.

Since the size of block B_2 is equal to $(2R_{cut}-R_{max})(t_1-t_m)$, the time it takes to transmit this block at the rate of $R_{max}-R_{cut}$ will be equal to $(2R_{cut}-R_{max})(t_1-t_m)/(R_{max}-R_{cut})$. To derive the duration of Phase 3, we note that (t_1-t_m) must be smaller than T_R and thus the duration will be bounded by $(2R_{cut}-R_{max})T_R/(R_{max}-R_{cut})$, which equals to T_A in (5) after simplification. After Phase 3 is completed, the client releases the dynamic channel and continues playback using data received from the static channel for the rest of the video session.

6. Performance Evaluation

In this section, we evaluate the three algorithms presented in Section IV and Section V using simulation. The simulator is developed using the CNCL simulation library [19] and the simulations are conducted using 50 VBR video bitrate traces measured from DVD videos. To facilitate comparison, we scaled the video bitrate traces so that all videos have the same average bitrate of 3Mbps. The server is configured with a total bandwidth of 150Mbps and the client an access bandwidth of 6Mbps. Each simulation run simulates a duration of 30 days, with the first day of data skipped to reduce initial condition effects. The client arrival rate is 1 request per second.

For the S&P algorithm, we set the parameter R_{cut} to equal to $1.1R_t$, where R_t is the average bitrate of the first $2T_R$ seconds of the video. Clearly this may not be optimal and we are investigating an efficient way to find the optimal R_{cut} value, without running a huge number of simulation runs.

Fig. 5 plots the simulation results for the three algorithms for 50 different videos. The vertical axis is the percentage increase in latency compared to the same system serving CBR video of the same length and average bitrate (i.e. 3Mbps). Thus this shows the cost of supporting VBR-encoded video instead of CBR-encoded video, although the VBR-encoded video will have better visual quality [12].

There are several observations from the simulation results. First, in terms of average latency increase computed from all 50 videos, S&P performs best at 50%, Dynamic Channel Priority second at 120%, and Static Channel Priority worst at 550%. Second, in terms of variations in latency increases, Dynamic Channel Priority is best with consistent latency increases across all 50 videos (the standard deviation is only 14%). S&P has more variations at a standard deviation of 40%. Static Channel Priority is the worst one with a huge standard deviation of 865%, and a maximum latency increase over 2,000%. The higher variation in Static Channel Priority is due to variations in the bit-rate of the video's initial portion. In particular, the algorithm allocates more bandwidth to cache from the static channel video data that cannot be immediately played back. Therefore if the initial video portion has a high bit-rate, then the dynamic

channel will take a longer time to cache sufficient video data to begin playback. By contrast, the Dynamic Channel Priority and the S&P algorithms are less sensitive to this effect because both algorithms allocate more bandwidth to cache video data that can be played back immediately.

Comparing the three algorithms, the S&P algorithm clearly performs best except for a few videos. Although on average the latency is still increased by 50%, this did not account for the improved visual quality due to the use of VBR encoding techniques. Given that VBR-encoded video could achieve visual quality comparable to CBR-encoded video at twice the video bitrate, this S&P algorithm can potentially achieve performance comparable to CBR-based systems with the proper choice of VBR encoding parameters.

The simulation results are obtained by simulating each video individually. In a real system with multiple videos, one can further improve system efficiency by allocating channels according to the video's popularity. Further investigations are needed to quantify the potential performance gains and also tradeoffs in applying non-uniform channel allocation policies.

7. Conclusions

This paper investigated a multicast VoD system supporting streaming of VBR-encoded video. Unlike unicast-based VoD system, existing algorithms such as temporal smoothing do not cater for the characteristics of multicast VoD systems, such as the use of periodic multicast and the limitation of client access bandwidth. Therefore new streaming algorithms are needed, and three such algorithms, namely Static Channel Priority, Dynamic Channel Priority, and Slice-and-Patch are studied in this paper. Using simulation, we found that the Slice-and-Patch algorithm generally outperforms the other two except for a few videos. We suspect that the exceptions are due to the non-optimal configuration of the R_{cut} parameter in S&P. We are currently running additional simulations to further investigate this issue. Nevertheless, with a modest 50% latency increase over the case with CBR-encoded video, this S&P algorithm has a very good potential to achieve performance comparable to CBR systems with similar resources.

Acknowledgements

This research is funded by a Direct Grant, and Earmarked Grants (CUHK 4328/02E) from the HKSAR Research Grant Council and the AoE-IT, a research grant from the HKSAR University Grants Council.

References

[1] A. Dan, D. Sitaram, and P. Shahabuddin, "Scheduling Policies for an On-Demand Video Server with Batching," *Proc. 2nd ACM Multimedia*, 1994, pp.15-23.

[2] H. Shachnai and P.S. Yu, "Exploring Waiting Tolerance in Effective Batching for Video-on-Demand Scheduling," *Proc. 8th Israeli Conference on Computer Systems and Software Engineering*, Jun 1997, pp.67-76.

[3] V.O.K. Li, W. Liao, X. Qui, and E.W.M. Wong, "Performance Model of Interactive Video-on-Demand Systems," *IEEE JSAC*, vol.14(6), Aug 1996, pp.1099-1109.

[4] W. Liao and V.O.K. Li, "The Split and Merge Protocol for Interactive Video-on-demand," *IEEE Multimedia*, vol.4(4), 1997, pp.51-62.

[5] K.A. Hua, Y. Cai, and S. Sheu, "Patching: A Multicast Technique For True Video-on-Demand Services," *Proc. 6th International Conf on Multimedia*, Sept 1998, pp.191-200.

[6] Y. Cai, K. Hua, and K. Vu, "Optimizing Patching Performance," *Proc. SPIE/ACM Conference on Multimedia Computing and Networking*, CA, Jan. 1999, pp.204-215.

[7] S.W. Carter, D.D.E. Long, K. Makki, L. M. Ni, M. Singhal, and N. Pissinou, "Improving Video-on-Demand Server Efficiency Through Stream Tapping," *Proc. 6th International Conference on Computer Communications and Networks*, Las Vegas, Sep 1997, pp.200-207.

[8] D. Saparilla, K.W. Ross, and M. Reisslein, "Periodic Broadcasting with VBR-Encoded Video" *Proc. IEEE Infocom 1999*, New York City, US, March 1999.

[9] S. Sen, Gao Lixin, and D. Towsley, "Frame-based Periodic Broadcast and Fundamental Resource Tradeoffs," *IEEE International Conference on Performance, Computing, and Communications*, 2001, pp.77-83.

[10] T.C. Chiueh and C.H. Lu, "A Periodic Broadcasting Approach to Video-on-demand Service," *Proc. of SPIE*, Philadelphia, 1996, pp.2615:162-9.

[11] A. Hu, I. Nikolaidis, and P. van Beek, "On the Design of Efficient Video-on-Demand Broadcast Schedules," *Proc. 7th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, Maryland, 1999, pp.262-269.

[12] W.S. Tan, N. Duong, and J. Princen, "A Comparison Study of Variable-bit-rate versus Fixed-bit-rate Video Transmission," *Proc. Australian Broadband Switching and Services Symposium*, Australia, 1991, pp.134-141.

[13] J.Y. B. Lee and C.H. Lee, "Design, Performance Analysis, and Implementation of a Super-Scalar Video-on-Demand System," *submitted for publication*.

[14] W. Feng and S. Sechrest, "Smoothing and Buffering for the Delivery of Pre-recorded Video," *Proc. ISET/SPIE Multimedia Computing and Networking*, San Jose, Feb 1995, pp.234-244.

[15] J.D. Salehi, Z.L. Zhang, J.F. Kuros and D. Towsley, "Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements through Optimal Smoothing", *Proc. of ACM SIGMETERICS*, Philadelphia, May 1996, pp.222-231.

[16] W. Feng, F. Jahanian, and S. Sechrest, "Optimal Buffering for the Delivery of Compressed Pre-recorded Video," *Proc. of the IASTED/ISMM Int'l Conf. on Networks*, Jan. 1995.

[17] W. Feng, Mishra, and Ramakishnan, "A Comparison of Bandwidth Smoothing Techniques for the Transmission of Pre-recorded Compressed Video," *Proc. INFOCOM '97*, vol. 1, Japan, 1997, pp.58-66.

[18] D.Y. Lee and H.Y. Yeom, "Tip Prefetching: Dealing with the Bit Rate Variability of Video Streams" *Proc. of the IEEE ICMCS 1999*, vol. II, Italy, 1999, pp.352-356.

[19] ComNets Class Library and Tools: http://www.comnets.rwth-aachen.de/doc/cncl.html