

Scheduling Video Stream Transmissions for Distributed Playback over Mobile Cellular Networks

Kam-Yiu Lam¹, Joe Yuen¹, Sang H. Son² and Edward Chan¹

Department of Computer Science¹
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{cskylam|csjyuen|csedchan}@cityu.edu.hk

Department of Computer Science²
University of Virginia
Charlottesville, VA22903
son@cs.virginia.edu

Abstract

In this paper, we present the Buffer Sensitive Rate-Based (BSRB) algorithm, which provides a fair scheduling scheme to serve video playback requests, while maximizing the performance of individual playback. In BSRB, the amount of bandwidth allocated to serve a video request depends on the buffer level of the requesting client, and the expected and minimum bandwidth requirements of the video. By maintaining a high video buffer level at a client, the quality of the playbacks can be maintained and its performance will be more adaptive to the changing playback conditions. By employing the rate-based policy in BSRB, the performance of the requests is less affected by the poor performance of a single client and hence fair services can be provided to the clients.

1. Introduction

In a distributed mobile video player system, mobile clients are connected to a video server through a mobile cellular network. They may request videos with different workload characteristics from the video server for playback while they are moving. If a mobile client enters into another cell while the video is being played back, a handoff procedure must be performed. Since the bandwidth of a mobile network is very limited, efficient allocation of the bandwidth to serve the requests from multiple mobile clients is an important design issue on the playback quality of the individual video as well as the overall system performance.

The playback status of a video at a client can be highly dynamic due to the mobility of clients and changing network qualities. To provide a high QoS in video playback, the system has to be adaptive in service to the changing playback status of individual video request [3, 5]. Even with an admission controller, the system may still be subjected to transient overloading due to variation in video traffics, communication errors and changing workloads in a cell as a result of the

movement of clients. A mobile client may enter into another cell after a handoff procedure while its video is being played. If it fails in the admission, its video being played has to be dropped.

In [1], a rate-based (RB) method has been proposed to serve the video requests based on the expected and minimum requirements of the requests. In RB, the service is divided into levels and all the clients within the same cell receive the same level of services no matter what their expected workloads are. Resource reservation is used to minimize the probability of request drops from handoff clients. Although RB is fair and has several nice features, it is not flexible in resource allocation and cannot adapt to the changing playback status of individual request. In this paper, we present a new method, called *Buffer Sensitive Rate-Based (BSRB)*, for scheduling the limited mobile bandwidth in a cell to transmit video streams to serve the requests from multiple mobile clients by integrating the ideas of fairness and adaptability. In BSRB, the amount of bandwidth allocated to serve a video request depends on the buffer level of the requesting client, and the expected and minimum bandwidth requirements of the video. By maintaining a high video buffer level at a client, the playback quality can be maintained and its performance will be more adaptive to the changing playback conditions of the system. In addition, by employing the rate-based policy in BSRB, the performance of the playbacks will be less affected by the performance of an unfortunate client, which has a high probability of error in communication.

2. Related Work

In the design of scheduling methods for multimedia systems, one of the most important concerns is to meet the urgency of each video packet since missing the playback deadline will make it useless. Therefore, many priority-based real-time scheduling algorithms, e.g., earliest deadline first and rate monotonic, have

been extended for scheduling the transmission of video packets in distributed multimedia systems. Other important concerns, specifically for mobile multimedia systems, are fairness in services, workload distribution problem due to mobility of mobile clients, and error problems in video packet transmission. In [7], the earliest deadline first scheduling algorithm is extended for scheduling of video packets in a mobile environment. In [6], the Server Based Fairness (SBFA) approach is proposed in which it uses a long-term fairness server to save the bandwidth of bad channels. Channel-Condition independent fair queuing (CIF-Q) [4] is another fair queuing approach for error posed mobile networks. It creates an error-free system on each system as a reference and can use any well-known algorithm as the fair queue algorithm. Session selection is based on the virtual time of each system in the ideal error-free system.

In [1], a rate-based (RB) method is proposed based on the concept of service levels in servicing the client requests in a cell aiming to resolve the admission problem from handoff clients. It is assumed that each client request specifies the minimum bandwidth requirement in addition to the mean bandwidth requirement of the requesting video. When overloading occurs due to admission, all the existing client requests in the system will be affected in the same proportion in order to minimize the probability of admission blocking. In [8], the buffer sensitive bandwidth allocation method is proposed by considering the buffer levels at the clients in bandwidth allocation in order to make the playback of the video more tolerate to the changing workload situation in the system.

3. System Model and Assumptions

We consider a distributed mobile video player system on a cellular network as shown in Figure 1. In each cell, there is a base station for communicating with the mobile clients in its cell through a mobile network and the base station connects to a video server through a reliable high-speed network. Since the mobile network bandwidth is much smaller than that of the high-speed network, it is assumed to be the bottleneck resource in the system.

The video server maintains a collection of videos compressed using the MPEG II standard. Each encoded video stream consists of a sequence of *group of picture (GOP)*. Due to the different sizes of the GOPs and fixed play rate of a video, the bandwidth requirement of a video stream is variable. The number of mobile clients in a cell is not fixed since the mobile clients may move into another cell or move in from other cells. The mobile clients generate video playback requests to the video server through the base stations of their current

cells. Some of the mobile clients are thin clients, e.g., the PDA and handheld PC, while others are more powerful clients, e.g., notebook computers. Due to the great differences in mobile machine capability, the sizes of the buffers at the mobile machines for video frames may be very different for different mobile clients. In addition, different clients may request videos with different workload characteristics. For example, the workload of a video requested from a high performance PC client is usually much higher than that from a pocket PC client.

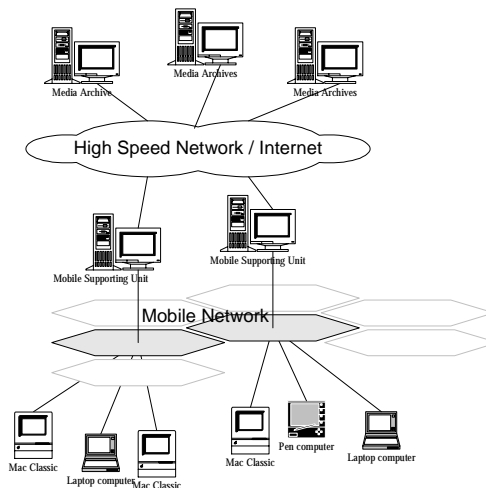


Figure 1: A distributed mobile video player system

The mobile client places the received frames at its video buffer. When the video buffer level has reached a pre-defined value, the decompression procedure starts and the playback of the video will begin by forwarding the decompressed frames to the video player one by one. If the playtime of a video frame has been missed, the frame will be dropped immediately.

When the server receives a video request, it will perform an admission test. If the request passes the test, the server will start to serve it. Upon the admission of a video request, the video server retrieves the required video file from the permanent storage and put the video frames into the video buffer specifically assigned to the request. The video frames are packed into video packets and then transmitted to the requesting mobile client through the base stations and the mobile network. While sending a video packet to the requesting client, communication errors may occur and retransmission may be needed. The communication error probability might be different for different clients since it normally depends on the location of the client and its surrounding environment.

4. Performance Objectives and Fairness in Services

There are various ways to quantify the quality of a video playback. The simplest one is based on the number of frames played (or number of video frames dropped). Each video request may associate with a minimum performance requirement, which may be defined on the percentages of played frames. In order to provide a guarantee to meet the minimum quality of a video request, an admission control procedure is usually employed to limit the total workload in the system and to prevent it from overloading.

In the design of admission control scheme and method for resources allocation, an important concern other than minimizing the number of frames dropped, especially in distributed mobile video player systems, is how to serve concurrent video requests in a fair manner. Since the mobile network is most likely to be the bottleneck resource, the problem is how to divide the limited mobile bandwidth between the base station and the mobile clients within the cell, to transmit video frames to the requesting clients. In this section, we first discuss the admission control issue and then examine the fairness issues in mobile cellular networks.

4.1. Admission Control and Transient Overloading

An admission controller decides whether a video request can be accepted or has to be rejected based on the existing video workload in the system and the workload characteristics of the requesting video. A simple way to define the admission condition is to use the average workload of the videos:

$$\frac{\sum_{i=1}^n \overline{BW_Consume_i}}{BW_Total} \leq 1 \quad (1)$$

where $\overline{BW_Consume_i}$ is the average bandwidth requirement of video stream i . BW_Total is the current total bandwidth available of the mobile network. For a cellular network, BW_Total is the current total bandwidth available at a base station to communicate with the mobile clients in its cell. The admission procedure needs to be applied on the newly created video requests from the mobile clients in the cell and also on the video admission request from the mobile clients, which are moving into the cell from neighboring cells.

With Eqn. 1, each client receives sufficient video frames for playback on average over a sufficient long period of time. However, transient overloading may still occur due to variable bandwidth requirements of videos. Transient overloading may also occur due to errors in communication, which makes the effective

bandwidth for video frames transmission much smaller than the allocated bandwidth. In order to minimize the probability of transient overloading, a tighter condition for admission control may be used. However, it will result in a higher probability of rejection or *request block probability* in admission, and a higher *request drop probability* for the requests from handoff clients.

4.2. Fairness in Services

The definition of *fairness* in a distributed mobile video player system is not trivial due to the dynamic properties and mixed workloads of the system [2]. Basically, we can divide the fairness issue into two parts: *fairness in admission* and *fairness in services*. In this paper, we concentrate on the fairness in services. If the system decides to admit a request, it is important to guarantee of the minimum quality to the client and to serve all the admitted ones fairly. The amount of bandwidth allocated to serve an admitted request should be proportional to the expected services specified by the clients even though its workload requirement may be much higher than other existing requests. When overloading situation occurs, the performance of all the existing requests should be affected in the same degree. If it is measured in terms of bandwidth allocation, the impact of overloading should affect the amount of bandwidth allocated to each client in the same proportion. If it is measured in terms of number of frames dropped, the dropped frames should be similar for different requests.

5. Scheduling Algorithms

In this section, we first examine two known scheduling methods in the literature, which are considered fair in bandwidth allocation. Then we present BSRB, which integrates the key concepts of those two methods to support the fairness while improving the adaptability.

5.1. Rate-Based Scheduling

In the Rate-Based (RB) method [1], each cell maintains a fixed pool of bandwidth reserved for serving requests from handoff clients to reduce the request drop probability. In order to lower the request block probability and the request drop probability, it has incorporated a borrowing scheme. When a request is generated, each request specifies its desirable bandwidth M and minimum bandwidth m to the system. The difference between M and m is called bandwidth loss tolerance (BLT). A fraction of BLT, called actual borrowable bandwidth (ABB) is divided into several service levels as shown in Figure 2.

All the requests in the same cell are served at the same level. If the cell does not have enough bandwidth

to accommodate an incoming request, the existing requests may temporarily give up a certain amount of bandwidth by moving down to a lower service level. As soon as bandwidth becomes available (by terminating requests or due to a mobile client leaving the cell), the borrowed bandwidth will be returned to the degraded requests. The most important feature of the Rate-Based method is that no request will be served below its minimum bandwidth requirement once it is admitted. It is a fair method since existing requests are served at the same service level and the impacts of overloading on all the existing requests are similar as they give up the same proportion of bandwidth to deal with the overloading situation. In addition, the probability of request drop probability is low by using bandwidth reservation.

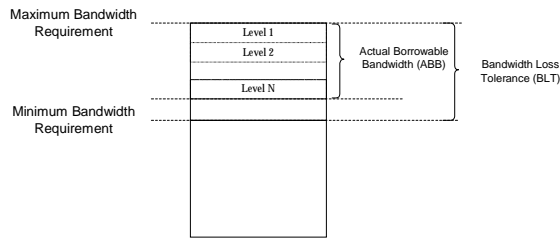


Figure 2: Service levels in the Rate-Based Method

5.2. Buffer Sensitive Bandwidth Allocation (BSBA)

To improve the adaptability of the system and minimize the impact of transient overloading, a client may maintain sufficient video frames at the buffer. The playback of a video will start only after a buffer level, called *preferred buffer level (PBL)*, has been reached. Based on the play rate of a video, the system determines the *buffer playback duration (BPD)* of the video frames at the buffer. The main idea of the Buffer Sensitive Bandwidth Allocation (BSBA) [8] method is to allocate the limited mobile bandwidth at a base station to serve the concurrent requests in the cell based on their playback buffer durations.

The allocation of bandwidth to serve a request in BSBA is divided into two phases. In the first phase, the bandwidth allocated to serve a client request is the minimum of the bandwidth that is equally divided from the available bandwidth, and the mean bandwidth requirement of its requested video:

$$BW_Ad_i = \min\left(\frac{BW_Available}{n}, BW_Consume_i\right)$$

where BW_Ad_i is the bandwidth allocated to serve request i and n is the number of concurrent requests which are not in error state. After first phase allocation, the remaining bandwidth, if any, will be allocated to the clients based on their buffer playback duration, which is determined as the playback time of the latest packet at

the client buffer minus the current time. In the second phase, more bandwidth will be allocated to the client whose buffer playback duration is shorter:

$$BW_AS_i = \frac{(Buffer_Playback_Duration_i)^{-1}}{\left(\sum_{i=1}^n Buffer_Playback_Duration_i^{-1}\right)} \times (1 - BW_AdFR)$$

where BW_AdFR is the total bandwidth allocated at the first phase.

5.3 Buffer Sensitive Rate-Based (BSRB)

By allocating bandwidth based on the agreement at admission such as in RB, the system can guarantee the amount of bandwidth to be allocated to the clients. Even though the probability of errors in communication is high for a mobile client at a position of low communication signal, the services to other clients will not be seriously affected. However, the main problem of such a fixed allocation scheme is that the services to the requests are non-adaptable to the changing playback status of the requests. A client may temporarily need the bandwidth more than its average workload for a short period of time due to transient overloading or communication errors. When its playback situation becomes better, the system may allocate smaller amount of bandwidth.

To allow higher flexibility in services and to adaptively serve the requests based on their playback status, we include the playback situations of requests as a factor in determining how to allocate bandwidth to serve each request. The playback status is reflected on the buffer level at the requesting client. Video buffer at a client can be considered as a reserved bandwidth of the client. The buffer level at a client can be easily calculated by the server based on the playback time of the last frame transmitted to the client. Based on the buffer level, the buffer playback duration (BPD) at a client is determined. If BPD is large, it can tolerate a longer period of transient overloading and more communication errors. Thus, the client with a high BPD and more buffered frames may *temporarily lease* some of its amount of bandwidth to serve the client whose playback status is poor. The calculation of the original amount of bandwidth to be assigned to each client may follow the principles suggested in the Rate-Based method. The system divides the services into levels based on the minimum and expected workload requirements of each request. All the requests in the same cell will be served at the same level. Then, the actual amount of bandwidth to be allocated to a request I at service level n is calculated using the following equation:

Amount of bandwidth for request I at level n – (buffer level at client I / bandwidth adjustment period)

Bandwidth adjustment period is a pre-defined tuning parameter and it indicates the amount of buffer bandwidth to be *leased* to other requests in each period. The definition of bandwidth adjustment period is based on the service cycle time which is the time required to serve all the requests once. The remaining bandwidth will be allocated according to the buffer playback durations of the clients with attempt to build up the buffer levels of the clients, especially the ones whose buffer levels are low. More bandwidth will be allocated to the client with smaller buffer playback duration. If an overloading situation occurs due to the migration of a client from other cell, or due to communication errors, the server may move to a lower level. Then, the new bandwidth allocated for each request will be used to calculate the actual bandwidth to be allocated to each client. To minimize the request drop probability from the handoff clients, certain amount of bandwidth is reserved for handoff requests as in the RB method. However, when considering the admission of handoff requests, the buffer level of the requesting client is also considered. If it is rejected, it will not be dropped immediately. It will wait until its buffer is empty. It is possible that by using the buffered video data, the request drop probability for handoff requests can be reduced.

6. Performance Evaluation and Results

In order to investigate the performance characteristics of the proposed BSRB method, we have developed a simulation program to simulate a distributed mobile video player system introduced in Section 3. In order to simplify the simulator, we only simulate a single cell with a base station and a number of mobile clients. The mobile clients may move into the simulated cell or move out of the cell while their requested videos are playing. To simulate a video request from a mobile client which is moving into the cell, we define the first video frame requested by the client and its initial buffer level, after a video request is generated. Similarly, we specify the last video frame to be requested by the client in order to simulate a client which is moving out of the cell while its requesting video is playing.

The set of mobile clients in the cell are divided into three groups and each group of clients has different buffer sizes and request videos with different workload characteristics, i.e., heavy workload, medium workload and light workload requests. We have included an error model to model the mobile communication problems between a client and a base station. The error model consists of two components: *error probability* and *error duration*. The error probability is used to model the probability of a client in communication error state. It may be a result of the interferences from the

surrounding buildings or a result of being too far away from the base station responsible for the cell. The error duration is the mean duration of a client in error state each time.

Mean stream length	10,000 GOPs
Network bandwidth	1Mbps
Number of client Groups	3
Mean bandwidth of video requested by Class 1 (BW_Class1)	64Kbps
Mean bandwidth of the video to be requested by Class 2	1.4Kbps
Mean bandwidth of the video to be requested by Class 3	640bps
Number of clients per group	50
Think time of the clients between each request	100 ~ 300 sec uniformly distributed
Error possibility	0.3
Error duration	100 sec
Preferred buffer level (PBL)	2 sec
Simulation length	500,000 sec
Mean cell stay time	1000 sec
Number of service level	3
Bandwidth reserved for handoff requests	10% of total bandwidth in a cell
Service factor	1.0

Table 1: Model parameters and baseline values

As shown in Figure 3, the frame lost rate of RB is consistently much higher than that of BSBA and BSRB. This is consistent with our expectation since RB is not adaptive to the changing playback status of individual request in bandwidth allocation although it is a fair policy. Considering the buffer levels of clients in bandwidth allocation can help to maintain the buffer levels at the clients to make the video playbacks less sensitive to the changes in video workload and communication errors. As depicted in Figure 3, the frame lost rates of BSBA and BSRB are similar.

As shown in Figure 4, the request drop rate of BSBA is higher than both RB and BSRB especially when the workload is heavy, i.e., large number of mobile clients. At a heavier workload, the probability of failure in admission is higher. The problem of admission failure due to heavy workload is more serious in BSBA than in RB and BSRB. In RB and BSRB, the services to a video request are divided into levels. If the workload is heavy, the service level of the existing requests in the cell is lowered in order to admit more new requests. Thus, their request drop rates are smaller. The request drop rate of BSRB is marginally better than RB since the buffer levels of the clients are in general higher under BSRB. They can tolerate a longer queuing time for admission. Also due to the higher buffer levels at the clients, the mean service level to the requests is higher in BSRB than in RB as shown in Figure 5.

7. Conclusions

In this paper, we have studied the problem of bandwidth allocation in serving video requests from mobile clients in a cellular mobile network. We first examined the bandwidth allocation problem in such an environment as compared with the problem in conventional distributed video player systems. The main objectives in bandwidth allocation are to provide fair services to all the concurrent video requests and at the same time to maximize the performance of individual video playback such as to minimize the number of dropped video frames. We have identified the performance problems of the Rate-based (RB) method and applied the concept of service levels into the Buffer Sensitive Bandwidth Allocation (BSBA) method in the design of the Buffer Sensitive Rate-Based (BSRB) algorithm. In BSRB, the allocation of bandwidth to serve a video request depends on the buffer levels of the clients in the same cell. If the buffer level of a client is low, more bandwidth will be allocated to serve it. In order to minimize the request drop probability, which is a failure in admission as a result of heavy workload at a cell, the service level concept is also adopted in BSRB. The simulation results are consistent with our expectation. BSRB not only can reduce the request drop rate as compared with BSBA, its frame lost rate is also lower comparing with RB.

References

- [1] Mona El-Kadi and Stephan Olariu, "A Rate-Based Borrowing Scheme for QoS Providing in Multimedia Wireless Networks", *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 2, pp. 156-166, 2002.
- [2] Songwu Lu, Vaduvur Bharghavan, R. Srikant, "Fair Scheduling in Wireless Packet Networks", *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473- 489, 2000.
- [3] Hang Liu and Magda El Zarki, "Adaptive source rate control for real-time wireless video transmission", *Mobile Networks and Applications*, vol. 3, pp. 49-60, 1998.
- [4] T. S. Ng, I. Stoica, H. Zhang, "Packet Fair Queuing Algorithms for Wireless Networks with Location-Dependent Errors", in *Proceedings of INFOCOM '98*, vol. 3, 1998. Pg 1103-1111.
- [5] Carlos Oliveira, Jaime Bae Kim, Tatsuya Suda, "An Adaptive Bandwidth Reservation Scheme for High-Speed Multimedia Wireless Networks", *IEEE Journal on Selected Areas in Communications*, vol. 16, no.6, pp. 858 – 874, 1998.
- [6] P. Rammanathan, P. Agrawal, "Adapting Packet Fair Queuing Algorithms to Wireless Networks", in *Proceedings*

of 4th Annual ACM/IEEE international conference on Mobile Computing and Networking, Oct 1998.

[7] S. Shakkottai and R. Srikant, "Scheduling Real-time Traffic with Deadlines over a Wireless Channel", in *Proceedings 2nd ACM Wireless Mobile Multimedia*, 1999.

[8] Joe Yuen, Kam-Yiu Lam, and Edward Chan, "A Fair and Adaptive Scheduling Protocol for Video Stream Transmission in Mobile Environment", in *Proceedings of IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 2002*.

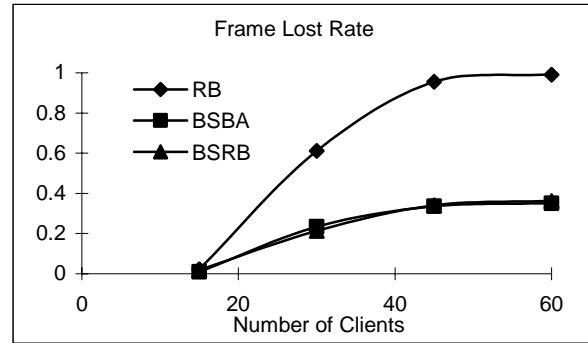


Figure 3: Frame lost rate Vs. Number of mobile client

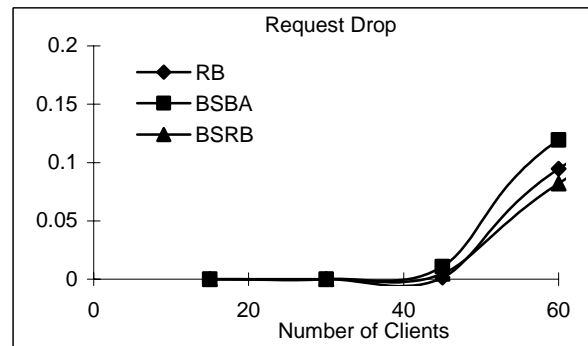


Figure 4: Request drop rate Vs. Number of mobile clients

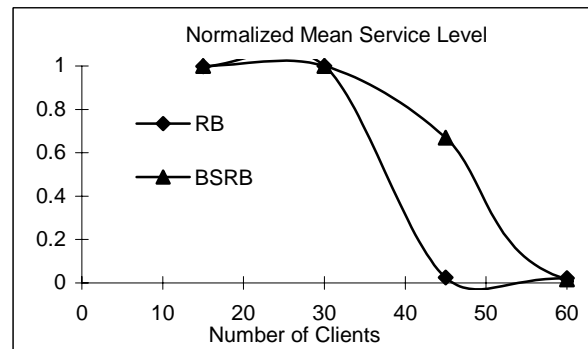


Figure 5: Normalized mean service level Vs. Number of mobile clients