

# 國立交通大學

電子工程學系

碩士論文

cdma2000 基於服務品質利用適應性視訊串流的無線



**QoS-based RRM Exploiting Adaptive Streaming**

**Video over cdma2000**

研究生：邱彥翔

指導教授：黃經堯 博士

中華民國九十四年八月

cdma2000 基於服務品質利用適應性視訊串流的無線資源管理

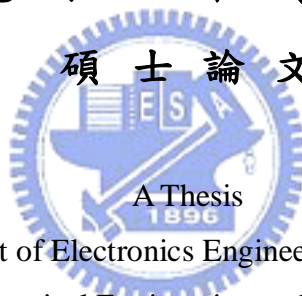
**QoS-based RRM Exploiting Adaptive Streaming  
Video over cdma2000**

研究生： 邱彥翔  
指導教授： 黃經堯

Student: Chiu Yang-Shiang  
Advisors : Huang Ching-Yao

國立交通大學  
電子工程學系

碩士論文



Submitted to Department of Electronics Engineering & Institute of Electronics  
College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Electronics Engineering

August 2005

HsinChu, Taiwan, Republic of China

中華民國九十四年八月

# cdma2000 基於服務品質利用適應性視訊串流的無線資源管理

研究生：邱彥翔

指導教授：

黃經堯 博士

國立交通大學  
電子工程學系 電子研究所碩士班

## 摘要

本論文旨在提出一個能夠同時增進系統資源使用率及提昇客戶服務品質的無線資源配置方案，稱之為客戶回饋輔助性無線資源管理。客戶回饋輔助性機制這樣的概念可以被延伸應用到企圖提供多媒體服務，特別是即時視訊串流服務，的所有無線通訊系統上。每一次的資源重新配置包含了對使用者在前端的無線系統使用的優先權排序及後端多媒體伺服器的串流壓縮率調整。客戶回饋輔助性機制利用了客戶緩衝器的資訊、目前基地台端的資料排隊量、無線頻帶的狀況以及系統的負載量來決定優先權跟壓縮率。相較於結合了客戶主導性(client-based)或無線網路回饋性(radio network feedback)調變方案的公平排程無線資源管理，模擬結果顯示了當系統負載被考量品質的傳呼允許機制(call admission)限制在合理的容量下，客戶回饋輔助性機制能夠提昇系統總處理資料量並降低客戶緩衝器匱乏的機率。

# QoS-based RRM Exploiting Adaptive Streaming Video over cdma2000

Student: Yang-Shiang Chiu

Advisor:

Dr. ChingYao Huang

Department of Electronic Engineering &  
Institute of Electronics  
National Chiao Tung University

## Abstract

This thesis proposes a new radio resource allocation (RRM) scheme, called client feedback assisted (CFA) RRM, to improve system resource utilization rate and at the same time to enhance individual quality of service (QoS). The concept of CFA is applicable to all wireless communication systems which tend to serve multimedia applications, especially real-time video streaming. The resource relocation includes rescheduling priorities in a front-end wireless system and reassigning source bit rate of each streaming in a back-end multimedia server. CFA exploits feedback information from a client buffer, current fullness of base station queue, RF condition, and system loading to decide the priority and source bit rate. As compared with a fairness scheduling RRM, which combines client-based and radio network feedback (RNF) as its streaming adaptation scheme, simulation results show that CFA solution can improve system throughput and decrease the probability of client buffer underflow.

## 誌謝

感謝黃經堯老師二年來的指導，除了在專業上教授了許多基本而重要的核心知識，也讓我從只會讀書的大學生，學習如何發掘問題、解決問題、最後解決問題。此外，在做研究的過中，不時指引修正適當的方向，也讓我學習了如何在關鍵之處要做出果斷的決定或向專家徵詢討論。感謝老師在百忙之中的諄諄教誨，才能使這篇研究能夠順利地完成。

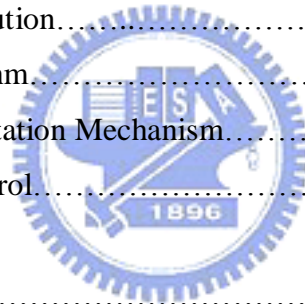
另外，感謝實驗室的夥伴們讓我在短短地兩年中體會了群體合作、協調、努力、相處的經驗。振坤學長、慧源學長、文嶽學長、振哲學長、宜霖學長、明原學長、宜鍵、雲懷、正達、裕隆、建銘、大瑜、勇嵐、鴻輝、昌叡、宗奇、盟翔、域晨，誠摯感謝你們陪我度過了這段有笑有淚的難忘研究生涯。

最後，我要感謝我的家人。有你們默默地付出支持、使我能夠安穩無憂地在交大完成二年的碩士學程。你們的關心、支持與祝福帶给了我無比的動力，讓我的心中充滿著溫暖與感動。目信未來的日子，我也能夠再你們的陪伴下繼續努力研究、進步。



# Contents

Chapter 1 Introduction.....	1
Chapter 2 Overview of RRM over cdma2000 and MPEG-4 FGS Technology.....	4
2.1. 3G Wireless System.....	4
2.2. MPEG-4 FGS Video Compression Technology.....	8
2.2.1. Video in MPEG-4.....	8
2.2.2. Fine Granularity Scalability (FGS) in MPEG-4 Video Standard.....	11
2.2.3. Encoding Adaptation Mechanisms.....	12
Chapter 3 Client Feedback Assisted Radio Resource Management and Encoding Adaptation Mechanism.....	16
3.1. Integrated CFA Solution.....	16
3.2. CFA RRM Algorithm.....	19
3.3. CFA Encoding Adaptation Mechanism.....	26
3.4. Call Admission Control.....	30
Chapter 4 Simulation Results.....	34
4.1. System Perspective Metrics and Performance Analysis.....	34
4.2. User Perspective Metrics and Performance Analysis.....	37
Chapter 5 Conclusion and Future Works.....	44



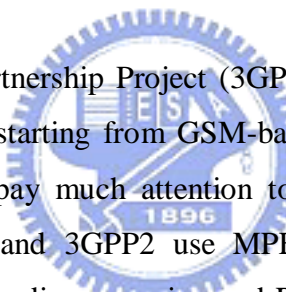
## List of Figures

Figure 2-1. 2G to 3G technology migration road map.....	5
Figure 2-2. FGS encoder structure.....	12
Figure 2-3. FGS decoder structure.....	12
Figure 2-4. client-based adaptation mechanism.....	14
Figure 2-5. RNF adaptation mechanism.....	15
Figure 3-1. CFA Architecture.....	18
Figure 3-2. CFA RRM control flow.....	19
Figure 3-3. three stage structures of client buffer.....	20
Figure 3-4. thresholds of output buffer fullness.....	23
Figure 3-5. value plane of weighting factor W in log scale.....	24
Figure 3-6. CFA RRM algorithm.....	25
Figure 3-7. representation of the timing indices.....	29
Figure 3-8. CFA adaptation algorithm.....	29
Figure 3-9. call admission control flow.....	31
Figure 3-10. call admission control algorithm.....	32
Figure 4-1 system capacity.....	36
Figure 4-2 system blocking rate.....	36
Figure 4-3 channel efficiency.....	37
Figure 4-4 power efficiency.....	37
Figure 4-5 system throughput.....	38
Figure 4-6 average number of buffer underflow during a 100sec clip.....	39
Figure 4-7 proportion of video suspension during a video session.....	39
Figure 4-8 average visual quality.....	41
Figure 4-9 distinguishability of visual quality among video users.....	41
Figure 4-10 vibration of visual quality during a video session.....	42
Figure 4-11 user standard deviation of timing standard deviation of personal throughput.....	42
Figure 4-12 mean of bandwidth switch ratio.....	43
Figure 4-13 standard deviation of bandwidth switch ratio.....	43

# Chapter 1

## Introduction

Evident growth of the needs of multimedia applications and the popularity of various portable devices make human accustomed to enjoy digital entertainment anytime and anywhere. This inevitable trend pushes telecom industry to propose the third generation wireless communication systems—cdma2000 by North America, Universal Mobile Telecommunications System (UMTS) by Europe, and even time division synchronous CDMA (TDSCDMA) by China; these have been designed with the capability of providing high speed data services, ranging from more than hundred kbps to several Mbps. Based on 3G technologies, operators, and contend providers a multimedia platform is created to that simultaneously support heterogeneous services, e.g. voice, video conference, video on demand (VOD), short message service (SMS), multimedia message service (MMS), and etc.



The 3rd Generation Partnership Project (3GPP) defines standards for 3rd generation mobile networks and services starting from GSM-based systems. 3GPP2 does the same for CDMA based systems. Both pay much attention to mobile multimedia. In their wireless terminal specification, 3GPP and 3GPP2 use MPEG-4 simple visual profile for video, MPEG-4 file format for multimedia messaging and RTP, RTSP for streaming protocols and control. The MPEG-4 standards consist of a set of tools providing improved compression efficiency and error resilience. In addition, MPEG-4 provides scalability—called fine granularity scalable (FGS) coding. Using FGS, the encoder generates a base layer and an enhance layer that can be truncated to any amount of bits within a Video Object Plane (VOP). The base layer keeps the most important information that characterizes a video object, and an object can't be reconstructed if base layer packets encounter loss or destruction. The remaining portion improves the quality of the VOP. This means the more FGS enhancement bits are received, the better the reconstructed visual quality is. Besides, no specific bitrate needs to be assigned to the encoder, but only a bitrate range. The encoder generates a base-layer to meet the lower bound of the bitrate range and an enhancement layer to meet the total bitrate constraint. The FGS enhancement bit stream can be sliced and packetized at the transmission time to satisfy the varying user bitrate. This makes FGS suitable for applications with varying transmission bandwidth, especially wireless environment. [3]



Real time streaming services will likely coexist with the basic voice and relatively delay insensitive data services in the near future. For a cdma2000 system without QoS [4], two basic fairness criteria of RRM are proposed—equal rate and equal time. Equal rate tries to ensure that all users can receive nearly the same data rate, irrespective of their RF conditions; equal time lets everyone be able to share parts of resources every time slot, no matter how many resources can each user share. Such kind of RRM fairly distributes system resources, but can't neither optimize utilization efficiency nor satisfy heterogeneous services with different QoS requirement. Thus it is necessary to develop a new RRM algorithm that considers both system performance and service quality. Moreover, media server can also adaptively tune its compression rate to match front-end request. There are two major adaptation schemes—client based, and radio network feedback (RNF). Client based method is the client that detects the bandwidth variations and orders the server to adapt the service by means of RTSP PAUSE/PLAY or new explicit message; RNF means the radio network has the main role to communicate to the server the bandwidth assigned to the connection. [5] However, this adaptation is relatively much slower than wireless variation, means any improper assignment due to insufficient information feedback can result in serious quality degradation. In our work, the proposed algorithm exploits feedback information from client buffer, current fullness of base station queue, RF condition, and system loading to decide the priority and source bit rate.

The rest of this thesis is organized as follows. In Chapter 2, radio resource management (RRM) mechanisms of 3G wireless system and MPEG-4 FGS encoding technology are briefly introduced. In Chapter 3, the proposed solution that integrate RRM and the encoding adaptation mechanism, called client feedback assisted (CFA) solution, is discussed. Simulation results are presented in Chapter 4. Finally, conclusions and discussions of this paper are provided in Chapter 5.

# Overview of RRM over cdma2000 and MPEG-4 FGS Technology

In this chapter, major RRM mechanisms over cdma2000 and MPEG-4 FGS video compression technology are introduced. Section 2.1 describes the main characteristics of CDMA technology and the core RRM mechanisms over cdma2000 1x specification. Section 2.2 depicts both MPEG-4 standard and fine granularity scalable (FGS) coding scheme.

### 2.1. 3G Wireless System

According to the ongoing growth in demand for multimedia applications and high-speed packet data services over mobile wireless networks, advanced system requirements and objectives for the next generation of air interface protocols and network architectures are necessary. Asymmetric and bursty nature of multimedia packet data traffic along with the variability of data rates and packet sizes and complexity of quality of service (QoS) management makes conventional voice-oriented channelization and access protocols of 2G systems inefficient, though the channelization, signaling, and access protocols of 2G cellular systems were designed to efficiently support symmetric circuit switched data and voice traffic, most of the new data applications are IP based with highly asymmetric and packet-switch traffic. [6]

The third generations (3G) of radio access technologies are expected to use new physical and logical channelization schemes with enhanced media and link access control protocols. Also, to maximize the spectrum efficiency, the physical layer designs must utilize advanced coding, link adaptation, and diversity schemes as well as power and interference control mechanisms. These observations and requirements motivated extended researches in various organizations. ITU activities on IMT2000 are comprised of international standardization, including frequency spectrum and technical specifications for radio and network components, technical assistance, and studies on regulatory and policy aspects. Most 2G-CDMA (IS95/B)-based networks have migrated to cdma2000-1X technology, primarily based on the IS2000 Release 0 standard. Some operators who decided to devote separate carriers for high-speed packet data to complement their existing voice-based services have opted to use 1xEV-DO (or HRPD) carriers as an overlay to their existing 2G carriers.

The major rationale behind the deployment of a CDMA system is its potential for high spectral efficiency, that is, the capability to support significantly more mobile subscribers within a given bandwidth. The core design concept of each component, such as power control and soft handoff, focuses on realizing and enhancing such potential while at the same time maintaining acceptable quality. In addition, the modulation concept permits the offering of such desirable system attributes as dynamic capacity and voice privacy.

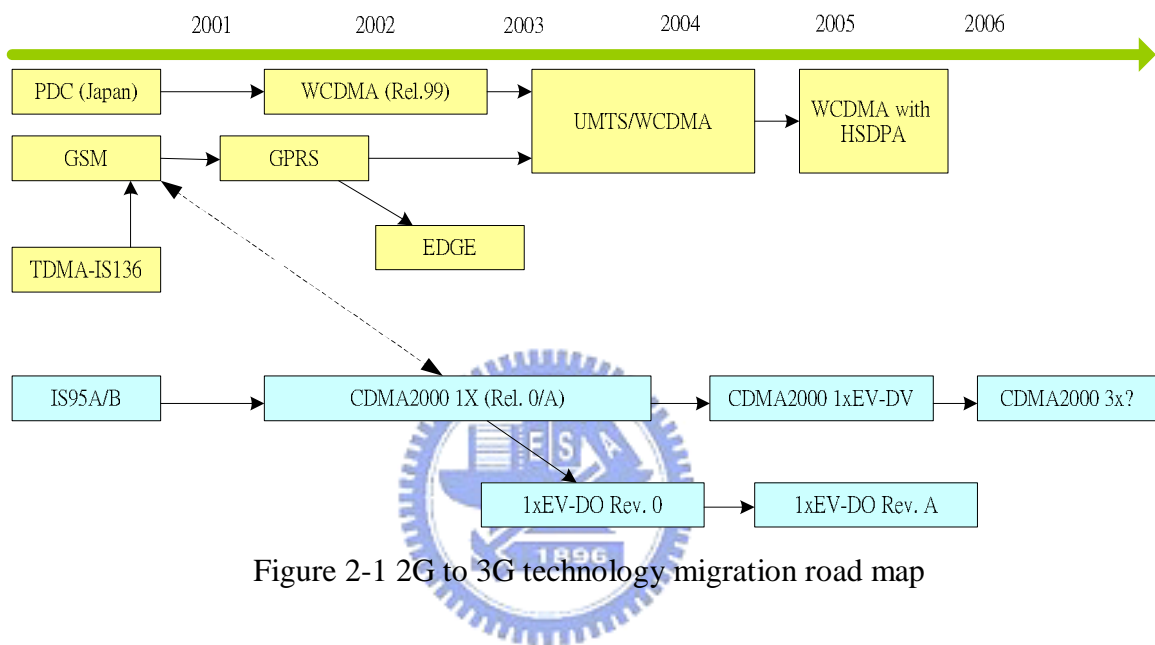


Figure 2-1 2G to 3G technology migration road map

### ***Fairness-based Bandwidth Allocation Algorithm***

From system points of view, always supplying resources to people under better air conditions can increase power efficiency and system throughput. However, individuals only care about personal service quality, thus maintaining long-term stable transmission rate, short-term delay constraints, and flexible burst transmission is essential to satisfy subscribers perceptually. For cdma2000 without QoS, the most intuitive fairness can be induced to two categories - equal rate and equal time scheduling. [4]

Equal rate tries to ensure that all users receive the same data rate. Depending on different kinds of media service, system will reserve relative sufficient power through weighting proportion to sustain corresponding SCH, irrespective of their RF conditions. Equal time uses an idle timer to record starvation of each user. The longer be idled, the urgent

bandwidth should be assigned to compensate previous throughput. This weighting factor represents how long does system serve the subscriber no resources and let him idle even though BSQ keeps on increasing.

The equal rate scheme may result in poorer sector throughput performance than the equal power scheme but is fairer from the user point of view. Equal time though can take care every user, round robin allocation guarantees neither system performance nor QoS satisfaction. In fairness-based RRM, the equal power scheme is the initial criteria, but also takes the minimum bandwidth (at least a fundamental channel is assigned if the system is not in overload) fairness, the amount of backlogged data, and Walsh code limitation into account.

If the users stay in poor RF conditions or with a small amount of data waiting in the queue, they will be assigned a low rate supplemental channel. If total requested power reaches the power limit, the SCH rate associated with the user requesting highest power will be reduced to the next lower rate. Besides, if the Walsh code resource becomes the bottleneck, the SCH rate associated with the user assigned the highest rate will be reduced to the next lower rate.



## ***Capacity***

Capacity [7] [8] considerations are fundamental to CDMA planning and operation. Here we will define capacity simply as the number of mobile subscribers that can be simultaneously supported to simplify the discussion. The forward link capacity will be analyzed below.

In each link, CDMA signals share the same spectrum. Each mobile station uses a unique code to make its signal appear as broadband interference to every other mobile station. Power control minimizes the impact of this interference by adjusting each signal level to the minimize necessary to achieve desired call quality, and applications of these principles result in the dynamics of CDMA capacity.

For forward link capacity, restrictions on base-station radiated power fundamentally determine upper limits on forward link capacity. The forward-link signal comprises message

traffic for subscribers, a sector-specific (pilot) signal used by all mobile stations and miscellaneous signals (sync, paging, etc.). The base station allocates the total power among these functions. Additional mobile stations cannot be supported when the sum of the allocations required exceeds the available transmit power.

The need for a minimum S/I at each mobile station governs the required allocations. The power allocated to other mobile stations within the cell as well as the received power from neighbor base stations contribute to interference. Use of orthogonal codes partially mitigates this interference because it allows the receiver to suppress signals intended for other mobile stations; however, multipath effects limit the extent to which this interference can be screened out. The requirement that a generous fraction of power must be allocated to the sector pilot further restricts forward-link power distributions. The sector pilot is important because all mobile stations use it in base-station acquisition and tracking. Therefore, Capacity limits are reached when the remaining power, distributed among all mobile stations, is insufficient to meet mobile station signal-to-interference ratio requirements.

### ***Soft Handoff***



CDMA provides various mechanisms to ensure a robust handoff, that is, to ensure call support when a mobile station crosses the boundary from one cell to another. The chief mechanism employed is soft handoff in which the mobile station's call is simultaneously supported by up to three sectors. This process enables the mobile station to establish contact with sectors it is likely to travel through well before it leaves its serving base station. [8]

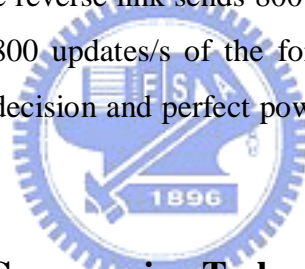
The FCH handoff procedure between cells (soft handoff) and between sectors (softer handoff) is based on the dynamic threshold defined in “cdma2000 1X TIA/EIA/IS-2000.2-A, Physical Layer Standard for cdma2000 Spread Spectrum Systems”. Soft handoff condition is determined from link-level simulation data without transmit diversity [7] [8] [9]. To avoid over preservative resources reservation that directly affect system capacity, no soft/softer handoff is considered for the SCH; only the base stations with strongest pilot power at the mobile terminal can transmit on the supplemental channel. This assumption is consistent with many of the cdma2000 1X system implementations.

## ***Power Control***

The fraction of total traffic power allocated for the mobile, which is a function of geometry, i.e., the total received power spectral density from desired base station divided by the total received power spectral density from all other base stations plus the thermal noise density, is determined from link-level simulation data without transmit diversity.

cdma2000 uses fast closed loop power control on the forward link dedicated channels with 800 updates per second. The closed loop power control compensates for medium to fast fading and for inaccuracies in open loop power control. [7] [8] Furthermore, fast forward link power control is effective for adaptation of dynamically changing interference conditions due to the activation and deactivation of high power high data rate users.

cdma2000 reverse link design also ensures the implementation of the fast power control of the forward link. The reverse link sends 800 bits per seconds of forward link power control information enabling 800 updates/s of the forward link transmitter power. To save simulation time, both handoff decision and perfect power control adjustment are made once a frame rate.



## **2.2. MPEG4 FGS Video Compression Technology**

In this section, the vision, concept and technology of MPEG-4 standard is introduced. Besides, fine granularity scalable (FGS) coding, the core functionality that makes MPEG-4 especially suitable for wireless applications, is also discussed in detail.

### **2.2.1. Video in MPEG4**

MPEG-4 [10] is an open standard, representing thousands of man-years of work shared by hundreds of companies. No one company can hope to match the technical and intellectual resources of an entire competitive market. No other technology has the potential to become as deeply developed and widely supported by multiple industries, vendors and service providers, and to be trusted by end users with their video and multimedia needs.

MPEG-4 can address the opportunities enabled by the digital revolution by easily deploying multimedia content for any platforms. MPEG-4 dramatically advances audio and video compression, enabling the distribution of content and services from low bandwidths to high-definition quality across broadcast, broadband, wireless and packaged media.

MPEG-4 provides a standardized framework for many other forms of media, including text, pictures, animation, 2D and 3D objects, which can be presented in interactive and personalized media experiences. To support the diversity of the future content market, MPEG-4 offers a variety of so-called profiles tool useful for specific applications. Users need only implement the profiles that support the functionality required. MPEG-4 is developed by the Moving Picture Experts Group (MPEG), a workgroup of the International Organization for Standardization (ISO) and the International Electro-technical Committee (IEC)— the group that designed MPEG-1, which includes MP3 digital audio and MPEG-2.

In an encoder using the MPEG-4 architecture, one or more audio-visual objects and their spatio-temporal relations are encoded separately, error-protected, multiplexed and then transmitted downstream. The transmission may use multiple channels offering various quality of service and various levels of interactivity. At the decoder, the audio-visual objects are demultiplexed, error-corrected, decompressed, combined and presented to the end user.

MPEG-4 [10] introduces several new video entities, called Video Object (VO), Video Object Layer (VOL) and Video Object Plane (VOP), and takes a different approach to video information, as compared to MPEG-1 and MPEG-2 standards, where the video information is of rectangular and fixed size and display at fixed intervals.

Video Object Planes represent instances of a given Video Object. Each VOP is described by its shape, texture/ color information, position within the image and its motion. At the encoder side, together with the VOP, composition information is sent to indicate where and when each VOP is to be displayed. Both VOP's and VO's have corresponding entities in the bitstream, which a user can access and manipulate. Thus a user may request to improve the quality of the foreground object at the cost of the background objects or vice versa. Such interactivity is very important when video objects are coded at very low-bitrate. Other examples of interactions include cut and paste operations, object zoom and freeze. The user may also be allowed to change the composition of the scene by manipulating the composition

information.

The introduction of the concept of Video Objects overcomes the limitation of existing standards such as JPEG, and MPEG and H.263 where images are represented as rectangular matrices. This type of representation and coding has proven to be very efficient for applications that do not require object level interactivity. However, pixel-based representation does not support the capability to distinguish and interact with different objects from which the picture is composed.

Object-based representation requires the compression scheme to be able to manipulate arbitrarily shaped regions. Though some scenes might be decomposed into objects during creation, automatic tools for segmentation must be developed for others. The task of the receiver is to reconstruct each of the visual objects and to compose the scene from the objects. Therefore, additional information has to be sent to the receiver specifying how to compose the scene from its objects. The System Layer is designed to support the capability to send the additional information for different objects and bind it with the objects and/ or to be able to retrieve additional information about the selected objects with a scene.

Spatial and temporal scalability and error robustness are the main functionalities supported at VOL and VOP level. To satisfy channels of varying bandwidth, receivers of different processing capability or to match different user requests, scalability is a critical feature when the same audio-visual objects are to be made. Robustness to error is also important as audio-visual communications on radio channels is a popular application of MPEG4.

An audio-visual object bitstream is scalable if at least one subset of the bit stream is sufficient for generating a useful presentation of the object. An audio-visual objects is a representation of a real or virtual object that can be manifested aurally and/ or visually. Audio-visual objects are generally hierarchical, in that they may be defined as composites of other audio-visual objects, which are called sub-objects; objects that are composites of sub-objects are called compound audio-visual objects; all other objects are called primitive audio-visual objects.



### 2.2.2. Fine Granularity Scalability (FGS) in MPEG-4 Video Standard

Fine granularity scalability (FGS) has been identified in MPEG-4 as a desired functionality, especially for streaming video applications. In the beginning, bit-plane coding of the DCT coefficients, wavelet coding of image residue, and matching pursuit coding of image residue are three proposed techniques for FGS in MPEG-4. Bit-plane coding of the DCT coefficients was finally chosen due to its comparably superior coding efficiency and implementation simplicity after core experiments. Some details of using bit-plane coding to achieve FGS are described in following subsections, including the overall FGS coding structure used in MPEG-4 is presented and a few details of FGS coding are discussed. [11]

#### *FGS Coding Structure*

The basic idea of FGS is to code a video sequence into a base layer and an enhancement layer. The base layer uses non-scalable coding to reach the lower bound of the bit-rate range. The enhancement layer is to code the difference between the original picture and the reconstructed picture using bit-plane coding of the DCT coefficients. Figure 2-2 and figure 2-3 show the FGS encoder and the decoder structures, respectively.

The bitstream of the FGS enhancement layer may be truncated into any number of bits per picture after encoding is completed. The decoder should be able to reconstruct an enhancement video from the base layer and the truncated enhancement layer bitstreams. The enhancement-layer video quality is proportional to the number of bits decoded by the decoder for each picture. The FGS decoder structure shown is the one standardized in the Amendment of MPEG-4.

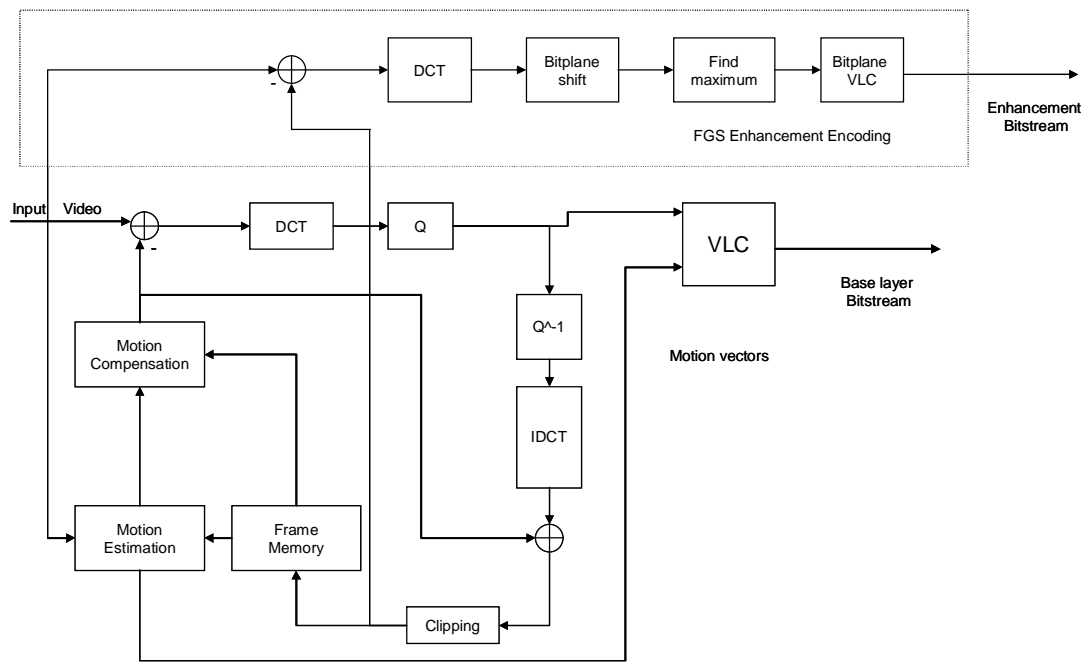


Figure 2-2 FGS encoder structure

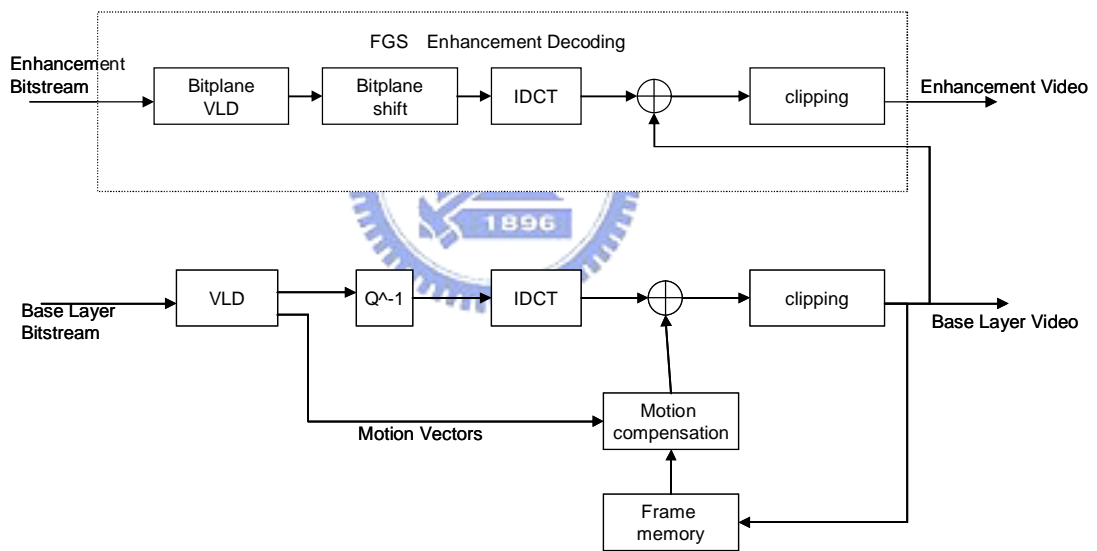


Figure 2-3 FGS decoder structure

### 2.2.3 Encoding Adaptation Mechanisms

Due to network congestion, channel variation, or handoff, the streaming server should have the ability to adapt the video clip to the new rate: if it keep encoding and sending the video packets at the fixed rule, like constant bit rate (CBR) delivery, base station queue (BSQ) in base station controller (BSC) would overflow and the buffer in the mobile device would underflow. To avoid continuous packet loss stem from BSQ overflow and frequent

rebuffering stem from client buffer underflow, the server performs the rate adaptation, and it has to be informed of the bandwidth change. This triggers system designer develop “feedback based” adaptation scheme. Except information of bandwidth change be detected or reported, client buffer fullness that can directly reflect user condition is also critical. Such feedback control concept creates another adaptation strategy, called “client based” adaptation. For CFA adaptation, media server merges both client feedback information and wireless control message, and exploits them to estimate the proper transmission rate, corresponding to some specific encoding bit rate.

On the contrast, adaptation also allows the reverse mechanism: the user can receive higher quality video frame if he enters unloaded cell, system can release more bandwidth to each existing user for the streaming connection, and the streaming server switches encoding bit rate from lower stage to higher stage after the adaptation. Both increasing and decreasing of the bandwidth is applied only to particular users, selected by congestion and admission control algorithms of BSC (e.g., the users performing handover to the cell where different bandwidth is available).

### ***Client-based Adaptation Mechanism***



Client based method [5], by means of a specific new RTSP message, consists of an event-triggered explicit adaptation request from the remote client to the media server, transparent front end wireless system. The client continuously estimates the incoming throughput, filters out the variations due to radio link jitter and monitors buffer fullness. This scenario implies a specific algorithm to be implemented in the client, and the proprietary protocol needs to be built to communicate the detected link rate from client to the server, transparently. Moreover, frequent switches of transmission bandwidth corresponding to high and low quality due to the trial and error algorithm may annoy the end user and cause additional signaling overhead.

Current network is possible to have adaptation by deputing the application layer to detect bandwidth variation and to perform service adaptation. In this respect, the application has the aim to detect the bandwidth variation and decide to adapt the service by means of application signaling. This method uses implicit knowledge about underlying transport layer

for detection and adaptation, and means a degraded QoS experience, mainly due to the long latency. At the application layer only the overall end-to-end information is available, thus, the application cannot distinguish between wireless and wired link conditions.

The main drawback is the difficulty to detect the increase of available transport resources since only client buffer fullness is actually known. Hence, to obtain reliable switch trigger, a complex mix of network probes or “trial and error” algorithms may be needed; this solution may be costly and complicated with the difficulty in adequately correlating the data to timely diagnose the need for an adaptation. Figure 2-4 shows the feedback message flow through protocols built in wireless and wireline network facilities from client portable device to remote media server. (Here UMTS but cdma2000 is used since the real implementation is done on UMTS protocol stack, and the same concept certainly can be applied to cdma2000.)

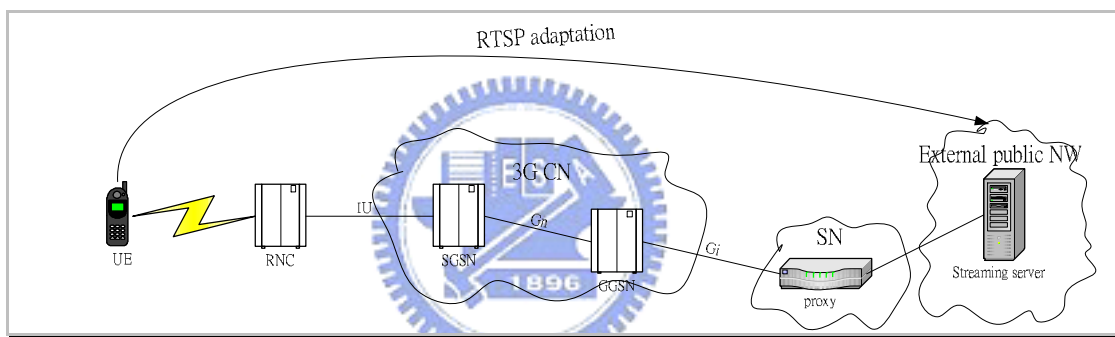


Figure 2-4 client-based adaptation mechanism

### ***Radio Network Feedback (RNF) Mechanism***

Immediate and explicit notification of the relatively guaranteed radio link parameters, such as bandwidth or power budget, that cdma2000 system can supply to the remote application server/proxy for video session is the main concept of the RNF adaptation. [5] Periodically or when the conditions over the RF condition changes, RRM algorithms located in the BSC exploits information, including uplink interference, available downlink power budget, Walsh codes, etc, to detect the modification and to relocate the resources among different users.

Besides, the application is also notified about this modification synchronously. This enables the media server to modify the service transmission rate and encoding bit rate

according to the assigned bandwidth, thus enhancing the service quality. Using RNF, the adaptation is precise and timely, as it is based on an explicit transport bit rate notification from the radio network to the server. On the other hand, to keep accuracy of RRM, it's intuitive to set stringent rate-switch criteria, like very long term observation of channel environment, to avoid insufficient resources in wireless bottleneck. This makes adaptation scenario conservative and degrade system resource utilization rate. Figure 2-5 shows the feedback message flow through protocols built in wireless and wireline network facilities from front end RRM controller to media server.

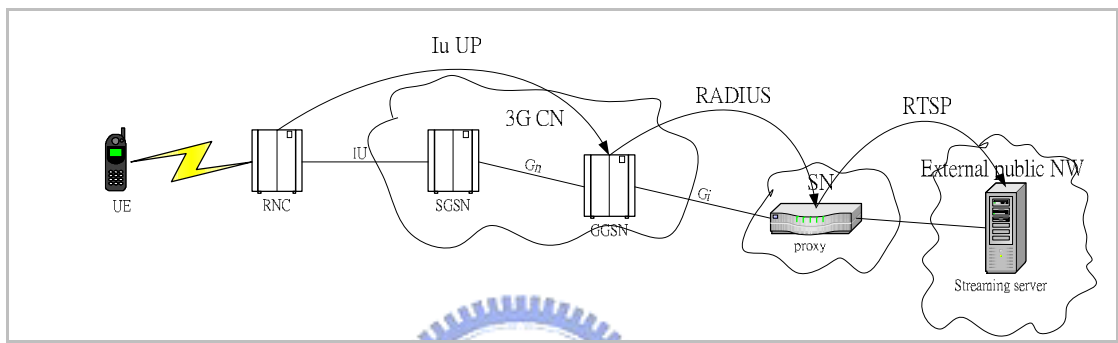
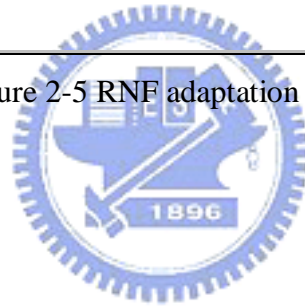


Figure 2-5 RNF adaptation mechanism



# Client Feedback Assisted Radio Resource Management and Encoding Adaptation Mechanism

The objective of the proposed solution is to maximize radio resource utilization rate while maintaining quality-of-service (QoS) of video streaming. The proposed solution, called integrated client feedback assisted (CFA) solution, consists of radio resource management (RRM) in a wireless system and encoding adaptation mechanism in a multimedia server. The CFA RRM could dynamically relocate limited radio resources to improve resource efficiency and to avoid client buffer underflow. The CFA encoding adaptation mechanism gives the flexibility to match varying radio frequency (RF) condition and system loading. The challenge here is how to produce the critical feedback information that can coordinate RRM and encoding adaptation mechanisms to make proper decisions. In this chapter, we first introduce the necessities of an integrated solution, and then the proposed CFA architecture is discussed. Because the feedback information and the CFA RRM algorithm are tightly related, they are described in the same section 3.2. The CFA encoding adaptation mechanism is discussed in section 3.3. Finally, since the system loading (capacity) directly affects the performance of the proposed CFA architecture, a cooperated call admission control is discussed in section 3.4.

### 3.1. Integrated CFA Solution

Video streaming is a high bandwidth, delay sensitive application. The QoS requirement consists of stable resolution, smooth adaptation, and continuous playout. Stable resolution and smooth adaptation are achieved if the coordination of RRM and the encoding adaptation mechanism can schedule sufficient resources to transmit video packets. The main difficulty to maintain continuous playout is the uncertain delay. The uncertainty could cause by RF variation, network queuing delay, insufficient radio resource, and improper adaptation of encoded bandwidth. Analyzing the delay factors, RF variation is dominated by surrounding environment, and network queuing delay can't be controlled if video source exists in public network. However, limited radio resource may be sufficient if RRM schedules resources smartly, and correct information can help encoder adapting its compression rate adequately.

First two factors are inevitable, but RRM and the encoding adaptation mechanism are possible to be enhanced.

The integration of different network components can effectively coordinate unified operations to optimize system performance. Without centralized coordination, front end RRM might not be able to schedule the resource assignment by utilizing the scalability of FGS video coding technology, and back end video adaptation mechanism might adopt conservative policy due to the lack of certain information of network condition. Hence the integrated CFA solution is proposed to combine both the front end wireless RRM and the back end encoding adaptation mechanism.

To design any effective control mechanisms, the most critical thing is to gather all kinds of information which are precise, instant, and related to this control. Those kinds of information could be retrieved from self probing, detection of current status, estimation of future condition, continuous reports by other components, and temporal trigger due to special events. For example, typical RRM cooperates with forward link fast power control (self probing and continuous reports), congestion avoidance (detection of current status), and channel estimation (estimation of future condition). Exploiting the information supplied by cooperated mechanisms, RRM can speculate each user's RF condition and system loading. RRM then decides the optimized allocation schedule according to its design criterion. Typical RRM mainly considers how to maximize resource efficiency, but the popularity of multimedia applications lets QoS become another important issue. To take QoS into account, the proposed solution needs to grasp the information which can represent the heterogeneous characteristics and technologies among multimedia applications. Such information is related to higher layer protocol (TCP/IP, UDP), specific function units (video/audio codec), or general hardware (receiver buffer), which exist only in end client devices. Hence exploiting feedback information from client devices can help the proposed solution considering not only resource efficiency but also QoS satisfaction.

General video codecs have buffers to alleviate all possible delay factors. Buffer is a specific memory space that can stack encoded video packets or decoded video frames. Each buffer retrieves and outputs its backlog regularly, depending on its functionality. Buffer can absorb burst data and can smooth out temporary idle due to uncertain delay. However, it can support only passive protection, not active prevention. If the information about buffer, such as

buffer fullness, is known by the proposed solution, the solution can transfer it as an emergency metric, which implies the risk of buffer underflow. The metric can affect RRM schedule, and can help the encoding adaptation mechanism judging the adaptation necessity. Hence we choose buffer status as the desired feedback information.

These observations encourage us to design an integrated control mechanism, which adopts feedback information as one of the most important reference, and we call it integrated client feedback assisted (CFA) solution. Using the feedback information as a core parameter, system can precisely know the emergency of resource starvation, and the satisfaction of current service quality. Resource starvation implies the priority a client should be served in schedule list, and the satisfaction implies the whether video compression rate should be adapted. Figure 3-1 shows the CFA architecture. The solution integrates both RRM and the encoding adaptation mechanism into base station controller (BSC). It continues monitoring every client by feedback information, relocates resources periodically, and sends adaptation command to remote multimedia server. In the multimedia server, the server controller accepts adaptation command, the streamer accepts adaptation parameter from server controller and segments demanded bitstream into video packets, and the packet buffer stacks the encoded video packets and output them based on traffic shaping rule. In the portable device, three buffers-network resynchronization (NR) buffer, processing resynchronization (PR) buffer, and output buffer are responsible for saving arrived packets, synchronized packets, and decoded frames, respectively. The detailed functionalities of these buffers will be further explained in the next section.

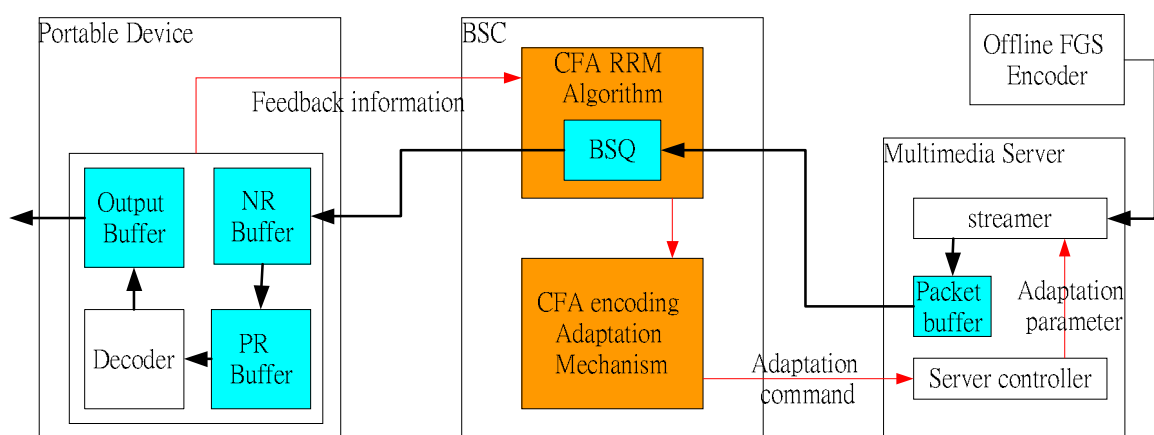


Figure 3-1 CFA Architecture



### 3.2. CFA RRM Algorithm

The objective of the proposed CFA RRM algorithm is to simultaneously improve system performance and to satisfy QoS requirement of real time video service. To improve system performance, RRM must schedule users in good RF condition first. To satisfy QoS requirement, RRM must give higher priorities to urgent users. The proposed CFA RRM control is depicted in figure 3-2. This algorithm can be separated into three steps—priority assignment, schedule, and resource allocation. Priority assignment treats estimated RF condition as efficiency notification, and treats feedback information as QoS notification. Merging these notifications, a proper priority can be calculated. The scheduler sorts active video users based on resulting priorities and decides the order video users requesting resource budgets. Resource estimation sends bandwidth requests to resource allocation according to the amount of backlog video packets stacked in base station queue. Since power budget is usually the bottleneck among various resources, resource allocation transfers original bandwidth requests to power requests by considering the associated RF condition and then allocates resources in the priority order. Hence, the better RF condition user with, the urgent possibility QoS be deteriorated, the more superiority user can fight for resources in CFA RRM algorithm.

To make the resulting priority representative, we should first declare the precise definitions of estimated RF condition and feedback information. Through functional analysis of buffer structures, QoS-related parameters can be derived and are integrated as feedback information. Besides, to further distinguish the level of risk that users can suffer, several thresholds are defined by the system and will be set in client buffer. Finally, the design concept and the practical algorithm of CFA RRM will be discussed.

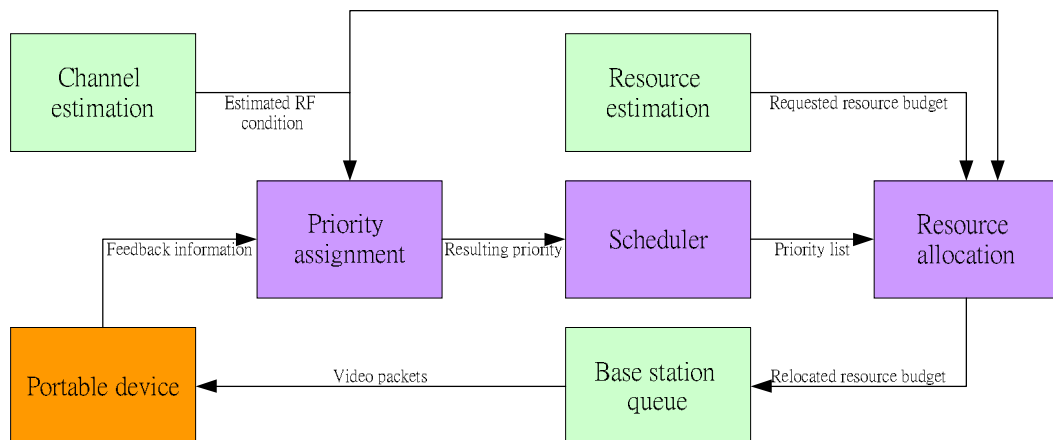


Figure 3-2 CFA RRM control flow

### ***Estimated RF Condition***

In cdma2000 1x physical layer, all channels in the same carrier (spectrum) with different Walsh code has roughly the same interference, fading, shadowing, and multipath. This implies if the RF condition of basic fundamental channel (FCH) has been estimated, other channels, like supplemental channels (SCH), can use the same RF condition estimation. Exploiting self probing, power control, and other prediction mechanism, channel estimation can predict how much power budget will a FCH needs, so the FCH power budget, called *FCH\_power*, is an important reference and is defined as estimated RF condition.

### ***Extracted feedback information***

To find the key parameters which are highly QoS relative, we first analyze the functionalities and operations of the client buffer. In our design, three-stage buffer structure, including network resynchronization (NR) buffer, processing resynchronization (PR) buffer, and output buffer, manipulates the data flow, likes video packet reception, video frame decoding, and video playout. Figure 3-3 shows this buffer structure with an internal data flow.

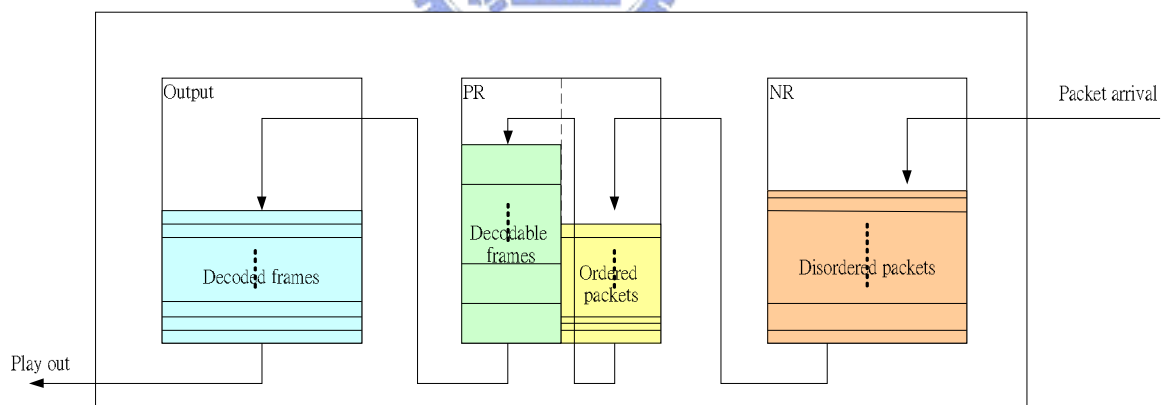


Figure 3-3 three stage structures of client buffer

The network architecture adopted is asynchronous packet switching public network because most video servers and content providers prefer to release their video source on the internet. The NR model mainly conceals disorder/delay jitter between video-to-video and video-to-audio packets due to the routing/queuing delay experienced in routers of wire line network and BSQ of the last mile wireless network. The goal of this model is to assign proper amount of packets to processing resynchronization model at proper timing, thus the decoding

frames can be played at a constant frame rate which have predefined by source. The physical meaning of the NR buffer fullness is the sensitivity reflecting how much bandwidth system assigned to this client, and this can be regarded as the major component of a *system related QoS* metric.

The task of the PR model focuses on monitoring whether video packets belonging to the same video frame are completely received and moves them to the decoder if necessary. The PR model is used to cancel the delay variation in generating the compressed video data units, thus the decoded frames are able to be played in predefined video frame rate constantly. The PR model contains two kinds of backlogged data—decodable frames and ordered packets. The video packets which are completely gathered for decoding can be treated as decodable video frames, and the remaining are ordered video packets.

The main task of output buffer is to save decoded video frame from the decoder and to supply video player for smooth playout. Since the most unacceptable event during a video session is frequent and unpredictable suspension, output buffer preloads some video frames to avoid accidental suspension. Thus suspension will happen only when output buffer becomes empty (underflow), and the physical meaning of output buffer fullness dominates a *user perceived QoS*.

To satisfy the QoS requirement, CFA RRM regards *the system related QoS metric* ( $QoS_{sys}$ ) and *the user perceived QoS metric* ( $QoS_{user}$ ) as feedback information.  $QoS_{sys}$  is defined as the sum of disordered packet size and ordered packet size;  $QoS_{user}$  is defined as the total number of decodable frames and decoded frames. In the beginning of a video connection, each user preloads the basic quality of video data as its backlog until total buffer fullness achieves predefined threshold. Excluding the extreme delay that rarely happens, the preload video packets are resynchronized and then are decoded as video frames, stacking as  $QoS_{user}$ . Only a little packets that are still disordered and packets that currently arrive are stacked as  $QoS_{sys}$ . Thus  $QoS_{user}$  is normally larger than  $QoS_{sys}$ . If radio resources are sufficient and video encoding bandwidth is adequate under acceptable network congestion, both  $QoS_{sys}$  and  $QoS_{user}$  are roughly constants.

Reversely analyzing the variation of  $QoS_{sys}$  and  $QoS_{user}$  can help designing the schedule policy of the CFA RRM. The stability of  $QoS_{sys}$  and  $QoS_{user}$  represents steady

resource budgets and reasonable network congestion. The  $QoS_{sys}$  variation of the multiple of two implies radio resource relocation (up-switch or down-switch the assigned bandwidth). The cumulative increment of  $QoS_{sys}$  means that some disordered video packets can't be resynchronized yet, and the resynchronization of these packets results in acute decrement of  $QoS_{sys}$  subsequently. The continuous degradation of  $QoS_{user}$  may reflect a severe disorder event as what the cumulative increment of  $QoS_{sys}$  means or the unfair RRM policy that schedules insufficient resources to the client. On the contrast, burst recovery and gradual recovery of  $QoS_{user}$  imply successful resynchronization and resource compensation, respectively. The CFA RRM grasps the demand of every user through the analysis. Additionally, if the risk of QoS degradation among video users can be clearly distinguished, the CFA RRM can utilize  $QoS_{sys}$  and  $QoS_{user}$  to schedule user priorities, corresponding to the degree of risk.

To clearly distinguish the risk of QoS degradation among users, user perceived buffer fullness ( $QoS_{user}$ ) is a representative metric because it directly reflects the possibility of video suspension (buffer underflow). The CFA RRM grades video users into three levels—unsatisfied, acceptable, and satisfied by setting three thresholds—*preload threshold* ( $T_{preload}$ ), *overflow threshold* ( $T_{over}$ ), and *underflow threshold* ( $T_{under}$ ) as shown in figure 3-4.  $T_{preload}$  is the initial threshold that preload operation should achieve before starting playout.  $T_{over}$  and  $T_{under}$  are purely used to separate users into three categories.  $T_{rebuffer}$  is the threshold rebuffer operation should achieve before continuous playout. Besides,  $QoS_{sys}$  and  $QoS_{user}$  are normalized to *rebuffer threshold* ( $T_{rebuffer}$ ) because we want to apply this CFA RRM methodology to any devices with different video codec, buffer size and preload threshold. The normalized  $QoS_{sys}$  is called *stack ratio* ( $r_{stk}$ ) and the normalized  $QoS_{user}$  is called *decoded ratio* ( $r_{dec}$ ). All thresholds can be self defined by the system depending on the client buffer size, the video codec, and the delay tolerance. Consequently, the level of risk (unsatisfied, acceptable, or satisfied) decides how the CFA RRM calculates user priorities exploiting  $QoS_{sys}$  and  $QoS_{user}$ .

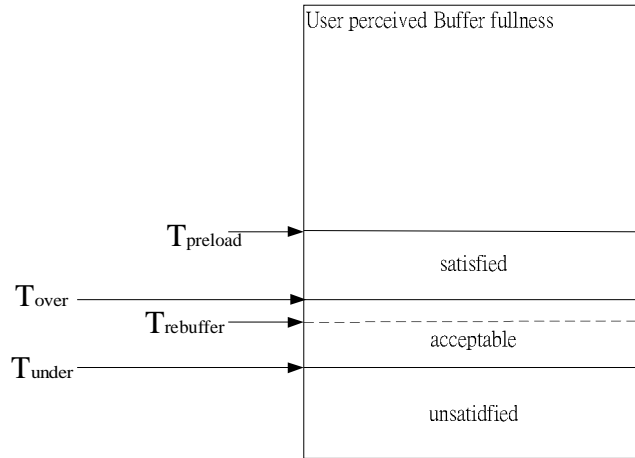


Figure 3-4 thresholds of output buffer fullness

### Algorithm Implementation

The principle of the CFA RRM is “allocating higher priority to users in good RF condition or users with nearly empty buffer”. Since estimated RF condition has been represented as  $FCH\_power$ , if  $QoS_{sys}$  and  $QoS_{user}$  are further integrated as an absolute weighting factor “ $W$ ”, the principle can be translated into practical implementation—sorting " $\frac{W_i}{FCH\_power_i} \quad \forall i \in \{video\_user\}$ ", and the resulting priority is defined as  $\frac{W}{FCH\_power}$ .

The CFA RRM algorithm is separated into three steps, as shown in figure 3-6. In step1, the weighting factor “ $W$ ” is calculated. Since what the CFA RRM tries to improve is the user perceived quality,  $r_{dec}$  is the most critical parameter and is chosen as the satisfaction index. If  $r_{dec}$  is above  $T_{over}/T_{rebuffer}$  (case1), the sufficient backlog implies low emergency and high tolerance, and the CFA RRM needs to assign a low “ $W$ ” that is stable but can still be distinguished. “ $W$ ” is thus calculated as the reversal of the sum of  $r_{dec}$  and  $r_{stk}$  because large  $r_{dec}$  maintains the stability and small  $r_{stk}$  enables “ $W$ ” to distinguish the light desire among users in this category. Consequently, considering the entire buffer fullness is more meaningful than emphasizing individual emergency under satisfied condition.

If  $r_{dec}$  drops between  $T_{over}/T_{rebuffer}$  and  $T_{under}/T_{rebuffer}$  (case2), medium emergency and fewer backlogs may be able to alleviate short term idle, but subsequent RF deterioration or system overload will result in buffer underflow. To sensitively detect any possible underflow event and to instantly raise the priority, “ $W$ ” is thus calculated as the sum of two independent

factors—the reversal of  $r_{dec}$  and the reversal of  $r_{stk}$ . Either gradually deterioration of  $r_{dec}$  caused by severe packet disorder or acutely degradation of  $r_{stk}$  caused by insufficient radio resources can trigger the CFA RRM assigning apparently large “ $W$ ” to the client in the acceptable condition.

If  $r_{dec}$  falls below  $T_{under}/T_{rebuffer}$  (case3), high emergency triggers weighting factor exponentially increasing in order to declare its high risk suffering buffer underflow, and the user must be served before very short deadline. “ $W$ ” is thus calculated as the square of the sum of two independent factors—the reversal of  $r_{dec}$  and the reversal of  $r_{stk}$ . The design concept is basically the same as case2, and the major difference is the square. This is because no more backlog can postpone the buffer underflow, and only the extremely growth of “ $W$ ” can guarantee the highest priority in the next schedule list.

In step2, the CFA RRM produces a schedule list that records the order of resource relocation. The schedule list comes from sorting the priorities all video users have, and  $W$  divided by  $FCH\_power$  equals the priority. In step3, the CFA RRM relocates radio resources to each user following the schedule list. Against traditional equal power or equal rate criterion, the CFA RRM admits a user to request arbitrary amount of resources unless the system runs out of the available resource budgets.

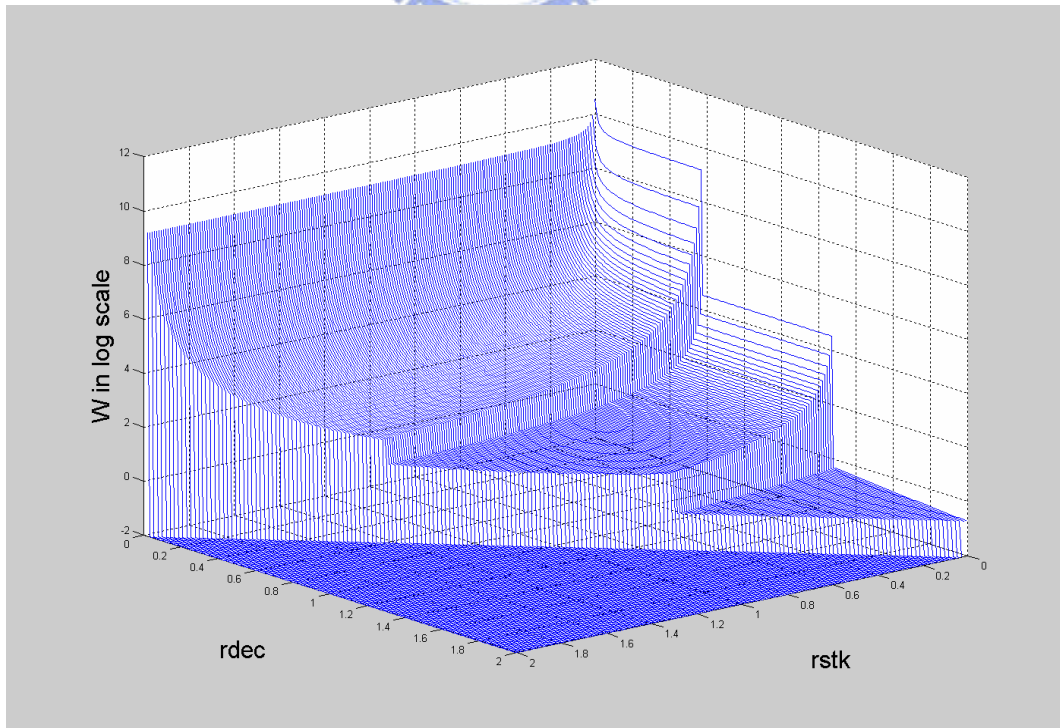


Figure 3-5 value plane of weighting factor  $W$  in log scale

The CFA RRM Algorithm

Step1. calculate weighting factor "W"

$$r_{stk} = \frac{QoS_{sys}}{T_{rebuffer}} \quad r_{dec} = \frac{QoS_{user}}{T_{rebuffer}}$$

$$W = [(r_{dec})^m + (r_{stk})^n]^p \quad \text{video\_frame\_rate} = 10 \text{ frames per second (fps)}$$

case1(satisfied):

$$\text{if } r_{dec} \geq \frac{T_{over}}{T_{rebuffer}} \Rightarrow \text{set } n = 1, m = 1, p = -1$$

$$\Rightarrow W = \left( \frac{1}{r_{dec} + r_{stk}} \right)$$

case2(acceptable):

$$\text{if } \frac{T_{over}}{T_{rebuffer}} > r_{dec} \geq \frac{T_{under}}{T_{rebuffer}} \Rightarrow \text{set } n = -1, m = -1, p = 1$$

$$\Rightarrow W = \frac{1}{r_{dec}} + \frac{1}{r_{stk}}$$

case3(unsatisfied):

$$\text{if } \frac{T_{under}}{T_{rebuffer}} > r_{dec} \Rightarrow \text{set } n = -1, m = -1, p = 2$$

$$\Rightarrow W = \left( \frac{1}{r_{dec}} + \frac{1}{r_{stk}} \right)^2$$

Step2. produce a schedule list that records the order of resource relocation

$$\text{Priority}_i = \frac{W_i}{FCH\_power_i} \quad \forall i \in \{\text{video\_user}\}$$

$$\text{Schedule\_list} = \text{sort}(\text{Priority}_i) \quad \forall i \in \{\text{video\_user}\}$$

Step3. relocate resources to each user following the schedule list

*available\_resource* = *total\_resource*

for *i* = 1:video\_user

if *request\_resource*(*Schedule\_list*(*i*)) < *available\_resource*

*assigned\_resource*(*Schedule\_list*(*i*)) = *request\_resource*(*Schedule\_list*(*i*))

*available\_resource* = *available\_resource* - *assigned\_resource*(*Schedule\_list*(*i*))

elseif *request\_resource*(*Schedule\_list*(*i*)) > *available\_resource* & *available\_resource* > 0

*assigned\_resource*(*Schedule\_list*(*i*)) = *available\_resource*

*available\_resource* = 0

esle

*assigned\_resource*(*Schedule\_list*(*i*)) = 0

Figure 3-6 CFA RRM algorithm

### 3.3. CFA Encoding Adaptation Mechanism

The encoding adaptation mechanism enables a video streaming server to adapt its source bit rate to cdma2000 radio link, whose bandwidth may vary in time due to system congestion or handoff. The benefits of the encoding adaptation mechanism are able to enhance system throughput by upgrading video resolution or to improve system capacity by supplying basic visual quality. Current adaptation mechanisms, such as the client based adaptation and the radio network feedback (RNF) adaptation, have their own superiorities and defects. Thus the CFA encoding adaptation mechanism is proposed to keep the superiorities and to avoid the defects from current mechanisms.

Client based adaptation mechanism consists of an event-triggered adaptation request from the client to the remote media server, and the front end wireless system is transparent. To send an adequate adaptation request, the client continuously estimates the incoming throughput, filters out the variations due to radio link jitter and monitors buffer fullness. This mechanism implies a specific algorithm to be implemented in the client, and the proprietary protocol needs to be built to communicate the detected link rate from the client to the server transparently. Frequent adaptations of transmission bandwidth due to the trial and error algorithm may not only cause additional signaling overhead but also annoy the end user because of the unstable visual quality. Without a unified controller to examine the necessity of each request, clients themselves would arbitrarily adapt required bandwidth. This results in buffer underflow due to system overload and throughput degradation due to system idle.

Immediate feedback notification of the guaranteed radio link parameters is the main concept of RNF mechanism. Parameters such as the available bandwidth, the power budget, and the estimated RF condition can help the remote application server/proxy making stringent adaptation decision. The main advantage is also the major disadvantage of the RNF mechanism because the guarantee of sufficient radio resources drives conservative up-switch policy and sensitive down-switch policy. These policies sacrifice the chance to utilize good RF condition and to fight for more resource budgets. Thus performance of such mechanism is stable in visual quality but low utilization rate in system performance.

The lack of critical information is the reason why typical mechanisms can't look after



both sides—user perceived visual quality and system performance. Without precise information about system available resources, the client based adaptation mechanism using trial and error algorithm may fit the maximum resource utilization rate but guarantee no QoS. On the contrast, knowing nothing about the client, the lowest QoS is guaranteed but system performance is sacrificed. This is because the RNF adaptation mechanism tries to absorb any accidental deficiency and to control all possible risks.

The CFA adaptation mechanism exploits both system and client information as its decision parameters. These parameters are categorized into historical record and current status. The historical record needs to memorize recent schedules and allocations, and the current status needs to reflect client emergency and system congestion. Monitoring the recent schedules and the emergency index can the CFA adaptation mechanism grasp user perceived quality. Analyzing the recent allocations and the congestion index can the CFA adaptation mechanism estimate the system loading and the available resource budgets. These knowledge and estimations enable the CFA adaptation mechanism to precisely speculate adaptation necessity. So the CFA adaptation mechanism can improve system performance by permitting up-switch request instantly and can alleviate uncertainty by monitoring available resources.

The CFA adaptation mechanism defines two historical records— $\overline{SCH}$  and  $\overline{W}$ .  $\overline{SCH}$  is the estimated available bandwidth depending on recently assigned bandwidth, and it is the resultant affected by visual quality, system loading, and RF condition. Using  $\overline{SCH}$ , the CFA adaptation mechanism can predict the risk of additional delay if the video stream with higher visual quality is sent.  $\overline{W}$  is the mean of weighting factor “W” during contiguous adaptation and indicates the average starvation to radio resources. The CFA adaptation mechanism can predict whether a user will be in high priority in next second by  $\overline{W}$ . The value of  $\overline{W}$ , which also distributes into three separating groups since it comes from  $W$ , indicates the corresponding level of risk to up-switch the encoding bit rate, and indicates the corresponding emergency to down-switch the encoding bit rate.

The CFA adaptation utilizes five timing indices— $\hat{t}_{trans}$ ,  $\hat{t}_{play}$ ,  $\hat{t}_{stk}$ ,  $t_{preload}$ , and  $t_{dec}$  to grasp current status, including resource emergency and system congestion.  $\hat{t}_{trans}$  is the *predicted transmit time*, estimating how long the system needs to completely transmit backlog

in BSQ.  $\hat{t}_{play}$  is the *predicted playout time*, estimating how long the queued video packets in BSQ can be played on the client after being received and decoded.  $\hat{t}_{stk}$  is the *predicted stack time*, estimating how long the packets not yet be decoded can be played. Without upper layer decapsulation, demultiplexing, and decoding, the precise encoding bit rate of each frame is not known, hence the CFA adaptation mechanism can only use current or average encoding bit rate to speculate.  $t_{preload}$  is the same as preload time  $T_{preload}$  and is the precise index.  $t_{dec}$  is the *decoded time*, representing how long the output buffer can sustain continuous playout until buffer underflow. It's also a precise index because video frame rate (10 frames per second in our study) is predefined, and the number of decoded frame is actually known. These CFA adaptation parameters are calculated as figure 3-7 shown.

The design concept of the CFA adaptation mechanism is “judging whether the estimated available bandwidth is timely sufficient to avoid the client buffer underflow”. Figure 3-7 and figure 3-8 shows the algorithm of the CFA adaptation mechanism. To declare how the encoding bit rate is adapted (up-switch, no-switch, or down-switch), the CFA adaptation mechanism first assigns an *authority parameter*  $q \in \{-1, 0, 1\}$  to each video session. “ $q$ ” releases specific authority for each video session to revise its encoding bit rate depending on the available resources. The smaller  $\bar{W}$ , the more backlogs in the client buffer. This implies the confidence to suffer higher risk, so the timing constraint is relatively relaxed. On the contrast, high  $\bar{W}$  implies no more risk should be suffered, and the crucial up-switch condition is set. The client who stays in good RF condition and has the authority parameter “1” can enhance (up-switch) the encoding bit rate first. The client who gets authority parameter “0” may not enhance (up-switch) the encoding bit rate unless we can make sure that the system has sufficient resources and the capacity is not congested. The client who gets authority parameter “-1” will degrade (down-switch) the encoding bit rate to avoid the risk of buffer underflow due to channel variation, sudden overload, and etc.

To improve client visual quality or system throughput, the media server up-switches the encoding bit rate if the sufficient backlog in the client buffer and the available resources can guarantee no buffer underflow. The media server down-switches the encoding bit rate if the limited radio resources can't guarantee a timely transmission following current encoding bit rate. In this algorithm, two kinds of threshold are used— $t_{preload} - t_{dec} - \hat{t}_{stk}$  and  $\hat{t}_{play}$ . They have analogous physical meaning but are considered from different point of view. Since the

major risk of up-switching the encoding bit rate stems from the buffer underflow, estimating the required transmission time by inspecting the buffer fullness through  $t_{preload} - t_{dec} - \hat{t}_{stk}$  is more meaningful and correct. On the other hand, the only reason a video session needs to down-switch its encoding bit rate is the lack of the radio resources. Thus estimating the required transmission time by inspecting the queued video packets ( $\hat{t}_{play}$ ) in BSQ is the most accurate. Finally, combining the authority parameter “ $q$ ” and the available resources, the system can decide whether the adaptation should be executed.

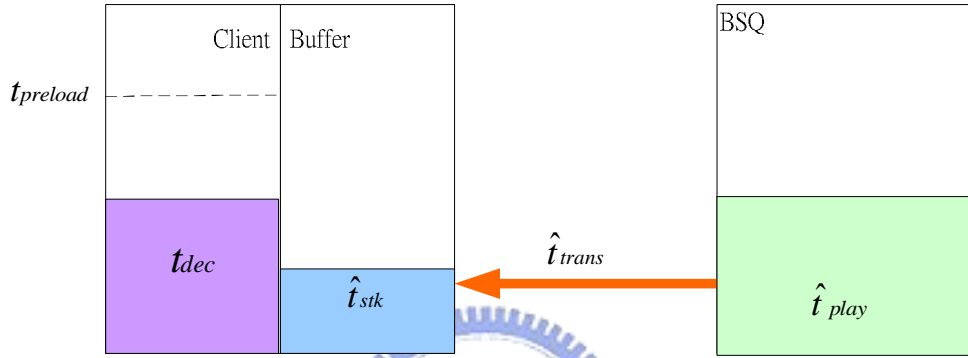


Figure 3-7 representation of the timing indices

$$\bar{w} = \frac{\sum W_i}{video\_frame\_rate}$$

$$\overline{BW} = \frac{\sum BW_i}{video\_frame\_rate}$$

$$\hat{t}_{trans} = \frac{BSQ}{\overline{BW}}$$

$$\hat{t}_{play} = \frac{BSQ}{encoding\_bitrate_{old}}$$

$$\hat{t}_{stk} = r_{stk} \times T_{rebuffer}$$

$$t_{preload} = T_{preload}$$

$$t_{dec} = r_{dec} \times T_{rebuffer}$$

case1 : if  $\bar{w} < \frac{T_{rebuffer}}{T_{over}}$

$$\Rightarrow \begin{cases} \text{if } \hat{t}_{trans} < (t_{preload} - t_{dec} - \hat{t}_{stk}) \times 2 & \Rightarrow q = 1 \\ \text{else} & \Rightarrow q = 0 \end{cases}$$

case2 : if  $\frac{T_{rebuffer}}{T_{over}} \leq \bar{w} < \frac{T_{rebuffer}}{T_{under}} + T_{rebuffer} \times frame\_rate$

$$\Rightarrow \begin{cases} \text{if } \hat{t}_{trans} < (t_{preload} - t_{dec} - \hat{t}_{stk}) \times 1 & \Rightarrow q = 1 \\ \text{elseif } \hat{t}_{trans} > \hat{t}_{play} \times 2 & \Rightarrow q = -1 \\ \text{else} & \Rightarrow q = 0 \end{cases}$$

case3 : if  $\frac{T_{rebuffer}}{T_{under}} + T_{rebuffer} \times frame\_rate \leq \bar{w}$

$$\Rightarrow \begin{cases} \text{if } \hat{t}_{trans} < \hat{t}_{play} & \Rightarrow q = 0 \\ \text{else} & \Rightarrow q = -1 \end{cases}$$

Figure 3-8 CFA adaptation algorithm

### 3.4. Call Admission Control

In traditional algorithm [4], if the current total traffic power consumed by active fundamental channels (FCHs) plus the power needed to support the incoming voice user exceeds the maximum power budget assigned for voice traffic, a voice call will be blocked to enter the system. If the current total traffic power consumed by active FCHs plus the FCH power that would be needed to support data users with non-empty queues but currently are not assigned data bursts plus the power needed to support an FCH for the incoming data user exceeds the upper power limit, a data user will not be assigned an FCH. Apparently, the data connection admission is more crucial than the voice call admission under power overload condition because voice service always has the highest priority. In practical implementation, once a data user is denied a FCH assignment to prevent system overload, this user will continuously retry to request resources through competition until success.

However, such admission control evidently sacrifices interests of data traffic and leaves voice traffic the highest priority, even needs to deprive the existing data users of the reserved resources. The QoS based call admission control, depending on the predefined QoS criteria, should positively reserve resources for existing users, and should release tolerable flexibilities to improve whole system performance. Table 1 shows the characteristics of three mainstream media services and is regarded as predefined QoS criteria. In QoS based call admission control flow, as shown in figure 3-9, QoS criteria, BW allocation, resource reservation, and resource estimation are four major components to produce three key factors—required resources, preservative resources, and reserved resources. These factors are the basis to make final admission decision. QoS criteria are stringent entry barriers that pessimistically estimate the required resources for a new incoming user or a handoff user. To protect the interests of existing users, system must reserve sufficient amount of resources. Thus bandwidth allocation is used to estimate these resource budgets. Resource reservation tries to preserve reasonable amount of resources for the predicted handoff traffic. The preservative resources are critical to control the risk of call dropping, though it may limit the system capacity and the resource efficiency. Resource estimation integrates reserved resources and preservative resources, and announces the available resources by subtracting them from the total resource budgets. Only when the required resources are fewer than available resources, this user can be served.

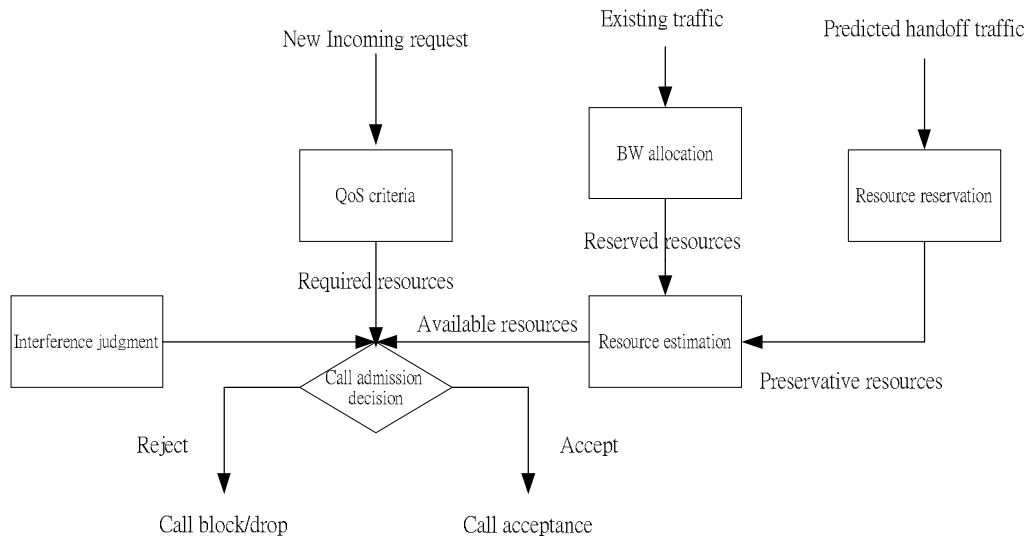


Figure 3-9 call admission control flow

In the real implementation, the call admission control algorithm, as shown in figure 3-10, tries to achieve two goals—guaranteeing sufficient resources for existing users and saving as more resources to incoming users as possible. Depending on the QoS (transmission rate) guarantee that the system have predefined, the corresponding amount of resource (power) must be reserved. Since the basic quality of MPEG4 scalable video streaming needs roughly 30kbps bandwidth, this algorithm treats the estimated power budget “ $P_{VQ,existing}$ ” as a reference, and the system designer can modify the QoS guarantee by resetting the parameter “ $r_{reserved}$ ”. The higher “ $r_{reserved}$ ”, the better visual quality the existing users enjoy, but the more difficult the incoming users be accepted. Hence the available power budget for incoming users can be calculated as the total power budget ( $P_{total}$ ) minus the sum of the overhead of control channels ( $P_{overhead}$ ), the preservative power ( $P_{handoff}$ ), and the reserved power ( $P_{existing}$ ). If the basic power requirement of an incoming user is less than the available power, system accepts this call, or it will be blocked.

Additionally, the video traffic arrival interval is modeled as a log normal distribution, with a density parameter “*video occupancy*”. “*video occupancy*” decides the average busy interval caused by video session in an hour. Thus “*video occupancy*” is a kind of data Erlang and can be modified by “ $k_{video}$ ” in our simulation. Changing “ $r_{reserved}$ ” and “ $k_{video}$ ” can we see the impact to the system performance under various QoS guarantee and traffic density.

$$P_{QoS} = \sum P_{VQ, existing} \times r_{reserved}$$

$$P_{available} = P_{total} - P_{overhead} - P_{handoff} - P_{QoS}$$

if  $P_{VQ, incomming} \leq P_{available}$

*call accept*

else

*call block*

$P_{QoS}$  : power that should be reserved for existing users to guarantee QoS

$P_{VQ}$  : power that supports guaranteed visual quality (specified transmission rate)

$r_{reserved}$  : reserved ratio that is decided corresponding to the QoS criteria, the characteristics of each application and the type of mixed traffic

$P_{available}$  : power that is available to serve new incomming users

$P_{total}$  : power that a base station can maximally supply as the total budget

$P_{overhead}$  : power that is assigned to control (Pilot, Paging, and Synchronization) channels

$P_{handoff}$  : power that is reserved for the predicted handoff traffic

$$PS: \text{ video session arrival rate} = \frac{60}{3600} \times k_{\text{video}}$$

$$\text{arrival interval} = \frac{-\log(\text{rand}(1))}{\text{video session arrival rate}}$$

Figure 3-10 call admission control algorithm

	<b>Voice traffic</b>
<b>Delay</b>	Real-time, circuit-switched design makes tiny delay between voice packets unacceptable.
<b>Bandwidth</b>	Fixed 9.6kbps is enough to supply voice codec adopted in 3G system.
<b>Walsh code</b>	Single, monopolized FCH is assigned.
<b>Power</b>	The power requirement depends on fading channel, mobile speed, cochannel interference, and media codec; thus it can't be set as expected mean value.
	<b>Video traffic</b>
<b>Delay</b>	It's hard to precisely quantize each end-to-end, application-based, codec-related, subjective delay constraint, especially in unstable wireless environment. To avoid quality degradation, several buffers are equipped in both BS and MS, and hundred milliseconds of end-to-end packet delay should be kept. For example, video conference has 100ms delay bound, and real time video depending on frame rate (40fps~10fps) may have 25ms~100ms delay constraints.
<b>Bandwidth</b>	Minimum average throughput should be 28.8kbps when MPEG4 scalable video coding techniques is adopted since at least base layer packets must be transmitted to play the lowest quality of image. Instantaneous BW variation is allowable to improve system performance but delay constraint must be obeyed.
<b>Walsh code</b>	Converting the radio BW into Walsh code can we easily know that the least necessity of Walsh code is three, and the variation of code number is proportional to the assigned BW. Though the BW of a SCH can be arbitrarily relocated slot by slot as a common traffic channel, basic FCH is a dedicated traffic channel and can't be reused unless current video session is terminated. This is the severe difference between BW and Walsh code and makes them not a direct mapping.
<b>Power</b>	The power requirement depends on fading channel, mobile speed, cochannel interference, and media codec; thus it can't be set as expected mean value.
	<b>Data traffic</b>
<b>Delay</b>	Besides upper layer (TCP/IP or above) timing constraints, non-real time data services have no stringent delay bounds between contiguous packets. To keep stable service quality, however, long term average transmission rate should be kept.
<b>Bandwidth</b>	Burst delivery of data packet is suitable to be transmitted at a short interval with large bandwidth, or to be consumed through a long period by the lowest transmission rate. Adequately alternating the transmission policies can help increasing resource utilization rate.
<b>Walsh code</b>	Once the transmission bandwidth is assigned, the number of Walsh code can be known by direct mapping.
<b>Power</b>	The power requirement depends on fading channel, mobile speed, cochannel interference, and media codec; thus it can't be set as expected mean value.

Table 1 QoS criteria

## Simulation Results

In this chapter, we explore the superiorities of the proposed CFA solution, and its performance is compared with typical solutions—fairness based RRM combined with client based adaptation mechanism and fairness based RRM combined with RNF adaptation mechanism. Since the objectives of the proposed CFA solution are to enhance system performance and to simultaneously guarantee QoS, performance metrics are naturally categorized into two types—system perspective metrics and user perspective metrics. In section 4.1 and section 4.2, the system perspective metrics and the user perspective metrics are respectively defined and the relative performance study with detailed analysis follows. Section 4.3 is the summary of this chapter.

#### 4.1. System Perspective Metrics and Performance Analysis

##### u System capacity:

The system capacity is defined as the maximum number of active video users that system can simultaneously serve under predefined QoS guarantee. Thus we will observe the degradation of capacity following the increasing *reserved\_ratio*. It depends on the design the RRM, power control, and resource reservation for handoff prediction. If the power control and handoff prediction are precise enough, only the RRM can affect system capacity because it decides the resource efficiency.

##### u Blocking rate:

The blocking rate is a probability that a new incoming user is refused to enter the system. Call blocking can avoid the system overload and protect existing users' interests. It is the resultant of the call admission control, which is related to the QoS guarantee (*reserved\_resource*), the handoff prediction (*preservative\_resource*), the entry barrier (*required\_resource*), and the total resource budgets (*total\_resource*). The handoff prediction is a fixed mechanism in our study. The entry barrier is set as the predefined QoS criteria. The total resource budget is an equipment-related constant. Thus we will observe the blocking rate by following the increasing video traffic arrival rate under the specific *reserved\_ratio* that is



used to reserve resources for existing users.

#### u Channel efficiency:

In our simulation, we set maximum available bandwidth is fixed to 614.4kbps in a cdma2000 carrier and quantize it as 64 9.6kbps channels. Thus removing 4 overhead control channels, how the system exploits the remaining 60 channels to maximize its allocation efficiency is quantified as channel efficiency. The design of RRM and the number of active video users are two major factors that can affect the channel efficiency. Thus the performance comparison among three different solutions following the increasing number of active video users is analyzed.

#### u Power efficiency:

Like channel efficiency, the maximum cell site transmit power is assumed to be 20 Watts, and 16 Watts can be used to support traffic channels while 4 Watts are reserved for Pilot, Paging and Synchronization channels. Thus the proportion of power usage of the remaining 16 Watts is defined as the power efficiency. Power control, power allocation of the RRM, and power reservation for handoff request are three major factors that can affect the power efficiency. Since the power control and handoff prediction are assumed to be precise enough, the power efficiency can be regarded as a performance metric to compare the proposed solution with typical solutions.

#### u System throughput:

Basically, if every user is always active, like real time video streaming (bandwidth requirement due to continuous transmission of new arrived video packets in BSQ is inevitable), the channel efficiency can be directly transfer to the system throughput. However, different application has different traffic characteristic. For example, a voice call has 40% that people really speak some words (active), and the remaining 60% is kept silent (inactive). Data applications, like web browsing and E-mail, request resources bursty (short interval but higher bandwidth). Actually, the channel efficiency is not equivalent to the system throughput in most cases.

Figure 4-1 shows the capacity (maximum number of active video users) that cdma2000 system can support. Evidently, capacity is decreasing following the increasing *reserved\_ratio* because more and more radio resources are reserved to existing users to

guarantee their QoS. The issue we should discuss here is how to choose a reasonable *reserved\_ratio*. If a low *reserved\_ratio* is chosen, system easily enters into overload congestion, and each user will suffer buffer underflow frequently. If a high *reserved\_ratio* is chosen, existing user's resources are over-guaranteed. Each existing user can request higher visual quality (encoding bandwidth), but this limits the system capacity which is directly related to vendors' profit. Figure 4-2 is an example showing system blocking rate under *reserved\_ratio* fixed to 1, and *k\_video* represents the density of video arrival rate.

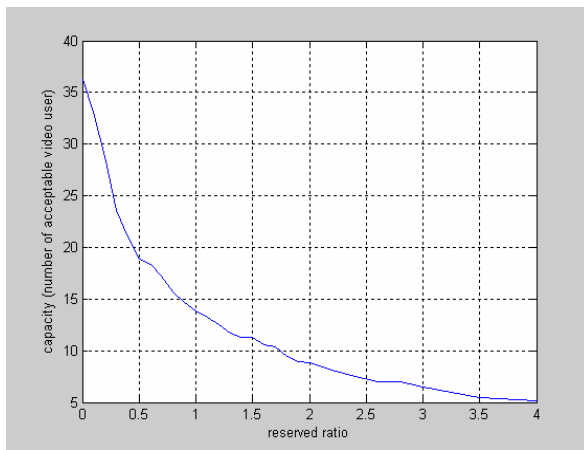


Figure 4-1 system capacity

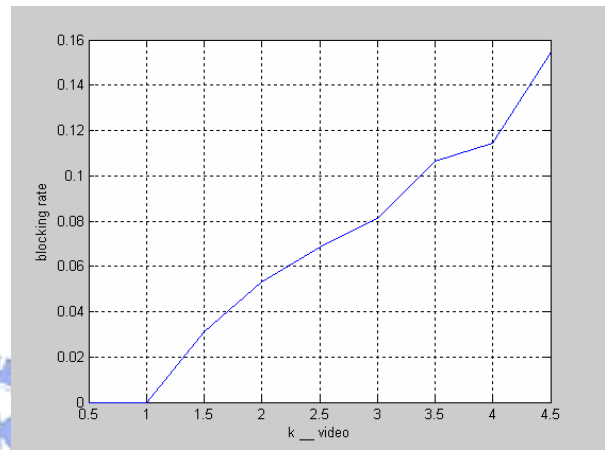


Figure 4-2 system blocking rate

Figure 4-3~4-5 represent three traditional metrics—channel efficiency, power efficiency, and system throughput. These metrics can prove that our proposed CFA solution is superior to the traditional solution from system point of view. Since different system loading (simultaneously active users) exploits different advantage of the proposed solution, the following discussion separates the metrics into three regions—light loading (1~4 users), medium loading (5~10 users), and heavy loading (11~15 users).

In the light loading scenario, system resources are sufficient to support arbitrary request from each user, and video users can always enjoy the optimal quality without suffering the risk of contention or scheduling. Hence the performance of three metrics in this region is a nearly linear growth, following the increasing number of video user. Since the power budget in this scenario is sufficient, all users would be served no matter their RF conditions. This is intuitively inefficient because the system must waste unnecessary power, and the slope of the power consumption becomes steep.

When system enters medium loading scenario, it can no more support every user in the optimal quality, but the resource budgets are still sufficient to handle the lowest quality. Though the long-term trend keeps on growing due to increasing number of existing user, each solution may vibrate its resultant channel efficiency under different video user existence. Such vibration mainly stems from the exponential growth of bandwidth up switch (9.6→19.2→38.4→76.8→153.6kbps). For example, if we suppose power budget to be infinite, six users maximize system throughput as  $163.2+163.2+86.4+86.4+48+28.8=576$ kbps; however, seven users can exploit only  $163.2+163.2+48+48+48+48+48=566.4$ kbps as the maximal system throughput. Consequently, channel efficiency and system throughput changes due to the number of video user, finite stages of the assigned bandwidth, and different combinations of the RRM and encoding adaptation mechanism. The proposed CFA solution achieves 81% channel efficiency, but two typical solutions can achieve only 65% and 35% channel efficiency in average, respectively. Power efficiency is still growing up but the slope tends to be flatter because system now can efficiently schedule users in good RF conditions first. The CFA solution is evidently superior to others because it can utilize analogous power budget to produce highest throughput.

If heavy loading scenario happens, half or above power budget is consumed to support each user's lowest visual quality, and the remaining budget limits the flexibility of quality adaptation. These result in stably high power utilization rate (80%). Simulation results prove that the CFA solution (85% channel efficiency) is superior to the typical solutions (the RNF solution with 35% channel efficiency and the client based solution with 80% channel efficiency). The 5% difference between the CFA solution and the client based solution reveals that the adaptation mechanism based on feedback information is better than blindly trial and error even rare flexibility is available.

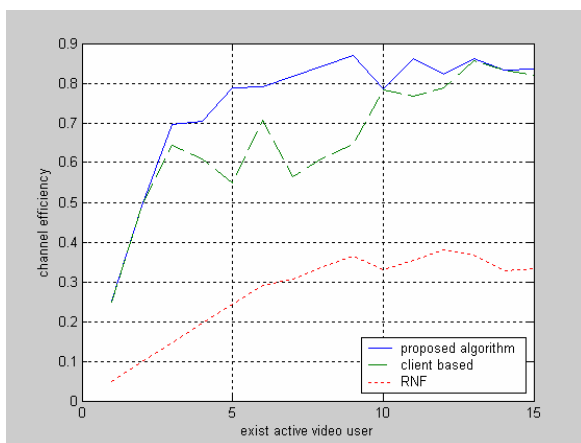


Figure 4-3 channel efficiency

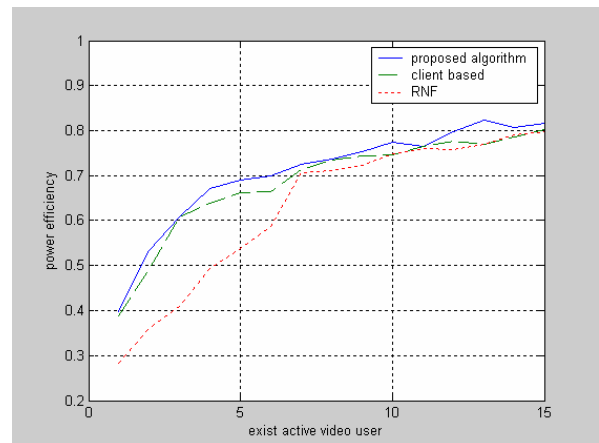


Figure 4-4 power efficiency

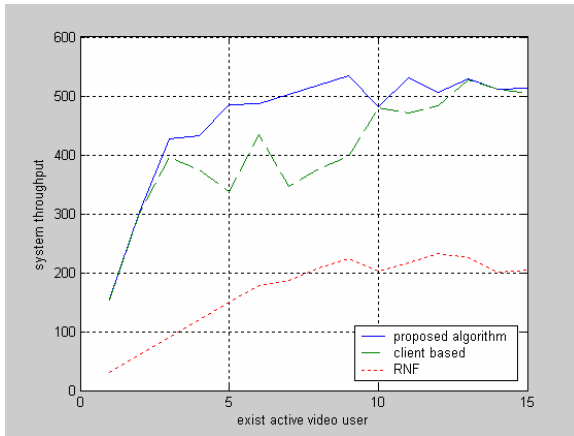


Figure 4-5 system throughput

## 4.2. User Perspective Metrics and Performance Analysis

### u Number of rebuffer event and rebuffer interval:

From the experience of serving a real time video streaming on the internet, the most unacceptable event that severely degrades the user perceived quality is the frequent suspension due to network congestion or client buffer underflow. Besides, how fast a buffer underflow event can be compensated and return to a normal clip playout is also critical. Number of rebuffer event shows the average number of rebuffer event during a 120 seconds video session, representing the frequency of buffer underflow. Rebuffer interval means the average suspend time to manipulate rebuffer events. Frequent rebuffer events and longer rebuffer interval result in the deterioration of user perceived QoS. The most effective solution is to reserve more resources for existing video users as the QoS guarantee. Thus we will observe these two metrics following the increasing *reserved\_ratio*. Only when a reasonable *reserved\_ratio* is chosen, the avoidance of buffer underflow and the maximization of resource efficiency can be optimally balanced.

### u Personal throughput:

The mean of personal throughput shows how much bandwidth a user can share following the increasing number of simultaneously active users. The standard deviation of personal throughput distinguishes the deviation of bandwidth allocation (QoS) among video users. This is a metric that shows the tradeoff between the fairness (everyone has the consistent visual quality no matter the RF condition) and the efficiency (a user in better RF condition can enjoy higher visual quality).

u Visual quality:

Visual quality can be defined as the combination of stability and resolution. Because of the lack of radio resources, to support a steadily high resolution video stream is impossible. High flexibility of the RRM and fast response of the video encoding adaptation mechanism intuitively results in relatively frequent switch of the resolution. Hence what we seek is the tradeoff between stability and resolution under the maximization of resource efficiency. Since we assume the media server can encode video frame in finite stages of scalability, corresponding to Walsh assignment in SCH (19.2 /38.4 /76.8 /153.6 kbps), the bandwidth adaptation ratio is twice per switch.

Figure 4-6 draws average number of rebuffer event and figure 4-7 draws average suspend interval of a rebuffer event during 120 seconds continuous video session. To analyze the performance of each solution, we separate performance curves into three regions—rarely reserved ( $reserved\_ratio=0\sim0.7$ ), proper reserved ( $reserved\_ratio=0.7\sim1.8$ ), and over reserved ( $reserved\_ratio=1.8\sim4$ ).

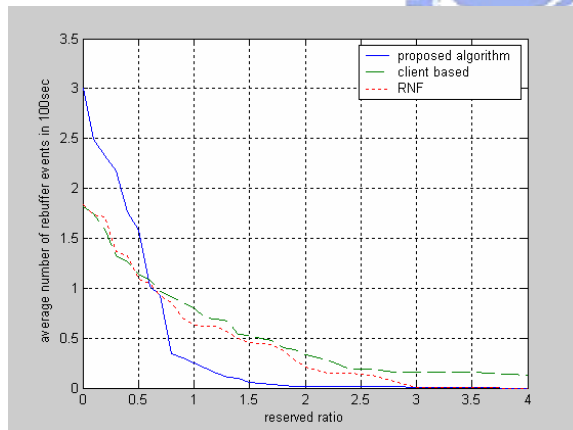


Figure 4-6 average number of buffer underflow during a 100sec clip

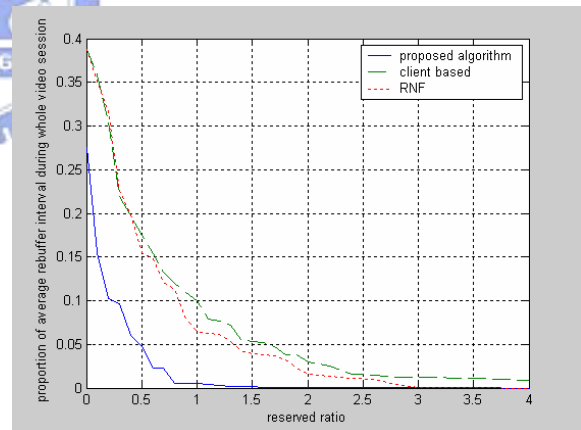


Figure 4-7 proportion of video suspension during a video session

Rarely reserved scenario, mapping to the capacity shown in figure 4-1, implies that 16~36 video users must simultaneously be served. Since the system bandwidth has been quantified into 64 channels (4 channels are control overhead, and other 60 channels are traffic channels), corresponding to 64 Walsh codes, the MPEG4 scalable video streaming, which requires roughly 30kbps to transmit base layer packets, needs 3 channels as the QoS lower bound. Thus to support 16~36 video sessions, at least 48~108 channels are necessary. We

have discussed that the maximum channel efficiencies of each solution are 81% (CFA), 65% (client based), and 35% (RNF), and the largest amount of simultaneously available channels is 48 under power budget limitation. Apparently, it is impossible for a system to sustain rarely reserved scenario, and the frequent rebuffer events are inevitable. Multiplying “average number of rebuffer event” by “average suspend interval of a rebuffer event” can we obtain the proportion of the suspend interval during a video session. Simulation results show that the CFA solution suffers 82.5%~ 2.5% suspensions, the client based solution suffers 70.2%~14% suspensions, and the RNF solution suffers 70%~12.5% suspensions. In this scenario, performances of three solutions are unsatisfied, and it is meaningless to compare the superiority. However, sharper degradation of the rebuffer event in figure 4-6 and lower proportion of the suspend interval in figure 4-7 can still help observing faster convergence of rebuffering elimination using the CFA solution.

In the proper reserved scenario, the suspend ratios of three solutions decrease from 2.5% to nearly 0% (the CFA solution), 14% to 1.6% (the client based solution), and 12.5% to 1.2% (the RNF solution), respectively. Though the performances of two traditional solutions may still not be acceptable, the CFA solution has successfully solved rebuffer problem and at the same time maintained system capacity without waste of over-reservation. Thus we call this scenario “proper reserved”. If a system designer adopts the CFA solution, it’s strongly recommended setting *reserved\_ratio* among this region (0.7~1.8). A vendor would tradeoff its QoS guarantee and capacity which affect the profit margin.

There are three phenomena can be discovered in over reserved scenario. First, reserved ratio higher than upper bound of proper region CFA is worthless unless system promises supplying advanced visual quality. Second, if the RNF solution must be used, *reserved\_ratio* above three is suggested and this implies that the RNF solution is adequate to handle light loading. Third, even a high *reserved\_ratio* can’t completely avoid rebuffer events when the client based solution is used unless the capacity shrinks to less than four users. This is because the maximum power budget (1/4 total traffic power per user) is absolutely sufficient to support a user transmitting its video clip in the lowest quality.

In our simulation database, bandwidth assigned to a user on every time slot (100ms) creates a row vector, and the records of multiple users are combined as a two-dimensional record matrix. To analyze per-user bandwidth distribution, timing average and timing standard

deviation are calculated first, and these two statistic factors among multiple users also result in probability distributions. Thus four final metrics--user expectation (mean) of timing average / timing standard deviation and user standard deviation of timing average/ timing standard deviation among different users can be calculated.

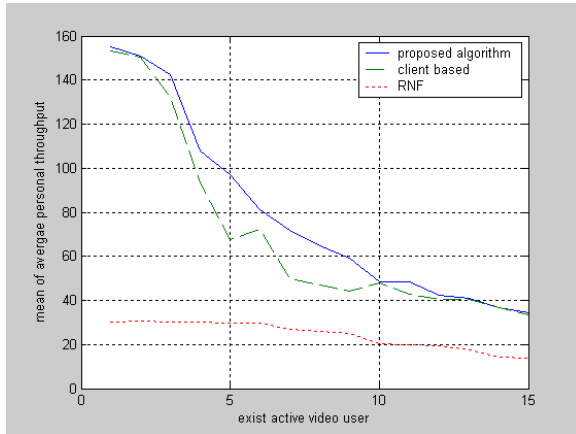


Figure 4-8 average visual quality

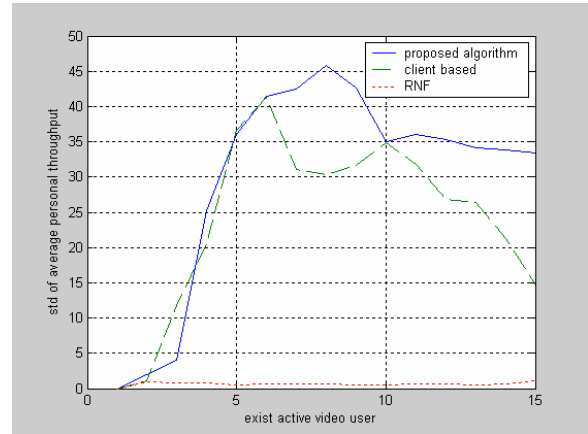


Figure 4-9 distinguishability of visual quality among video users

In figure 4-8, user expectation of timing average represents the average bandwidth a user can be guaranteed in long term. As expected, the more coexisting users, the less bandwidth one can share and two valuable messages are revealed. One is that the CFA solution enhances personal throughput (visual quality), especially in medium loading scenario, and this doubly proves the CFA solution as an efficient solution utilizing available resources. The other message tells the over preservative policy of the RNF solution, which rejects almost up-switch requests and results in the lowest stable throughput. Besides, average throughput falls below 28.8kbps means resource inefficiency and inevitable rebuffer events.

In figure 4-9, user standard deviation of timing average represents how different solutions distinguish available visual quality for users. Apparently, the RNF solution with nearly unchangeable switch policy keeps no difference among video users. Both the client based solution and the CFA solution, however, can effectively distinguish available visual quality for users due to RF conditions and client buffer fullness. In the light loading scenario, each user receives optimal quality equally, thus little difference is produced relative to their high transmission bandwidth. In medium loading scenario, the deviation rapidly increases from 5kbps to 45kbps. This means not only each user's basic quality can be taken care but

also remaining resources can be allocated to good RF users to improve system utilization rate and individual visual quality. In heavy loading scenario, all users tend to request the lowest amount of bandwidth (28.8kbps) corresponding to basic quality, but the CFA solution can still collect sparse resources to several users with better RF conditions, resulting in large deviation of visual quality.

In figure 4-10, user expectation of timing standard deviation shows possible quality vibration during whole video session. In figure 4-11, user standard deviation of timing standard deviation indicates that users under the same scenario may suffer different degree of quality change. Integrating these two metrics, the conclusive analysis displays some phenomena. First, the RNF solution has the lowest and concentrated bandwidth variation. This implies the stability during the session and the analogy among users. Second, other two solutions shrink their flexibility following the increasing users, especially at the beginning of medium loading scenario. Third, when heavy loading scenario happens, a user with bad RF condition can receive only basic but stable quality, but others who attempt to exploit remaining resources would suffer exquisite variation. The violence produced by rarely good RF users results in large deviation.

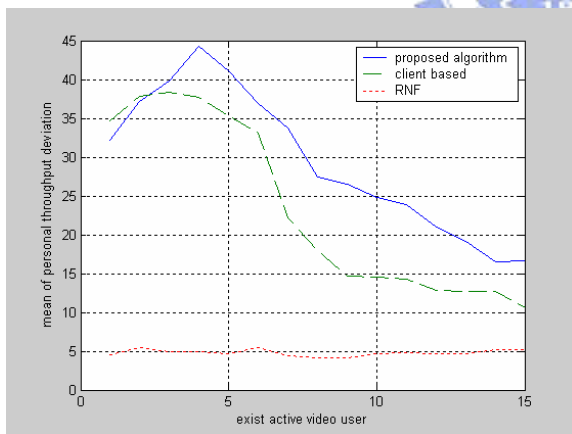


Figure 4-10 vibration of visual quality during a video session

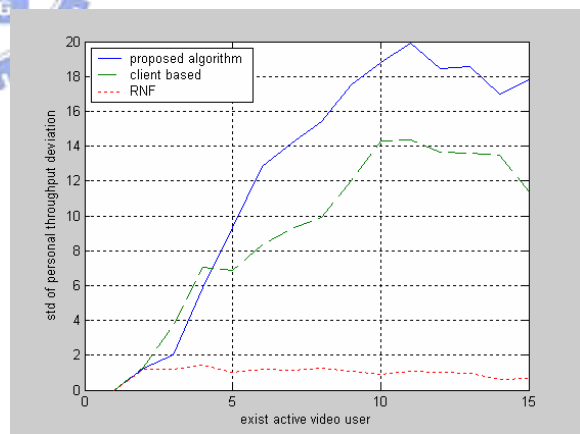


Figure 4-11 user standard deviation of timing standard deviation of personal throughput

Final two figures show mean and standard deviation of bandwidth switch of media server between contiguous judgments. Though mean and standard deviation of the adaptation ratio can somehow represent stability of a visual quality, we can't use it as an absolute judgment which adaptation algorithm is better. As we have discussed previously, it's very tricky to precisely define visual quality. For example, psychologically speaking, is encoding



bandwidth kept in the lowest stage but steady state better than it be switched to upper stage during an interval if available? How is the switch frequency acceptable for human eyes? These uncertain issues that highly depend on human perspective are difficult to be quantified and we only shows the trend and results of encoding bandwidth variation in our study, without any comment which adaptation algorithm is better.

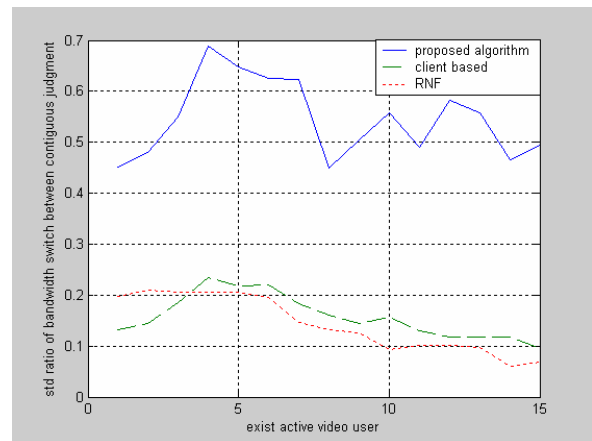
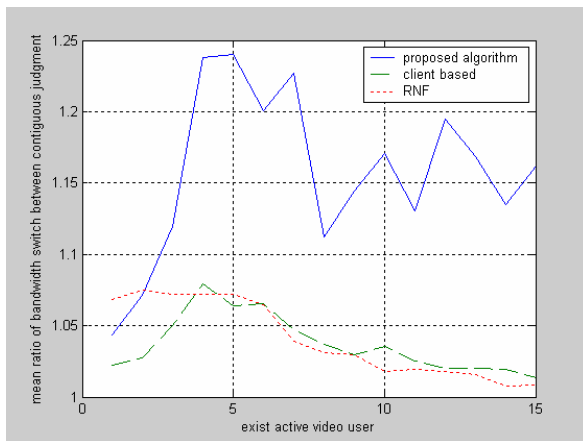
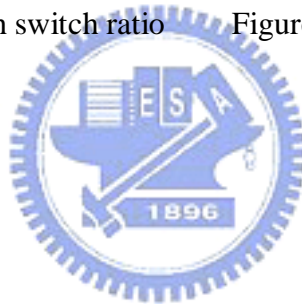


Figure 4-12 mean of bandwidth switch ratio

Figure 4-13 standard deviation of bandwidth switch ratio



### Conclusion and Future Works

In our study, client feedback assisted (CFA) solution is proposed to improve system resource utilization rate and to enhance individual QoS. The core concept focuses on relocating resources to scalable streaming services by simultaneously rescheduling priorities in front-end wireless system and reassigned source bit rate of each streaming in back-end multimedia sources. CFA exploits feedback information from client buffer, current fullness of base station queue, RF condition, and system loading to decide the priority and source bit rate.

Consequently, from user perceived performance, CFA solution can avoid most rebuffering conditions that client-based method alone can't detect. From resource efficiency of view, CFA solution is superior to RNF solution although RNF can precisely adapt back-end encoding bit rate with front-end channel information.

More and more real-time, interactive, and heterogeneous applications will be served in wireless system. Thus collecting core information that critically affects the quality and merging them into system RRM will be the design trend. Furthermore, how to implement analogous concept when cross system migration is capable (for example, 3G cellular system to WLAN/WMAN) is also a challenge.

## References

- [1] Roger L. Peterson, Rodger E. Ziemer, David E. Borth, "Introduction to Spread Spectrum Communications", Prentice Hall International Editions, 1995
- [2] Theodore S. Rappaport, "Wireless Communications", second editions, Prentice Hall, 2002
- [3] MPEG-4 Industry Forum, "MPEG-4 – The Media Standard", m4-out-20027-R3.pdf
- [4] Joe Huang, Yuqi Yao, Yong Bai, and Szu-Wei Wang, "Performance of a Mixed-Traffic CDMA2000 Wireless Network with Scalable Streaming Video", IEEE Transaction on Circuits and systems for video technology, Vol. 13, NO. 10, October 2003
- [5] Svetlana Chemiakina, Luigi D'Antonio, Francesco Forti, Roberto Lalli, Justus Petersson, and Alessio Terzani, "QoS Enhancement for Adaptive Streaming Services over WCDMA", IEEE journal on selected areas in communications, Vol. 21, NO. 10, December 2003
- [6] Kamran Etemad, "CDMA2000 Evolution : system concepts and design principles", 2004 by WILEY-INTERSCIENCE
- [7] David J. Goodman, "Wireless Personal Communications Systems", 1997 by Addison Wesley Longman, Inc.
- [8] The Members of Technical Staff, Bell Labs, "Handbook of CDMA System Design, Engineering, and Optimization", 2000 by Prentice Hall PTR
- [9] "The cdma2000 ITU-R RTT Candidate Submission", v0.18, July 1998
- [10] "Overview of the MPEG-4 Standard", March 2001 ISO/IEC JTC1/SC29/WG11 N4030
- [11] Miroslaw Bober and Josef Kittler, "Video Coding for Mobile Communications – MPEG4 Perspective", 1996 the Institution of Electrical Engineers
- [12] "Coding of Moving Pictures and Audio", ISO/IEC JTC1/SC29/WG11, MPEG2003/M10298, Hawaii, December 2003
- [13] Demitri Bertsekas and Robert Gallager, "Data Networks", second edition, 1992 by Prentice-Hall, Inc.
- [14] Mario Baldi, and Yoram Ofek, "End-to-End Delay Analysis of Videoconferencing over Packet-Switched Networks", IEEE/ACM transactions on networking, Vol. 8, NO. 4, August 2000
- [15] Savo Glisic and Branka Vucetic, "Spread Spectrum CDMA Systems for Wireless Communications", 1997 by Artech House, Inc
- [16] Kim K. Leung and Branimir Vojcic, "Multiaccess, Mobility and Teletraffic for Wireless Communications: Volume 3", 1999 by Kluwer Academic Publishers

- [17] Takeshi Yoshimura, Tomoyuki Ohya, Hosei Matsuoka, and Minoru Etoh, "Design and Implementation of Mobile QoS Testbed "MOBIQ" for Multimedia Delivery Services",
- [18] Jinwei Cao, Dongsong Zhang, Kevin M. McNeill, and Jay F. Nunamaker, Jr., "An Overview of Network-Aware Applications for Mobile Multimedia Delivery", Proceedings of the 37th Hawaii International Conference on System Sciences - 2004
- [19] Chia-Hui Wang, Jan-Ming Ho, Ray-I Chang, and Shun-Chin Hsu, "A Control-theoretic Rate-based Control of Buffer-Occupancy Feedback for Real-time Multimedia Communications"
- [20] C. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over burst-error wireless channels," *IEEE J. Select. Areas Commun.*, vol. 17, May 1999.
- [21] Ana-Belén García, Manuel Alvarez-Campana, Enrique Vázquez, Julio Berrocal, "Quality of Service Support in the UMTS Terrestrial Radio Access Network"
- [22] TIA/EIA INTERIM STANDARD, G3G CDMA-MC to GSM-MAP, TIA/EIA/IS-833, MARCH 2000, TELECOMMUNICATIONS INDUSTRY ASSOCIATION
- [23] Thomas Stockhammer, "IS FINE-GRANULAR SCALABLE VIDEO CODING BENEFICIAL FOR WIRELESS VIDEO APPLICATIONS?", IEEE ICME 2003
- [24] Weiping Li, Fan Ling, and Xuemin Chen, "Fine Granularity Scalability in MPEG-4 for Streaming Video", ISCAS 2000- IEEE international Symposium on Circuits and Systems, May 28-31, 2000
- [25] Susie Wee, Wai-tian Tan, John Apostolopoulos, Minoru Etoh, "Optimized Video Streaming for Networks with Varying Delay"
- [26] Weiping Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard", IEEE Transactions on Circuits and Systems for Video Technology, VOL. 11, NO. 3, MARCH 2001
- [27] Chun Yuan, Bin B. Zhu, Yidong Wang, Shipeng Li, Yuzhuo Zhong, "Efficient and Fully Scalable Encryption for MPEG-4 FGS"