

# Scheduling Video Stream Transmissions for Distributed Playback over Mobile Cellular Networks

Kam-Yiu Lam<sup>1</sup>, Joe Yuen<sup>1</sup>, Sang H. Son<sup>2</sup> and Edward Chan<sup>1</sup>

Department of Computer Science<sup>1</sup>  
City University of Hong Kong  
Tat Chee Avenue, Kowloon, Hong Kong  
{cskylam|csjyuen|csedchan}@cityu.edu.hk

Department of Computer Science<sup>2</sup>  
University of Virginia  
Charlottesville, VA22903  
son@cs.virginia.edu

## Abstract

In this paper, we address the issues in mobile bandwidth scheduling for video stream transmissions in a distributed video player system supported on cellular network. Two important concerns in the design of the scheduling algorithms are fairness in services and adaptability to the changing playback condition of individual video. 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. Adaptability is an important issue in video streaming over mobile cellular networks, since different clients may be subjected to different probabilities of communication errors and also the workload in each cell is highly dynamic due to the mobility of clients. 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. It is expected that 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. In addition, by employing the rate-based policy in BSRB, the performance of the requests will be less affected by the poor performance of a single client and hence fair services can be provided to the clients. Extensive simulation experiments have been performed to investigate the performance characteristics of the proposed BSRB algorithm as comparing with other algorithms under different system settings and network failure probabilities.

*Keywords: mobile multimedia, cellular mobile network, buffering, bandwidth, and real-time scheduling*

# 1 Introduction

In a distributed mobile video player system, mobile clients are connected to a video sever 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.

Although various efficient techniques, e.g., video frame buffering, feedback control mechanisms, and video stream smoothing techniques, have been proposed to improve the performance of distributed video player systems supported on wired networks [1, 3, 10, 11, 13], these techniques may not be suitable to mobile video player systems due to the unique characteristics of the mobile systems. The asymmetric bandwidth property of a mobile network limits the amount of feedback messages from mobile clients to the server and poses challenges on the flow-control mechanisms for video stream transmission. Mobility of clients may seriously affect the distribution of workloads in the system and the effectiveness of admission control mechanisms in maintaining the quality of video playbacks. It is difficult to guarantee the playback quality in such a poor and unstable network environment [6]. Even though the bandwidths of mobile networks are improving, the demand on high video quality is also growing. It is believed that even with the support of the 3G mobile networks, how to allocate the mobile network bandwidth to serve different video requests is still a big concern on the performance of a mobile video player system.

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 [6, 8]. 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. Since multiple requests may exist in a cell at the same time, how to serve them concurrently in using the limited mobile bandwidth in a fair way is an important issue for the video playback performance. As different clients may request videos with very different workload requirements, a simple first come first serve method is obviously unsuitable and unfair. In [4], 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 it 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 [4]. 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. It is expected that 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.

The remaining parts of the paper are organized as followings. Section 2 reviews the related work in the area. Section 3 defines the system model and summarizes the system characteristics. Section 4 defines the system performance objectives and the importance of fairness and adaptability in this kind of systems. Section 5 discusses two methods proposed in the literature and then presents our novel Buffer Sensitive Rate-Based method. Section 6 presents the performance studies and the discussion of simulation results. Finally, Section 7 concludes the paper.

## 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, i.e., earliest deadline first and rate monotonic, have been extended for scheduling the transmission of video packets in distributed multimedia systems. Other important concerns, which are specifically important 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 [12], the earliest deadline first scheduling algorithm is extended for scheduling of video packets in a mobile environment. In [9], 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) [7] 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 system, i.e., an error-free system. In [2], a hierarchical admission scheme is proposed to solve the admission problem when the system is a cellular system and the mobile clients have high mobility. Although they have addressed some of the issues, they have not considered both fairness and adaptability using client buffers. In this paper, we concentrate on the design of efficient scheduling methods based on the demand of each request and the buffer levels of the clients.

In [4], a rate-based (RB) method is proposed by introducing the concept of service levels in servicing the client requests in a cell with the objective 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 [14], 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. In Section 5, we provide more detail descriptions on the rate-based method and the buffer sensitive bandwidth allocation method.

### **3 System Model and Characteristics**

#### **3.1 System Model**

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. It is the objective of the paper to design efficient methods for serving the requests from mobile clients in a cell to effectively use the limited mobile bandwidth.

The video server maintains a collection of compressed videos, i.e., compressed in the MPEG II standard. Each encoded video stream consists of a sequence of *group of picture (GOP)* and each GOP is composed of three types of frames: P, B and I frames. Mostly, the size of I frame is much larger than a B frame or a P frame. Due to the different sizes of the GOPs and fixed play rate of a video, the bandwidth requirement of a video stream is variable.

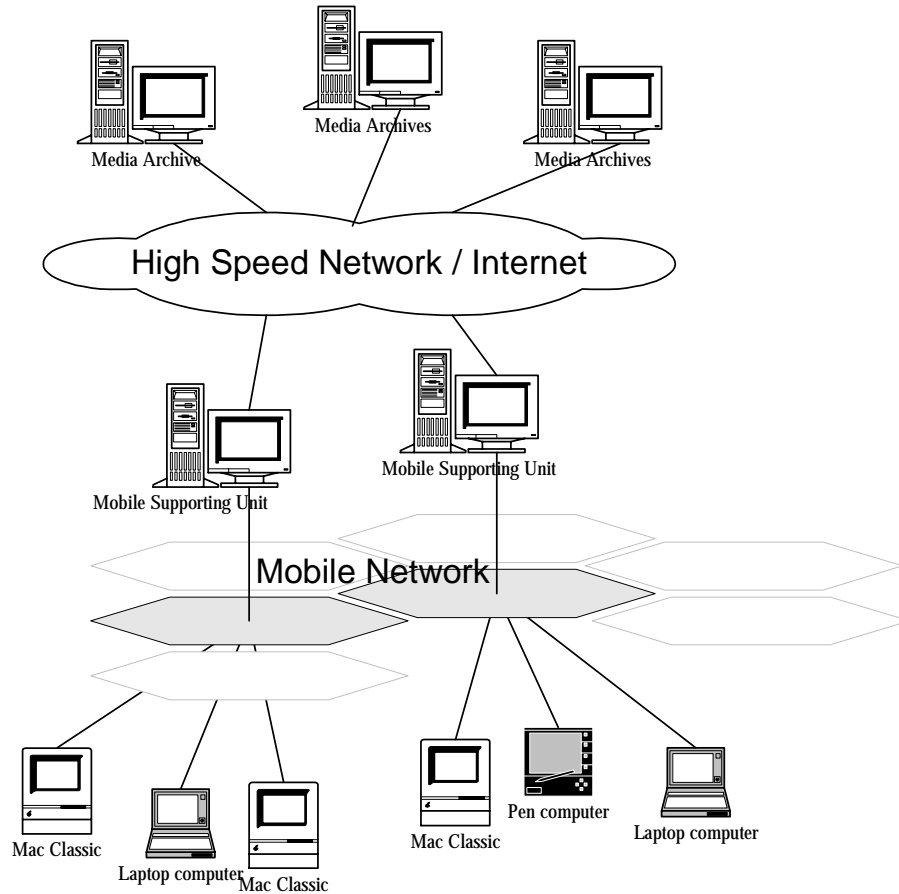


Figure 1: A distributed mobile video player system on a cellular network

The number of mobile clients in a cell is not fixed since the mobile clients may move into another cell or move from other cells into the cell. The mobile clients may 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. For a thin client, the video buffer size may be very limited, i.e., 50 kbytes or even less. 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.

The mobile client places the received frames at its video buffer. The playback of a video will start after the video buffer level has reached a pre-defined value. Then, the decompression procedure will start and the playback of the video will begin by forwarding the decompressed frames to the video player one

by one. However, if the playtime of a video frame has been missed, the frame will be dropped immediately. This will affect the performance of the video player as this could seriously affect the smoothness of the video playback.

When the server receives a video request, it will perform an admission test. If it can pass the test, the server will start to serve it. As shown in Figure 2, the video server may be serving multiple client requests at the same time. 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.

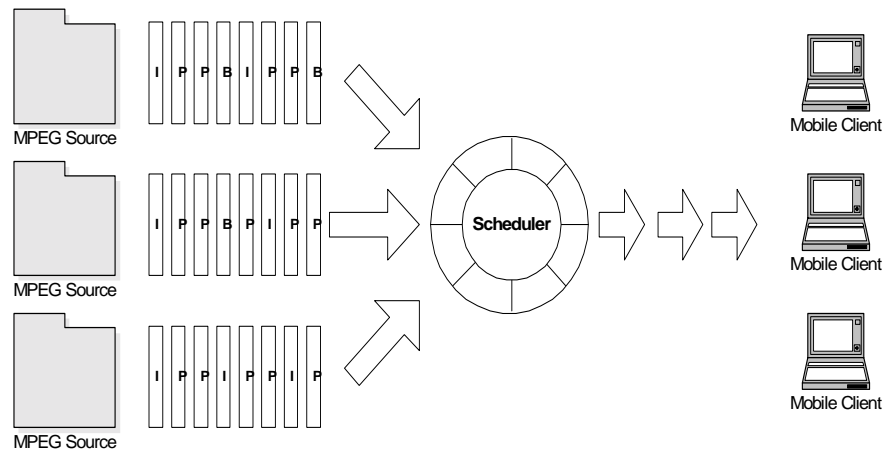


Figure 2: A server is serving multiple mobile client requests

### 3.2 System Characteristics

In the followings, we will discuss the main differences between a distributed mobile video player system and a distributed video player system supported on a wired network. The unique system characteristics make the efficient methods designed for conventional distributed video player systems unsuitable for mobile video player systems.

*Low Bandwidth:* Although the bandwidth of a mobile network is improving, it is still much lower than that of a wired network. For example, for a wireless LAN, although the bandwidth is up to several

Mbps, it is still much less than that of a wired LAN. The bandwidth of cellular radio network is even much lower. The GSM systems can only support up to 9.6 kbps (or 14.4 kbps) for data transmission. Although new technologies such as GPRS and 3G networks are emerging, the 3G networks is expected to have a maximum of 2Mbps which is still extremely low compared to conventional wired networks. In this paper, we are assuming a cellular radio network, which can support up to 2Mbps, e.g., 3G, in which the total bandwidth supported in each cell is fixed. Note that the speed of a mobile client may affect the effective bandwidth in serving the client.

*High Transmission Error and Disconnection:* The probability of transmission errors in a mobile network highly depends on the surrounding environment conditions of the mobile clients. Therefore, different mobile clients may observe different error probabilities and disconnections may be quite frequent. Disconnection may be temporarily and can be resolved after several re-transmissions. Sometimes, the disconnection is more permanent and the communication channel will only be re-established after a long period, i.e., after the mobile client has moved to a new location where the communication signal is stronger. In serving the clients, it is important to minimize the impact of the high errors of a particular client on the services provided to other clients.

*Asymmetric network bandwidth:* Normally, in a cellular network, the bandwidths of the down-link channels are comparatively much higher than the up-link channels. In the past few years, various feedback control techniques are proposed in which the scheduling for transmission of video frames to serve a client is based on the feedback playback status of the client. However, due to the asymmetric bandwidth of a mobile network, it is important to minimize up-link messages and this makes such feedback mechanisms less effective.

*Mobility of clients:* Mobility of clients affects the workload distribution in the system and can also have serious impact to the admission control mechanism. Normally, admission control procedure defined over the existing workload and the workload characteristics of the requesting video to limit the workload of the system. However, most of the proposed admission control procedures have ignored the situation where the admission request is generated from a mobile client which is moving into the cell while its video playback has already started. If the new cell of a client does not have sufficient bandwidth to continue the video playback, the possible consequence may be a drop of the video playback. This is called request drop and is highly undesirable.

*Different client buffer sizes:* In order to minimize the impact of network jitter, different buffer techniques have been proposed. However, in a mobile environment, some of the mobile machines may be thin-clients, e.g., PDA or handheld PC, while others are thick clients such as notebook computers. It is common that the different clients may have different buffer spaces for the playback of their requested videos. Therefore, the impact of data transmission and overloading on the performance of the video playback on different clients can also be very different. The clients, whose video buffer size is small, will be more sensitive to network errors than the clients with larger buffers.

*High variations in video workload characteristics:* One of the main causes of the problems in supporting high quality of video playback is due to the variable bit rates of the videos. The problem will be more serious in a mobile multimedia system. It is because the clients may request videos with great differences in their workloads. Some of the clients, e.g., those heavy clients, may request video with a greater workload, i.e., greater size and higher resolution, while the thin clients may requests small video clips whose workload is much smaller. The great differences in the video workloads of different clients could create problems in maintaining the fairness in serving the requests especially under an overloading condition.

#### **4 Performance Objectives and Fairness in Services**

The prime performance objective of a video player system is to support high quality video playbacks. 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). Dropping video frames affects the smoothness of a video playback and has to be minimized. In addition, each video request may associate with a minimum performance requirement, which may be defined on the percentages of played frames. If the system cannot provide an “acceptable” performance in the playback, i.e., at least meeting the minimum performance requirement of the request, it is better not to start the video playback and the system should reject the request. 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 the followings, we first discuss the admission control issue and then examine the fairness issues in such an environment.

#### 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. For example, Figure 3 shows the frame size variation of three video streams over a minute. The mean frame sizes of the three video streams are 5, 25 and 56 Kbytes. Transient overloading occurs at times 5, 12, 16, 19, 26, 45, 49, 53 and 60, i.e., when the summation of the frame sizes of all the streams is higher than the summation of the line Mean Total. The overloading situation is transient since it may only last for a short period of time depending on how the frame sizes vary. 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.

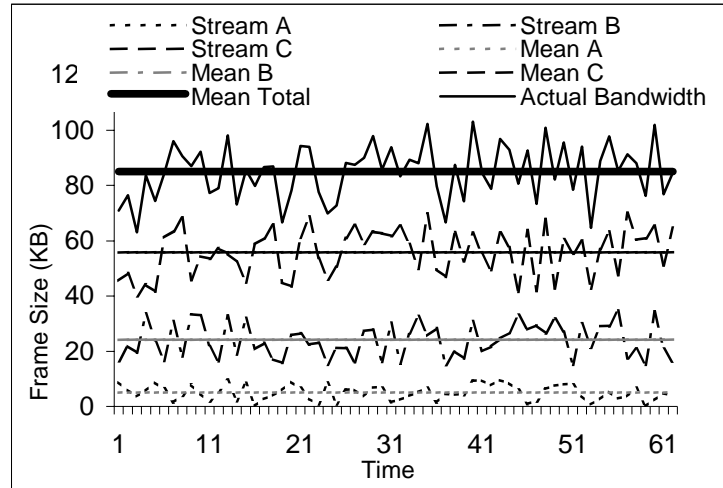


Figure 3: Variable video workloads

## 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 (as discussed in Section 3) [5]. Basically, we can divide the fairness issue into two parts: *fairness in admission* and *fairness in services*. When a request comes in, it needs to pass through an admission procedure. If it fails in the admission, it may be rejected or it may wait and queue for checking admission later. If the admission condition is defined on system workload and the expected workload of the request, a heavier workload request will have a higher probability to be rejected than a lighter workload request. If a rejected request is allowed to wait, the waiting time of a request depends on the queuing disciplines for admission. If the system is allowed to admit a light workload request even a heavy workload request is waiting, the waiting time of the heavy workload request will become indefinite. However, if the first-in-first-out discipline is used for queuing to check for admission, the utilization of the system resources could be seriously affected since some of the resources have to be reserved for the heavy workload request. To be fair, the waiting time of the requests should be proportional to the amount of resources required by them. For example, consider a client requesting a video with mean workload of  $W$  and another client requesting a video with mean workload of  $W/5$ . It can be considered fair if the waiting time of the first client should be around 5 times of the second client.

Another issue in admission control is the admission of handoff clients. When a mobile client enters a new cell while its requested video is playing, it has to pass through an admission procedure. Since it is highly undesirable to drop the playback due to admission failure, the admission condition defined for handoff clients may be comparably less restrictive than the conditions for new requests. The system may

reserve certain amount of bandwidth specifically for handoff requests. However, the problem is how much bandwidth should be reserved. If the amount is large, it will significantly affect the services to new requests and the total system utilization.

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 scale. 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.

Communication errors makes the actual amount of video data received by a client much smaller than the amount of bandwidth allocated to serve it. If errors occur in communicating with a client, the system may need to allocate more bandwidth to compensate the errors if it wants to maintain the minimum performance of the request. In this case, fairness in services may not be able to be maintained since the performance of one request may seriously affect the performance of other existing requests. If the communication errors are serious and have been lasted for a long time, it is impossible to guarantee the minimum performance. In that case, it may be preferable to reduce the amount of bandwidth allocated to the client with communication errors in order to provide a fair service to other existing requests.

In this paper, we concentrate on the problem of fairness in services. In the next section, we present a novel method called *Buffer Sensitive Rate-Based (BSRB)* algorithm, which aims to maximize the playback performance of individual request, while serving all the requests in a fair manner.

## 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. To deal with the issue of fairness in admission of handoff requests, we use a reservation scheme as the one proposed in the rate-based method [4]. The system reserves certain amount of bandwidth for the admission of handoff requests in order to reduce the request drop probability.

## 5.1 Rate-Based Scheduling

In the Rate-Based (RB) method [4], 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 4.

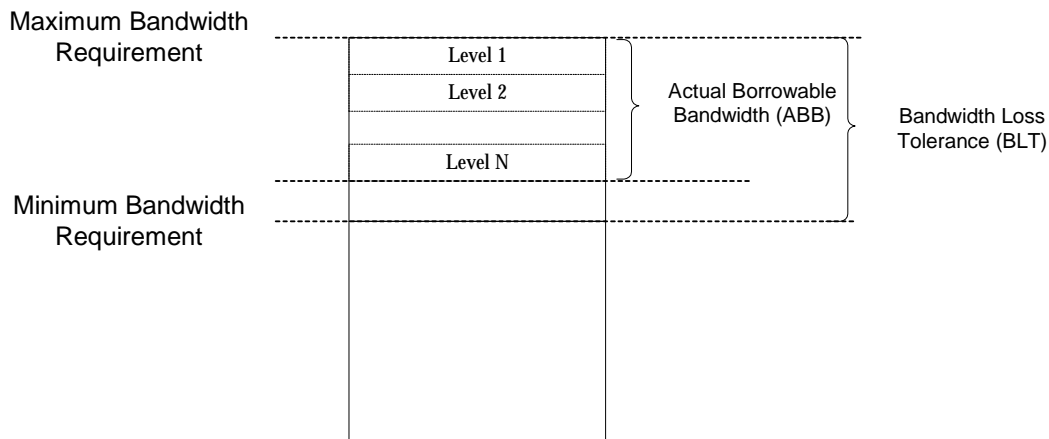


Figure 4: Service levels in the Rate-Based Method

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, i.e., due to terminating request or 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.

## 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. It is the period of time that the buffer will become empty if no frame arrives during that period of time. The main idea of the Buffer Sensitive Bandwidth Allocation (BSBA) [14] 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:

$$BW\_Ad_i = \min\left(\frac{BW\_Available}{n}, \overline{BW\_Consume_i}\right)$$

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. The rationale of such allocation scheme is to provide a fair treatment to all the clients independent of the workload characteristics of their requested videos. If the average required bandwidth of a client is smaller than the mean value,  $(BW\_Avaliable/n)$ , its average bandwidth will be allocated. Although the size of the next packet for transmission may be larger than the average workload of the video, allocating average bandwidth of the video to serve a client should not significantly affect the playback if the client buffer level is high enough. The buffer level can be maintained at the preferred level in a sufficiently long period of time when the client receives the average workload of its requested video.

The total bandwidth allocated at the first phase,  $BW\_Ad_{FR}$ , is:

$$BW\_Ad_{FR} = \sum_{i \in N} \min\left(\frac{BW\_Available}{n}, \overline{BW\_Consume_i}\right),$$

where  $N$  is the set of clients 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. Obviously, the playback will be less affected by transient workloads if its playback duration is longer. Therefore, in the allocation of the bandwidth in the second phase, more bandwidth will be allocated to the client whose buffer playback duration is shorter:

$$BW_{ASi} = \frac{(Buffer\_Playback\_Duration_i)^{-1}}{(\sum_{i=1}^n Buffer\_Playback\_Duration_i^{-1})} \times (1 - BW\_AdFR)$$

It is expected that by allocating more bandwidth to serve a client whose video buffer level is low, we can restore its buffer level to the preferred buffer level.

### 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 need more bandwidth temporarily 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 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 which 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 one which buffer level is low. More bandwidth will be allocated to the client with smaller buffer playback duration. If 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 using the simulation language called CSIM 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. After a video request has been generated from a mobile client, we will define the first video frame to be requested by the client and its initial buffer level to simulate that it is a video request from a mobile client, which is moving into the cell. Similarly, we specify the last video frame to be requested by the client in order to simulate that the client may move 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.

In addition to modeling BSRB, we also have developed two other simulation programs to simulate the same system using RB and BSBA for comparison purposes. Similar to BSRB, in both RB and BSBA, certain amount of bandwidth is reserved for handoff requests in order to minimize the drop probability. The admission control in BSRB follows the principles used in the BSRB model. It is defined based on the average workload requirement of a request and the existing workload in the system.

The following table shows the key model parameters and their baseline setting values.

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

The main performance measures are the frame lost rate, request block time, request drop rate and normalized mean service level. The frame lost rate is an indicator of the playback performance of the videos. It is defined as the number of the frames dropped over the total number of frames requested. If a

request is failed in admission, it will wait in the queue for admission in a first come first served manner. A request will be dropped if it is still failed in admission until its buffer is empty. Request drop rate is defined as the number of requests dropped due to failure in admission over the total number of admission requests. The normalized mean service level is only for RB and BSRB, since they are the ones that divide the services into levels. Normalized mean service level is calculated as the mean service level to the clients divided by the total service levels in the system.

As shown in Figure 5, 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 and can make the video playbacks less sensitive to the changes in video workload and communication errors. As depicted in Figure 5, the frame lost rates of BSBA and BSRB are similar.

As shown in Figure 6, 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 video 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 7.

## **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 problems 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. We have developed a simulation system to simulate a distributed mobile video player system and simulation experiments have been performed to compare BSRB with RB and BSBA. The simulation results are consistent with our expectation. BSRB not only can reduce the request drop rate as compared with BSRB, its frame lost rate is also lower comparing with RB. Currently, we are implementing BSRB in a distributed mobile video player system. Another important future work we are considering is how to include the client buffer levels of handoff clients for admission control to reduce the request drop probability. If the buffer level of a client is high, a looser condition may be used.

## References

- [1] R. Agarwal and A. M. K. Cheng, "Reducing Variation in Bit-Rate Produced by Encoder in MPEG Video", in *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, June 1999.
- [2] Sunghyun Choi and Kang G. Shin, "A comparative study of bandwidth reservation and admission control schemes in QoS-sensitive cellular networks", *ACM Wireless Networks*, vol. 6, no. 4, pp. 289-306, 2000.
- [3] Shanwei Cen, Calton Pu and Richard Staehli, "A Distributed Real-time MPEG Video Audio Player", in *Proceedings of the 5th International Workshop on Network and Operating System Support of Digital Audio and Video*, 1995.
- [4] 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.
- [5] Songwu Lu, Vaduvur Bharghavan, R. Srikant, "Fair Scheduling in Wireless Packet Networks", *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473- 489, 1999.
- [6] 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.
- [7] 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.
- [8] 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.

- [9] P. Rammanathan, P. Agrawal, “Adapting Packet Fair Queuing Algorithms to Wireless Networks”, in *Proceedings of 4<sup>th</sup> Annual ACM/IEEE international conference on Mobile Computing and Networking*, Oct 1998.
- [10] S. Rao and A. M. K. Cheng, “Scheduling and Routing of Real-Time Multimedia Traffic in Packet-Switched Networks”, in *Proceedings of IEEE International Conference on Multimedia*, July-Aug. 2000.
- [11] Cormac J. Screenan, Jyh-Cheng Chen, Prathima Agrawal, B. Narendran, “Delay Reduction Techniques for Playout Buffering”, *IEEE Multimedia*, vol. 2, no. 2, pp. 88-100, 2000.
- [12] S. Shakkottai and R. Srikant, “Scheduling Real-time Traffic with Deadlines over a Wireless Channel”, in *Proceedings 2nd ACM Wireless Mobile Multimedia*, August 1999.
- [13] David K. Y. Yau, and Simon S. Lam, “Adaptive Rate-Controlled Scheduling for Multimedia Application”, *IEEE/ACM Transactions on Networking*, vol. 5, no. 4, 1997.
- [14] 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.

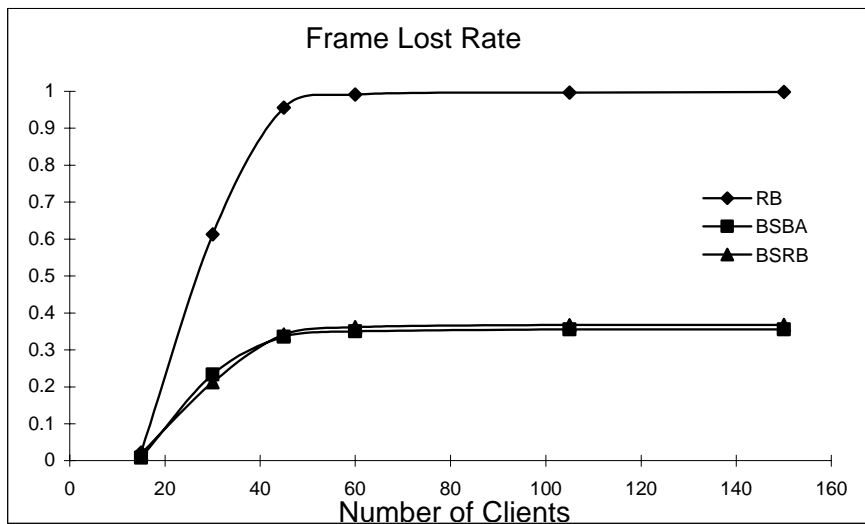


Figure 5: Frame lost rate Vs. Number of mobile clients

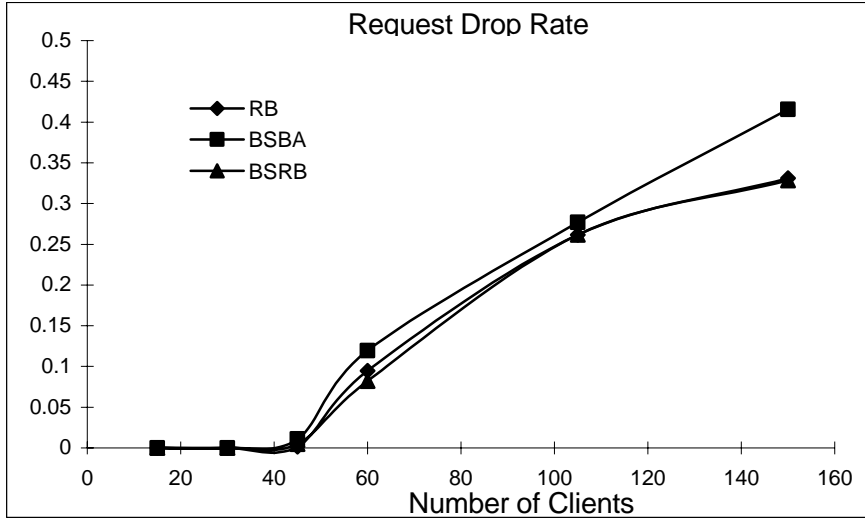


Figure 6: Request drop rate Vs. Number of mobile clients

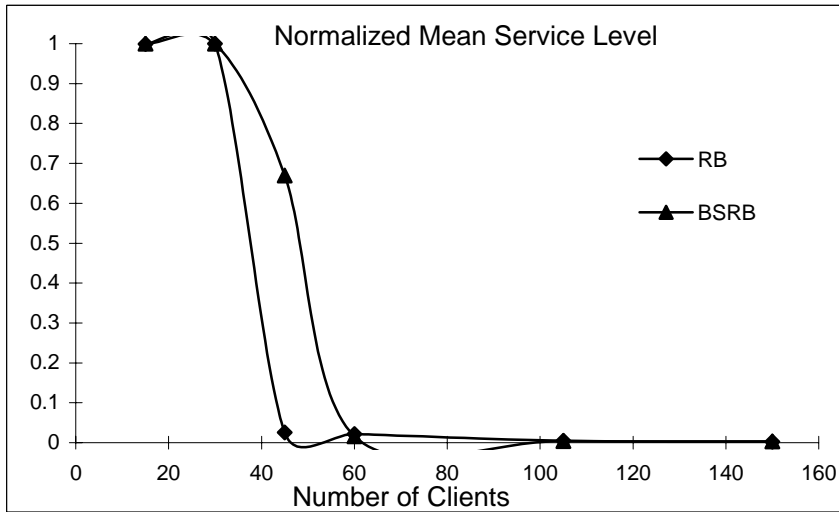


Figure 7: Normalized mean service level Vs. Number of mobile clients