Scalable Multiple Description Coding and Distributed Video Streaming over 3G Mobile Networks

by

Ruobin Zheng

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2003

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, a novel Scalable Multiple Description Coding (SMDC) framework is proposed. To address the bandwidth fluctuation, packet loss and heterogeneity problems in the wireless networks and further enhance the error resilience tools in Moving Pictures Experts Group 4 (MPEG-4), the joint design of layered coding (LC) and multiple description coding (MDC) is explored. It leverages a proposed distributed multimedia delivery mobile network (D-MDMN) to provide path diversity to combat streaming video outage due to handoff in Universal Mobile Telecommunications System (UMTS). The corresponding intra-RAN (Radio Access Network) handoff and inter-RAN handoff procedures in D-MDMN are studied in details, which employ the principle of video stream re-establishing to replace the principle of data forwarding in UMTS. Furthermore, a new IP (Internet Protocol) Differentiated Services (DiffServ) video marking algorithm is proposed to support the unequal error protection (UEP) of LC components of SMDC. Performance evaluation is carried through simulation using OPNET Modeler 9.0. Simulation results show that the proposed handoff procedures in D-MDMN have better performance in terms of handoff latency, end-to-end delay and handoff scalability than that in UMTS. Performance evaluation of our proposed IP DiffServ video marking algorithm is also undertaken, which shows that it is more suitable for video streaming in IP mobile networks compared with the DiffServ video marking algorithm (DVMA) proposed in [71] [72].

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor W. Zhuang, for her guidance throughout this work. She was always available, encouraging and professional. From her, I have learned not only how to do good research, but also how to be a better person. Her encouragement, patience and support made my thesis possible. In addition, I would like to thank Professors X. Shen and P. Ho for their careful and thorough reading of this thesis, and for their valuable and helpful comments for my research work.

Contents

List of Tables

List of Figures

1 Introduction

With the emergence of broadband wireless networks and increasing demand of multimedia information on the Internet, wireless video communications have received great interest from both industry and academia, and wireless multimedia services are foreseen to become widely deployed in this decade.

Real-time transport of live video or stored video is the predominant part of real-time multimedia. There are two concepts for delivery of stored video over the Internet or the wireless networks, namely the downloadable video and the streaming video [1].

The video download is the same concept as the file download, but a large file. The entire video file is expected to be downloaded on the local machine, where it could be played back using the standard media software. It allows simple delivery mechanisms, e.g., Transmission Control Protocol (TCP). However, it usually suffers long and perhaps unacceptable transfer time and large storage space. Also, download before viewing requires the user's patient.

In contrast, the streaming video is partitioned into packets. It needs not be downloaded in full, but is being played out simultaneously during video delivery. There is relatively low delay (e.g., Real and Microsoft use 5-15 seconds) of starting playback before viewing. It also minimizes the storage requirement.

In this thesis, we are more concerned with the streaming video, which refers to real-time transmission of stored video. Its relevant techniques are also applicable to the delivery issues of live video.

1.1 Motivation of This Research

1. Bandwidth, packet loss and heterogeneity problems

Due to its real-time nature of transmission of stored video, video streaming typically has quality of service (QoS) requirements, e.g., bandwidth, delay and error requirements. However, unreliability, bandwidth fluctuations and high bit error rate of wireless

channels can cause severe degradation to video quality. In addition, for video multicast, network heterogeneity and receiver heterogeneity make it difficult to achieve efficiency and flexibility. The bandwidth problems, packet loss and transmission error, and heterogeneity problems will be discussed in detail as follows.

To address the bandwidth fluctuation, packet loss and heterogeneity problems, scalable coding (i.e., layered coding) and rate shaping are employed for transport delay- and bandwidth-sensitive video. Furthermore, to enhance the video quality in the presence of unavoidable packet loss and/or bit error, open-loop error control (e.g., multiple description coding) and close-loop error control (e.g., delay-constrained retransmission) are explored in this thesis.

1) Bandwidth problems

To achieve acceptable presentation quality, a streaming application typically has minimum bandwidth requirement. However, the current Internet offers only the best-effort service and does not provide any bandwidth reservation mechanism. It is well known that fluctuations of the Internet traffic have a fractal-like scaling behavior over time scales [25]. Due to the self-similar nature of traffic fluctuations in the Internet, the available bandwidth is unknown and dynamic.

In the wireless networks, the wireless channel suffers from both bandwidth fluctuation and bandwidth limitation: (1) The throughput of a wireless channel may be reduced due to multipath fading, shadowing, co-channel interference, and noise disturbances; (2) When a mobile terminal moves between different networks (e.g. from a wireless local area network (LAN) to a wireless wide area network (WAN)), the available bandwidth may vary drastically (e.g., from a few megabits per second to a few kilobits per second); (3) When a handoff takes place, a base station may not have enough unused radio resources to meet the demand of a newly admitted mobile host. Thus, the available bandwidth of wireless channel is time-varying and even unknown.

If the transmission rate of streaming video is faster than the available bandwidth, the congestion will occur, resulting in bursty loss, excessive delay and severe drop in video quality. On the contrary, it invokes the inefficient utilization of available bandwidth and the sub-optimal video quality. Thus, it is desirable for streaming video application to employ congestion control mechanisms to match video bit rate with available bandwidth.

2) Packet loss and transmission error

In the wired link of a mobile network or the Internet, most errors are caused by packet loss due to congestion or misrouting. The effects of packet loss greatly depend on the types of packet loss [47]: isolated single packet loss, burst packet loss and temporary outage (loss of communication). Misrouting may occur in the downlink during handoffs, which will incur bursty packet loss and temporary outage.

The wireless channels are typically more error-prone at the bit level. The wireless link of a mobile network suffers from very high bit error rate (BER) due to multipath fading, shadowing, co-channel interference, noise disturbances and handoff. The types of transmission bit error also can be classified into three groups: isolated single bit error, burst bit error, and temporary outage (e.g., due to handoff).

The effects of packet loss or bit error are significant for video streaming due to error propagation [27]. Predictive video-encoding algorithms employ motion compensation to achieve high compression by reducing temporal redundancies between successive frames. When this motion information is lost to the decoder, a reconstruction error can occur. Such errors can propagate temporally and spatially if the affected region is subsequently used as a prediction for motion compensation. Furthermore, differential encoding is also employed to reduce statistical redundancies, for example in the encoding of motion and quantizer information. Loss of such information can cause additional spatial degradation throughout the affected frames by producing incorrectly predicted motion vectors and quantizer levels. Because of motion compensation, these errors also can propagate temporally and spatially.

Because of error propagation of streaming video, isolated single packet loss or bit error is converted to burst packet loss or bit error. Also, the video packet which arrives beyond a delay bound is useless and has to be considered lost. Such loss or error can potentially make the visual presentation displeasing to human eyes or even make the presentation impossible.

From a video communication perspective, it is important to reduce or eliminate the effects of burst loss/error and outage. The error characteristics of video communication in different environments are roughly summarized in Table 1-1 [28] [38].

Table 1-1 Error characteristics of video communication

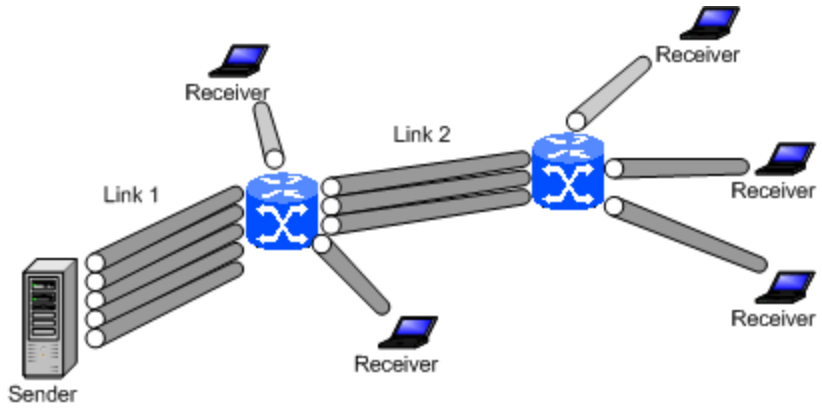
Application	Error Characteristics
Video phone over PSTN (H.324)	Very few bit errors and packet losses
Video conferencing over ISDN (H.320)	Practically error free (BER=10^{-10}~10^{-8})
Video conferencing over ATM (H.310, H.321)	Almost error free (CLR=10^{-6}~10^{-4})
Digital television	Almost error free (after FEC)
Terrestrial/cable/satellite TV	Almost error free (depend on weather)
Video phone over the Internet (H.323)	BER = 0, Packet loss of 0~30%
Mobile video phone (H.324 wireless)	BER = 10^{-5}~10^{-3}, burst errors

To enhance the video quality in presence of unavoidable packet loss or bit error, error control mechanisms should be used.

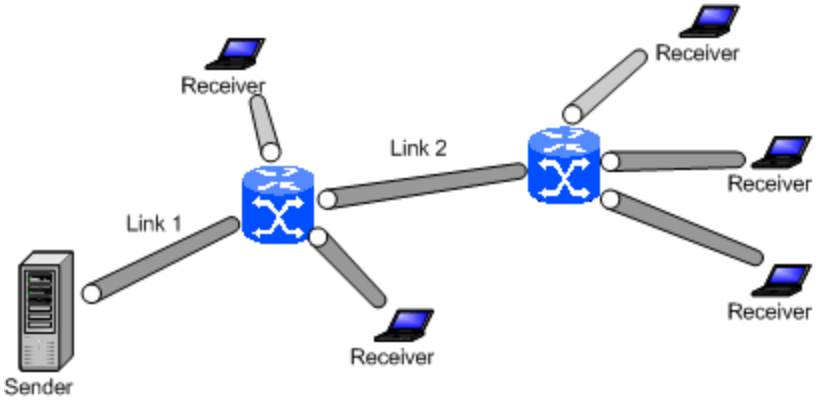
3) Heterogeneity problems

Before addressing the heterogeneity problems, we first compare unicast with multicast. The unicast delivers streaming media through point-to-point transmission, where only one sender and one receiver are involved. In contrast, multicast delivers streaming media through point-to-multipoint transmission, where one sender and multiple receivers are involved. For streaming applications such as video conferencing and Internet television, multicast delivery can achieve high bandwidth efficiency since the receivers can share links. On the other hand, unicast delivery of such applications is inefficient in terms of bandwidth utilization. An example is given in Figure 1-1, where, for unicast, five copies of the video content flow across Link 1 and three copies flow across Link 2 as shown in Figure 1-1 (a); For multicast in Figure 1-1 (b), there is only one copy of the video content traversing any link in the network, resulting in substantial bandwidth savings. However, the efficiency of multicast is achieved at the cost of losing the service flexibility of unicast (i.e., in unicast, each receiver can individually negotiate service parameters with the source). Such lack of flexibility in multicast can be problematic in a heterogeneous network environment. For example, the receivers in Figure 1-1 (b) may attempt to request different video quality with different bandwidth.

But only one copy of the video content is sent out from the source. As a result, all the receivers have to receive the same video content with the same quality.



(a) Unicast video distribution using multiple point-to-point connections



(b) Multicast video distribution using point-to-multipoint transmission

Figure 1-1 Unicast and Multicast video distribution

In a public land mobile network, there are two kinds of heterogeneity, namely, network heterogeneity and receiver heterogeneity [2]. Network heterogeneity refers to the different domains (e.g., wireless domain and wired domain) having unevenly distributed resources (e.g., processing, bandwidth, storage, and congestion control policies). Network heterogeneity can make different users experience different packet loss/delay characteristics. Receiver heterogeneity means that receivers have different or even varying latency requirements, visual quality requirements, and/or processing capability.

It is a challenge to design a multicast mechanism that not only achieves efficiency in network bandwidth, but also meets the various requirements of the receivers.

2. Handoff design issues in media delivery

The problem of handoff in the wireless network is well-know, however it is largely unexplored in the applications of streaming media. There are a serial of problems or requirements associated with media streaming during seamless handoff, such as handoff latency (or media stream interruption), end-to-end delay (or service delivery time), media synchronization and Handoff scalability. However, the handoff procedures in UMTS R99 and Release 4 [57] [58] [59]may not satisfy the requirements of seamless handoff for media streaming services. The handoff design issues of media streaming will be discussed in details in Chapter 2.

To address the handoff problems in media streaming, a Scalable Multiple Description Coding framework is proposed together with distributed video storage in the DiffServ mobile network to support streaming video handoff. It leverages the distributed multimedia delivery mobile network to provide path diversity to combat outage due to handoff. Since the media streaming services are pushed to the edge of core network so that the streaming media is sent over a shorter network path, it also reduces the media service delivery time, the probability of packet loss, and the total network resource occupation with relatively consistent QoS in all scenarios.

3. Error resilience enhancement for MPEG-4 video

To further explore the error resilience and concealment tools in MPEG-4, the shape, motion, and texture information in the bit-stream of an object-based video are re-organized into different layers in the proposed SMDC scheme to support the classification and priority assignment in the DiffServ network. Moreover, due to the joint design of LC and MDC, it is possible to overcome the drawbacks of LC and MDC

4. UMTS QoS and IP DiffServ

It is challenging to provide QoS attribute translation and mapping between the IP networks and the UMTS systems and to implement the IP differentiated services for the traffic encapsulated and isolated by tunneling in UMTS. In order to support the unequal error protection for layered video, a UMTS-to-DiffServ QoS mapping scheme and its marking algorithm for MPEG-4 scalable video are proposed in the thesis. Furthermore, it spurs the evolution of UMTS toward its final all-IP phase for the purpose of addressing the DiffServ tunneling issue in UMTS.

This thesis studies the architecture of third generation mobile networks and the handoff procedures for video delivery in UMTS. In order to address the handoff issues in video streaming, as well as the bandwidth fluctuation, packet loss and heterogeneity problems in the wireless networks, and to further enhance the error resilience tools in MPEG-4, a scalable multiple description coding framework together with a distributed multimedia delivery mobile network is proposed. The corresponding intra-RAN handoff and inter-RAN handoff procedures in D-MDMN are studied. Furthermore, a new IP DiffServ video marking algorithm is explored to support the UEP of SMDC. Simulation results show that the proposed scheme achieves performance improvements compared with the original UMTS and DVMA solutions.

1.2 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 overviews and discusses the General Packet Radio Service (GPRS) and UMTS mobile network architecture, handoff protocols and handoff problems for media streaming.

In Chapter 3, the concepts of layered coding and multiple description coding are introduced, followed by the proposed mobile system model (i.e., the distributed multimedia delivery mobile network) and the protocol stack of network-aware end system.

Chapter 4 presents the details of the proposed scalable multiple description coding, the corresponding handoff procedures in the D-MDMN and a novel IP DiffServ MPEG-4 video marking algorithm.

The simulation models, system setup, test conditions, and simulation results are presented and analyzed in Chapter 5.

Finally, Chapter 6 gives conclusions of this work and suggestions for further research.

2 Handoff Design Issues in Media Delivery

Handoff is a basic mobile network capability for dynamic support of terminal migration. In order to illustrate the proposed solution to solve the handoff problems of media streaming, the GPRS/UMTS mobile systems and handoff protocols are first studied.

2.1 GPRS Network Architecture

The GPRS network configuration is outlined in TS 23.002 [60] [64] and illustrated in Figure 2-1.

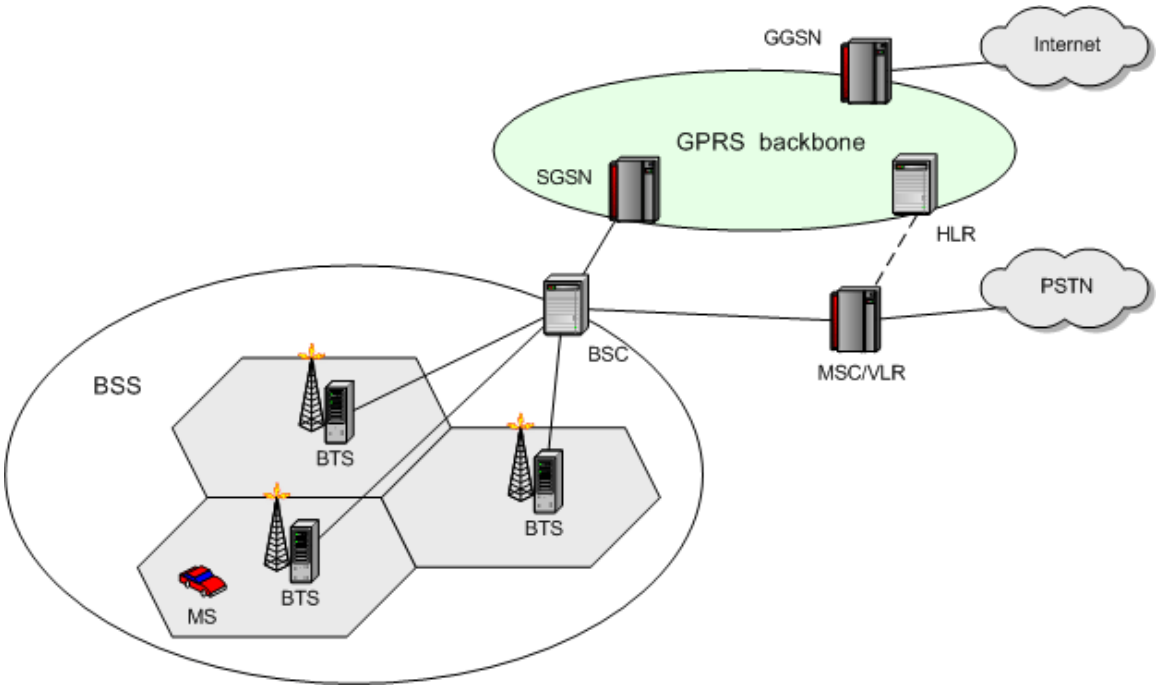


Figure 2-1 GPRS Network Architecture

Current services (voice and circuit-switched data) are supported via the base station subsystem (BSS) and network subsystem (NSS). The BSS consists of the base transceiver station (BTS) that handles the radio physical layer and the base station controller (BSC)

that deals with radio resource management and handover. The NSS for circuit-switched services consists of the mobile switching center (MSC), the visitor location register (VLR) integrated in the MSC, and the home location register (HLR).

GPRS provides packet-switched services over the Global System for Mobile communications (GSM) radio. The major new element introduced by GPRS is an NSS (GPRS backbone) that processes all the data traffic. It comprises two network elements:

- Serving GPRS support node (SGSN), which keeps track of the location of individual mobile stations and performs security functions and access control.
- Gateway GPRS support node (GGSN), which encapsulates packets received from external packet networks (Internet) and routes them toward the SGSN.

The interface between the BSS and the SGSN is based on the frame relay transport protocol. The SGSN and GGSN are interconnected via an IP network. No layer-two technology has been specified.

2.2 3GPP UMTS Network Architecture

The telecommunication system standardised by the Third Generation Partnership Project (3GPP) consists of a core network and a radio access network that may be either GERAN or UMTS Terrestrial Radio Access Network (UTRAN), or both. The UMTS network [60] [64], shown in Figure 2-2, consists of two independent subsystems connected over a standard interface:

- Radio access network, which may be either GSM/EDGE radio access network (GERAN) or UMTS terrestrial radio access network (UTRAN), or both. UTRAN composes of node B and a radio network controller (RNC). Node B is functionally similar to the GSM BTS, and RNC is similar to the GSM BSC.
- UMTS core network (CN), which is equivalent to the GSM/GPRS NSS.

The separation of the radio access network (RAN) from the core network is the fundamental concept of the cellular system.

The UMTS core network reuses as much as possible the GSM/GPRS NSS:

- Packet switched (PS): an evolution of the GPRS SGSN/GGSN with a more optimized functional split between the UTRAN and core network.
- Circuit switched (CS): an evolution of the NSS with the transcoder function moved from the BSS to the core network.

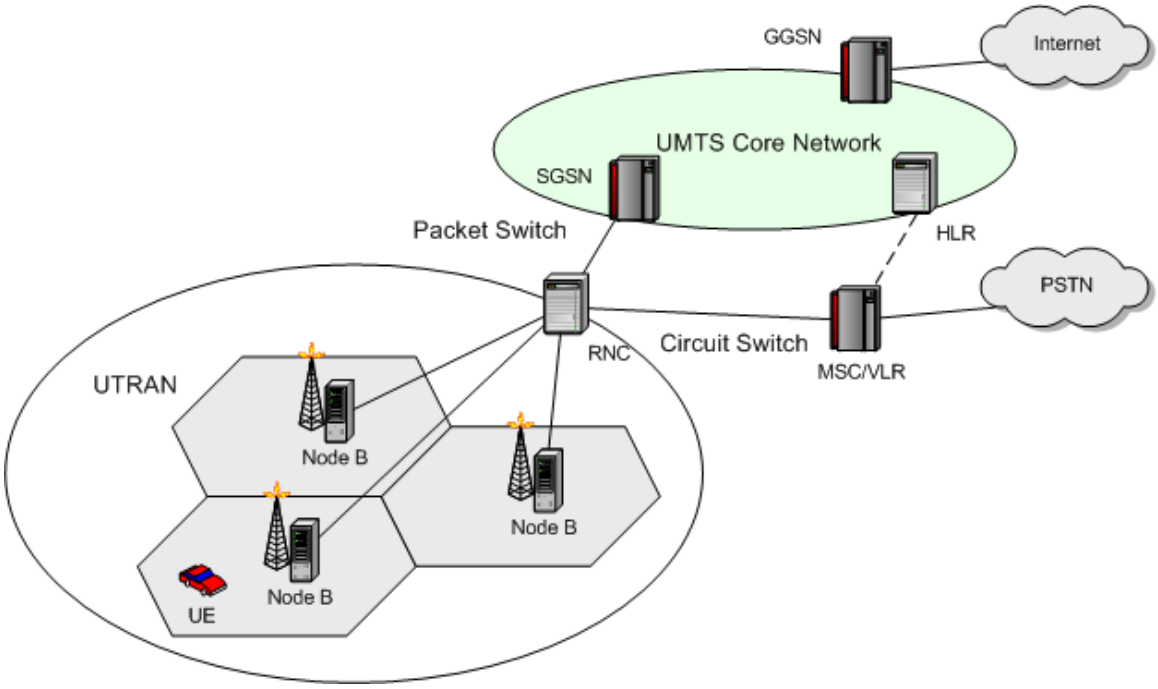


Figure 2-2 UMTS Network Architecture

The UTRAN consists of several possibly interconnected radio network subsystems (RNSs). An RNS contains one RNC and at least one node B. The RNC is in charge of the overall control of logical resources provided by the node Bs. RNCs can be interconnected in the UTRAN (i.e., an RNC can use resources controlled by another RNC). In case of a WCDMA RAN, the RNC provides soft handover, combining and splitting between streams from different base stations belonging to the same mobile station.

Node B provides logical resources, corresponding to the resources of one or more cells, to the RNC. It is responsible for radio transmission and reception in the cells maintained by this node B. A node B controls several cells.

2.3 UMTS Handoff Procedures for Media Streaming

2.3.1 Assumptions

Before the presentation of the handoff procedures in UMTS Release 4 for media streaming, the following assumptions are made [52].

1. The handoff procedures are designed specially for media streaming services in packet switched service domain. The QoS attributes required by audio and video media streams are defined in 3GPP TS 23.107 V5.6.0 [61], as discussed in Section 2.4.1.
2. This research focuses on the hard handoff procedures. Soft handoff may provide better performance for media streaming. However, hard handoff is required when there are no connections between source RNC and target RNC within the mobile network, especially under the consideration of network heterogeneity and receiver heterogeneity (such as interworking between UTRAN and GERAN or UTRAN and IEEE 802.11 (WLAN)).
3. The handoff procedures are instantiated in GPRS. However, they are also applicable to either GSM (with the SGSN/GGSN being replaced by MSC/GMSC), or UMTS (with the BSC/BTS being replaced by RNC/Node B).
4. Mobile assisted handoff is adopted. That is, the mobile station (MS) assists the network by taking periodic measurements on the downlink and relaying them back to the network for handoff decision making.
5. The inherent GPRS tunneling protocol (GTP) of GSM/GPRS/UMTS [51] [70] is used to support mobility.
6. Based on the measurements on both downlink and uplink which are performed by MS and BTS/Node B respectively, the BSC/RNC makes handoff decisions.
7. In both intra-RAN and inter-RAN handoffs, the MSC/SGSN determines the readiness of the new access point to accommodate the handoff; in inter-cell, intra-RNS handoff, the BSC/RNC does so; while in intra-cell handoff, the BTS/Node B does so.

8. For media streaming, the BTS/Node B ensures that the handoff algorithm maintains packet sequencing after handoff.
9. The handoff procedures in UMTS Release 4 are presented as follows only in the scenarios of intra-RAN handoff and inter-RAN handoff. The details of other scenarios, such as inter-cell, intra-RNS handoff, and intra-cell handoff are similar and will not be repeated in this section.

2.3.2 Handoff Procedures in UMTS Release 4

2.3.2.1 Intra-RAN Handoff Procedure

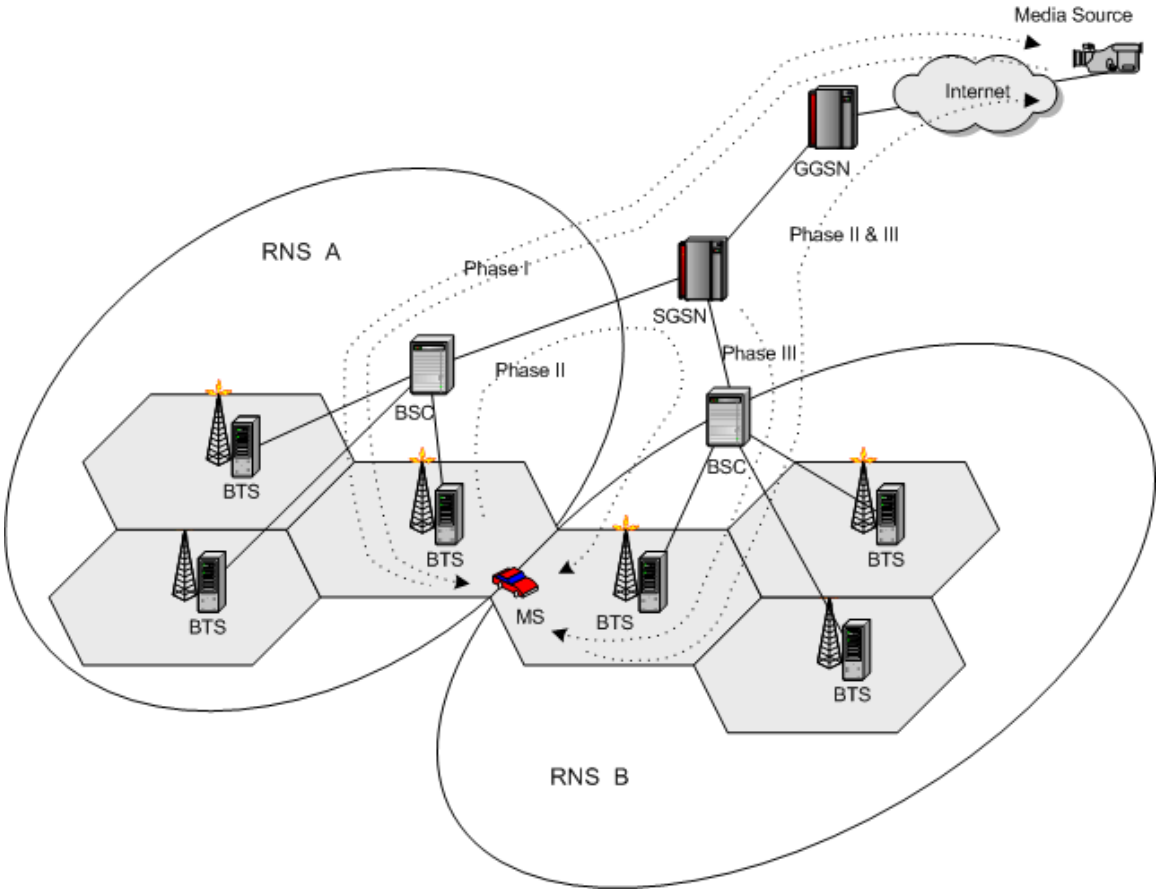


Figure 2-3 UMTS Rel4 network model of intra-RAN handoff (Data plane)

Figure 2-3 illustrates a typical UMTS network model under the IP transport mode in the scenario of intra-RAN handoff. In UMTS Release 4, the media service provider is

outside the CN which consists of SGSNs and GGSN, and far from the MS. The media streams should first get through CN and then can feed into the RNS A or B.

In the IP transport mode, the destination IP address of an end-user packet is not used to make the packet forwarding decision. Instead, the packets are encapsulated in an intermediate layer (e.g., frame relay transport protocol (FP) in the RAN and GTP in the CN, which may be specific to the chosen wireless technology. The encapsulated data units are then transported, between the nodes in the segment, over another IP layer. Most of the existing proposals espouse this approach, which allows the mobile operator to keep many of the legacy components of the 2G network untouched while upgrading just the transport layer from point-to-point lines or an ATM network to an IP-based network.

The control plane of handoff procedure [57] [58] [59] consists of three phases, as shown in Figure 2-4, while the data plane of these three phases is shown in Figure 2-3.

- **Phase I: Preparation of RNS handoff and resource allocation**

The MS sends its periodic measurement reports (signal #1). Based on these reports and its own measurement and on current traffic conditions, the source RNS (sRNS) makes the decision to perform a handoff and sends an HO-required message (signal #2) to inform the SGSN about the identifier of the target RNS (tRNS) to which the MS attempts to make a handoff. The SGSN then shall generate an HO-request message (signal #3) to the selected tRNS and requests the allocation of resources for the MS. The tRNS checks if enough radio resources are available and activates a physical channel at the tRNS to prepare for the arrival of the MS. Once resource allocation has been completed by the tRNS, it shall return an HO-request-ack message (signal #4) to the SGSN. When this message is received by the SGSN, it starts to set up a link (i.e., GTP tunnel) to the tRNS, indicates the completion of the preparation phase on the core network side for the handoff by sending an HO-command message (signal #5) to the sRNS.

Note that: The HO-request-ack (signal #4) from the tRNS contains the complete radio interface message that shall be sent by the sRNS to the MS in the HO-command (signal #6), the SGSN transparently passes this radio interface message onto the sRNS.

For the data plane of handoff phase I, upon receiving the signal #5 at the end of the preparation phase, the sRNS stops transmitting downlink data to the MS and should store all downlink data which continue to arrive from the SGSN to the sRNC.

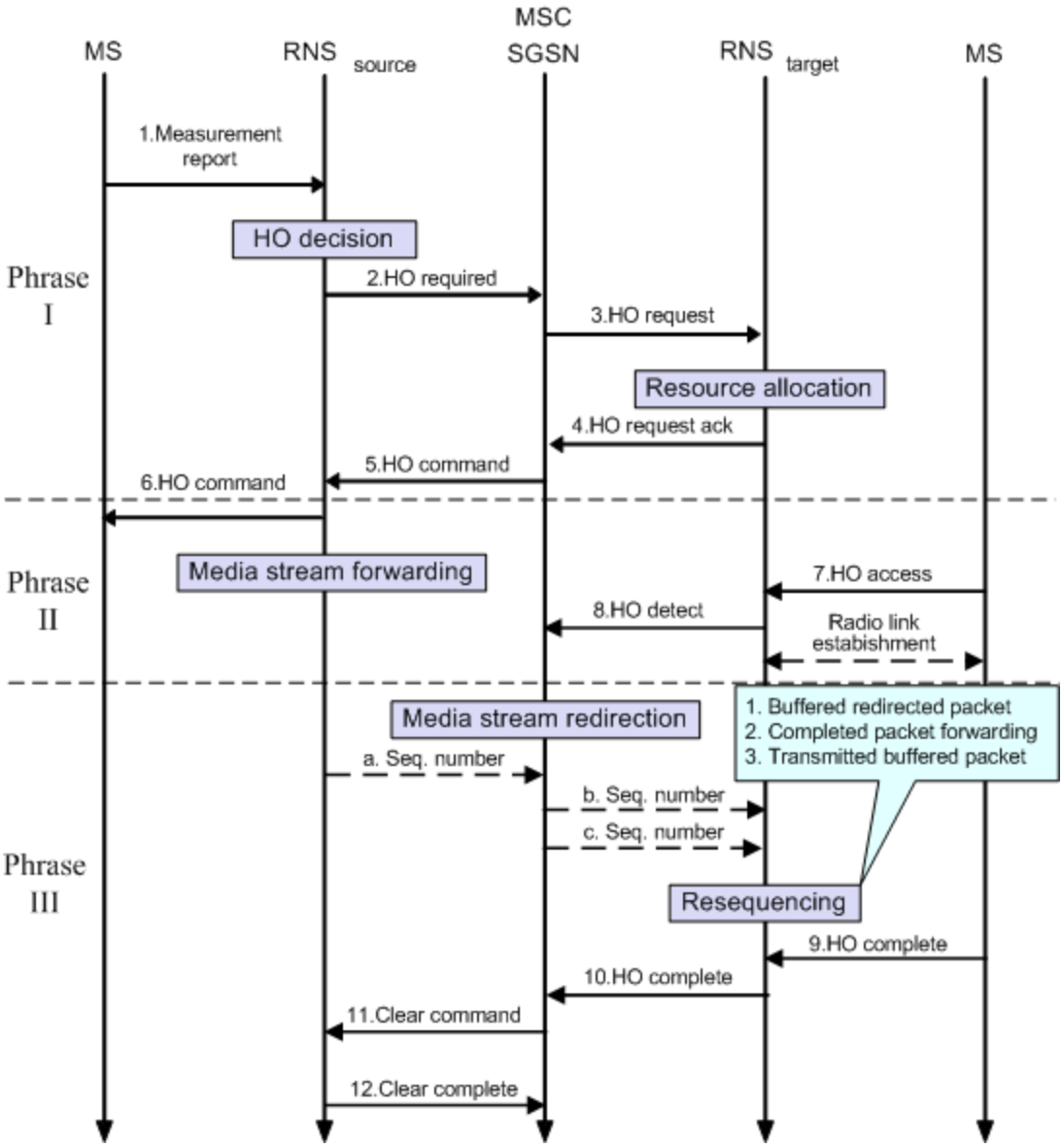


Figure 2-4 Intra-RAN handoff procedure in UMTS Rel4 (Control plane)

- **Phase II: Moving the serving RNS role to target RNS**

On receipt of the HO-command (signal #5), the sRNS will issue the radio interface message HO-command (signal #6), containing a *Handover Reference Number* previously allocated by tRNS, to the MS. The MS will then break its old radio link and access the new radio resource using the *Handover Reference Number* contained in the HO-access message (signal #7). The number will be checked by the tRNS to ensure that it is as expected and that the correct MS has been captured. If this is the correct MS, the tRNS shall send an HO-detect message (signal #8) to the SGSN.

For the data plane of handoff phase II, upon receiving the signal #5 in Phrase I and at the beginning of the execution phase, the sRNS starts to forward all the buffered data (including state information for session migration) to the tRNS, via SGSN.

Once data forwarding is started, the tRNS stores all GTP Protocol Data Unit (GTP-PDUs) forwarded from the sRNS. When Serving RNS operation is initiated, the tRNS starts the downlink data processing and transmission from the first forwarded GTP-PDU.

After the GTP tunnel is created between the tRNS and the SGSN, the uplink flow is switched from the old path to the new path.

- **Phase III: Releasing resource reservation in the old path**

For correct resequencing, the sRNS and the SGSN should forward the *Sequence Number* information respectively to the tRNS as defined in Release 99, so that the tRNS can judge whether the data forwarding has been completed or not. This requires triggering the GTP Sequence Number field in the GTP header for each video packet.

When the MS is successfully communicating with the tRNS and after the data forwarding is complete, an HO-complete message (signal #9) will be sent by the MS to the tRNS. The tRNS will then send an HO-complete message (signal #10) to the SGSN. After the SGSN has received the signal #10 from the tRNS, it shall begin to release the resources reserved on sRNS for the MS in the old path. In Figure 2-4 the resources are released by using the Clear-command message (signal #11) and Clear-complete message (signal #12).

On the data plane of handoff phase III, at the beginning of the releasing phase, the downlink media flow is redirected from the old path to the new path. The tRNS should store the redirected data until the transmission of all the forwarded data to the MS is completed, such that the correct packet sequencing can be ensured. The functionality of resequencing is implemented in the BSs.

2.3.2.2 Inter-RAN Handoff Procedure

Figure 2-5 illustrates a typical UMTS Release 4 network model under the IP transport mode in the scenario of inter-RAN handoff. Similarly, the media service provider is outside the CN which consists of SGSNs and GGSN, and far from the MS. The media streams should first get through CN and then feed into the Domain A or B.

Note that the handoff latency gets worse since the data forwarding has to take place between the RNSs which are separated by a transport network.

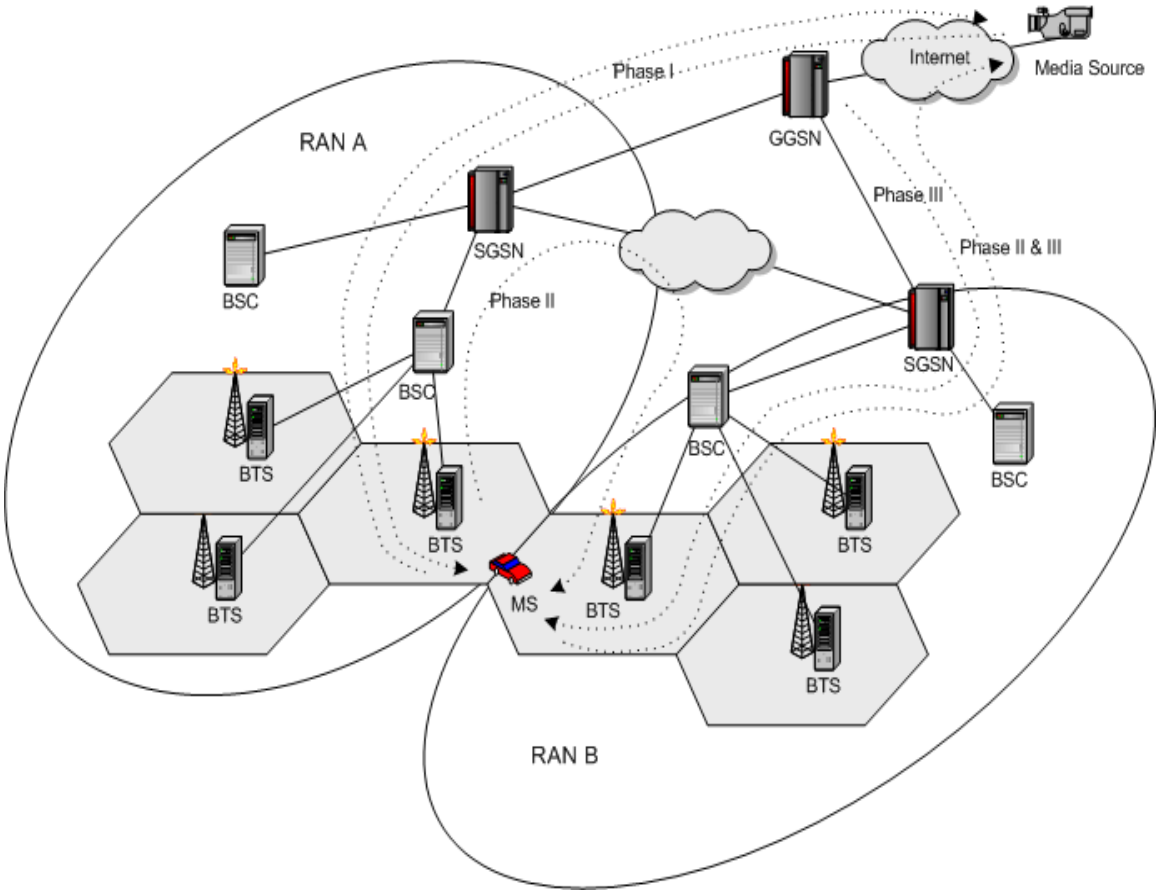


Figure 2-5 UMTS Rel4 network model of inter-RAN handoff (Data plane)

The control plane of handoff procedure [57] [58] [59] also consists of three phases, as shown in Figure 2-6 and Figure 2-7, while the data plane of these three phases is shown in Figure 2-5. The differences are described as follows in contrast with the intra-RAN handoff.

- **Phase I: Preparation of RNS handoff and resource allocation**

The sRNS in RAN A informs the target SGSN (tSGSN) about the MS which attempts to make a handoff to the tRNS in RAN B. The source SGSN (sSGSN) sets up a link (i.e., GTP tunnel) to the tRNS through the tSGSN, and requests the allocation of resources for the MS.

For the data plane of handoff phase I, at the end of the preparation phase, the sRNS stops transmitting downlink data to MS and should store all downlink data which continue to arrive from the sSGSN to the source RNC.

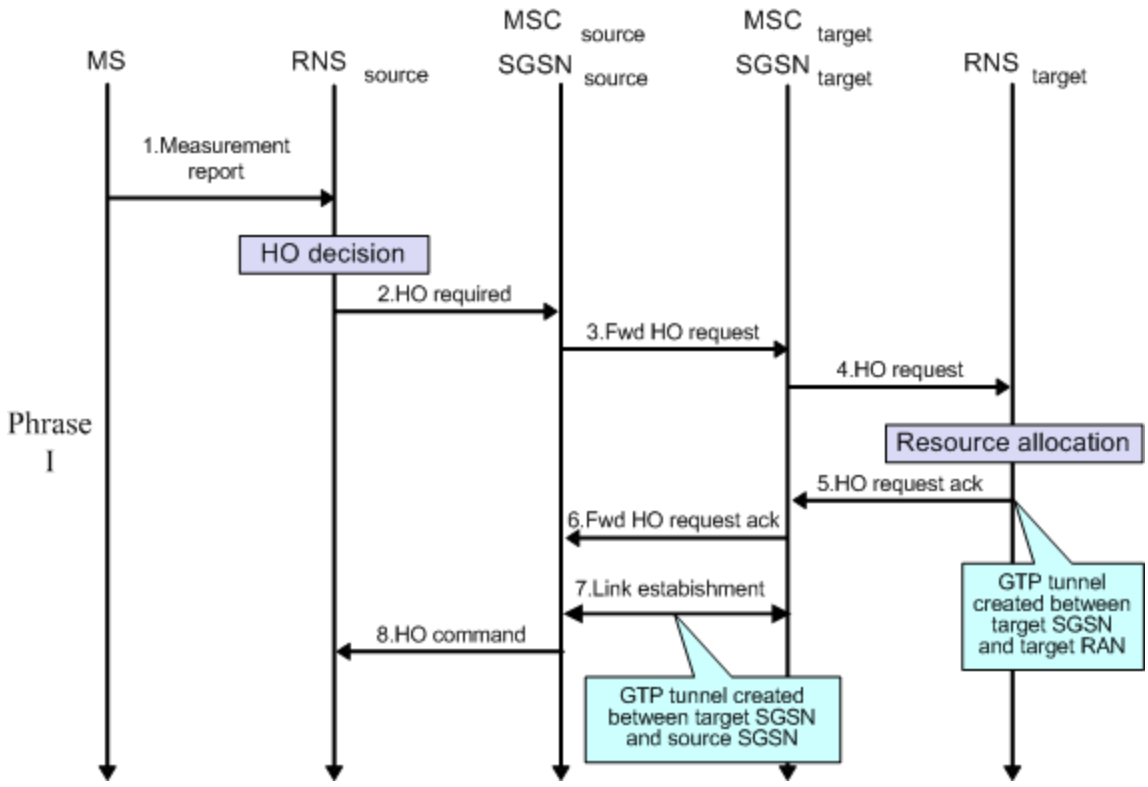


Figure 2-6 Inter-RAN handoff procedure in UMTS (Control plane: Phase I)

- **Phase II: Moving the serving RNS role to target RNS**

Under the handoff command from the sRNS, the MS access to the tRNS. Meanwhile, the downlink and uplink GTP tunnel is updated between the tSGSN and the GGSN through Update-PDP-context-request message (signal #12) and Update-PDP-context-response message (signal #13), so that the downlink and uplink flow can use the new route in the next phase.

On the data plane of handoff phase II, at the beginning of the execution phase, the sRNS starts to forward all the buffered data including state information for session migration to the tRNS, via a GTP tunnel between the RNSs. When data forwarding starts, the tRNS stores all GTP-PDUs forwarded from sRNS. When serving RNS operation is initiated, the tRNS starts the downlink data processing and transmission from the first

forwarded GTP-PDU. After updating of the uplink GTP tunnel, the uplink flow is switched from the old path to the new path.

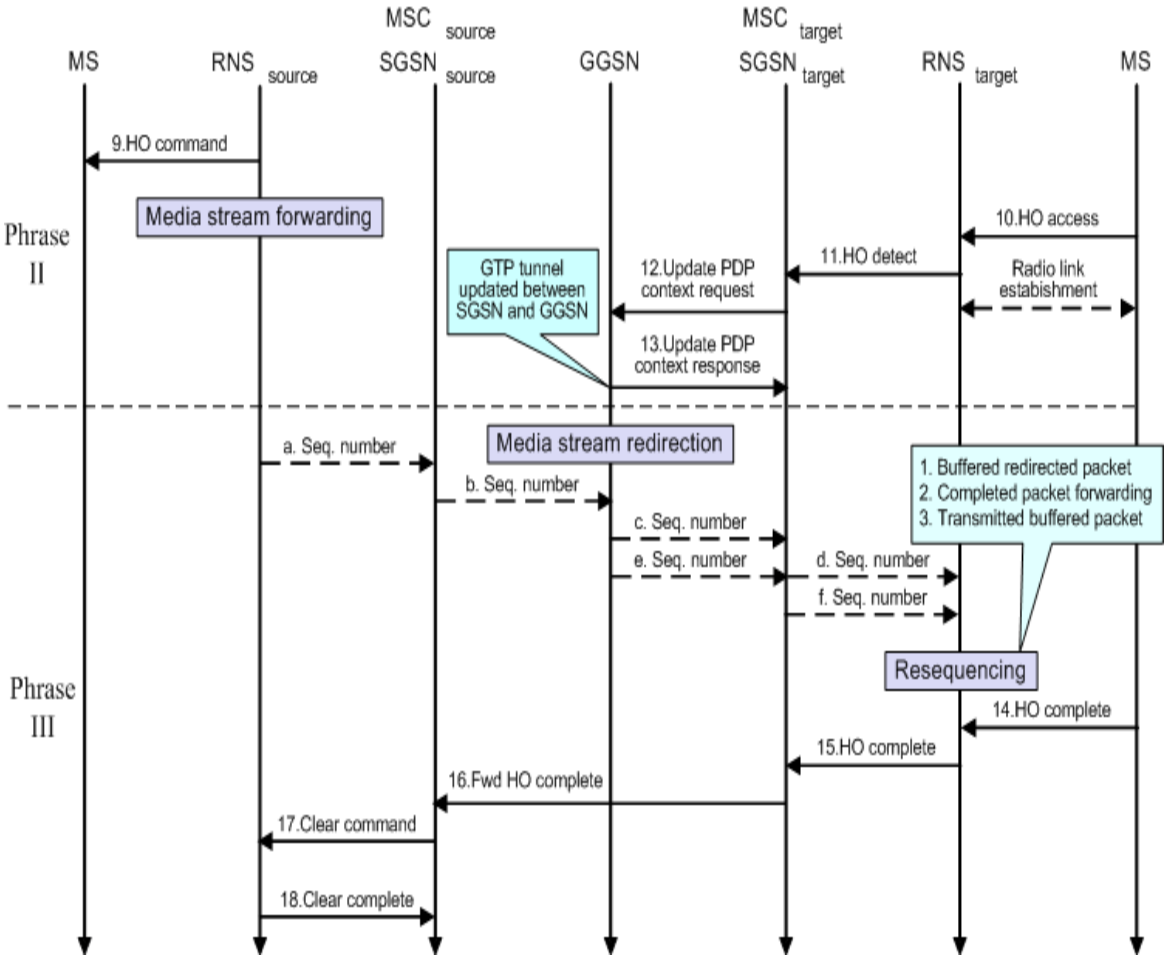


Figure 2-7 Inter-RAN handoff procedure in UMTS (Control plane: Phase II&III)

- **Phase III: Switching of downlink flow in CN**

The handoff is completed and the sRNS then releases the resources reserved for the MS in the old path.

For the data plane of handoff phase III, at the beginning of the path-optimization phase, the downlink media flow is redirected from the old path to the new path. The tRNS should store the redirected data until the transmission of all the forwarded data to the MS is completed, such that the correct packet sequencing can be ensured.

For correct resequencing, the sRNS and the GGSN should forward the Sequence Number information respectively to the tRNS as defined in Release 99, so that tRNS can

judge whether the data forwarding has been completed or not. This requires triggering the GTP Sequence Number field in the GTP header for each video packet.

Note that: The mechanism shown in Figure 2-5 assumes that the downlink GTP port used for a given media stream in tRNS is the same for all arriving GTP-PDUs regardless of their arrival routes.

2.4 Handoff problems in GPRS/UMTS

There are several problems or requirements as follows associated with media streaming during handoff.

2.4.1 Bearer Service QoS and Seamless Handoff

Handoff management for streaming applications is the process of initiating and ensuring a *seamless* handoff, in which the radio access network changes the radio transmitters or radio access mode or radio system used to provide the bearer services, while maintaining a defined bearer service QoS. "Seamless handoff" means a handoff without perceptible interruption of the radio connection according to the definition given in 3GPP in [37]. For seamless handoff, it assumes that there is no need to buffer any downlink or uplink traffic in the involved nodes considering that packet loss (or frame loss) is tolerated to some degree in the streaming application.

Because of the limitation of the cost and the physical size, a mobile handset generally can afford only limited buffer size. It is liable to extend the playout buffer in base stations. Also due to the mismatch between high transmission rate over wired links and low transmission rate over wireless links, the packet buffer in base stations can be used for rate matching and packet resequencing. When handoffs occur, the buffered data are then forwarded from the sRNS buffers to the tRNS.

In addition, packet loss can already occur over the radio or due to congestion in the wired link. From Table 1-1, the BER of wireless video can be up to 10^{-3}. Therefore any packet loss due to handoff is in addition to the packets lost over the radio or in the wired link.

In order to maintaining a defined bearer service QoS ($BER \leq 10^{-3}$) in Table 2-1, a data buffering and forwarding mechanism in UMTS R99 should be reused also in UMTS Release 4 for streaming services requiring seamless handoff.

The values in Table 2-1 are indicative of the QoS attributes required by audio and video media streams in 3GPP [61], including BER and frame erasure rates (FER), for the Adaptive Multi Rate (AMR) speech codec and the MPEG-4 video codec as examples.

However, the handoff procedures in UMTS R99 and Release 4 under above considerations may not satisfy the requirements of transfer delay, media stream interruption, signalling traffic and scalability, as follows.

Table 2-1 QoS attributes required by audio and video media streams in 3GPP

Type of payload	QoS attributes
AMR speech codec payload	Bit rate: 4.75 ~ 12.2 kbit/s
	Delay: end-to-end delay not to exceed 100ms (codec frame length is 20ms)
	BER: 10^{-4} for Class 1 bits 10^{-3} for Class 2 bits For some applications, a higher BER class (~10^{-2}) might be feasible.
	FER < 0.5% (with graceful degradation for higher erasure rates)
MPEG-4 video payload	Bit rate: variable, average rate scalable from 24 to 128 kbit/s and higher
	Delay: end-to-end delay between 150 and 400ms video codec delay is typically less than 200 ms
	BER: 10^{-6} - no visible degradation 10^{-5} - little visible degradation 10^{-4} - some visible artefacts 10^{-3} - limited practical application
	Packet loss rate: for further study

2.4.2 Transfer Delay

Transfer delay (i.e., end-to-end delay of media service delivery) is used to specify the delay tolerated by the media application. It allows UTRAN to set transport formats and Automatic Repeat reQuest (ARQ) parameters. For interactive media services, it should be minimized. For streaming audio/video, the transfer delay depends on the playout buffer length in the Internet. However, as presented in Table 2-1, the transfer delay of UMTS bearer service was bounded stringently in 3GPP. In the UMTS network model, shown in Figure 2-5, the media providers are separated from the UTRAN by the CN. The media streams should first get through CN and then feed into the Domain A or B. Thus, the transfer delay requirement may not be satisfied under the current model and should be studied further.

2.4.3 Handoff Latency (Media Stream Interruption)

Handoff latency is defined as time between the last packet transmitted from the old base station (BS) and the first packet transmitted from the new BS. As mentioned previously, buffered data forwarding during seamless handoff may result in relatively large handoff latency (or media stream interruption) and a large amount of additional traffic.

In the case of inter-RAN handoff, for downlink media streams, there are two possible situations wheren media stream gap or overlapping may happen:

1. The media stream overlap/gap may be introduced when tRNS takes the serving RNS role and starts to produce the downlink data from forwarded GTP-PDUs.

In this case the estimated gap/overlap for hard handoff is equal to the delay of the GTP tunnel used for data forwarding. This first instance of media stream overlap coincides with radio hard handoff.

If the transport bearer delay difference is smaller than the air interface Transmission Time Interval (TTI) (10, 20, 40 or 80 ms depending on the service), the amount of gap/overlap most likely does not exist.

2. The additional media stream gap may be introduced when the CN transport is optimized.

In this case the gap will exist only if the delay via the optimized route is larger than the delay via the forwarding route.

The above two types of media stream overlap/gap can also happen during a soft handoff. Due to the existence of the Iur interface (i.e., the interface between sRNC and tRNC), the first type of overlap/gap during soft handoff should be smaller in theory. However, note that the Iur interface is only logical interface, which may be provided via a transport network. Thus, in the real world, the handoff latency may be worse due to the introduction of the transport network between RNCs.

The effects of media stream interruption get worse when the network heterogeneity and the receiver heterogeneity are considered. Usually, video transcoders [33] have to be deployed at the edge of networks, such as at base stations, in order to address the heterogeneity problems. However, it incurs the extra state information migration besides data forwarding and the temporal dependence issue because of the nature of predictive encoding and motion compensation.

Temporal Dependence

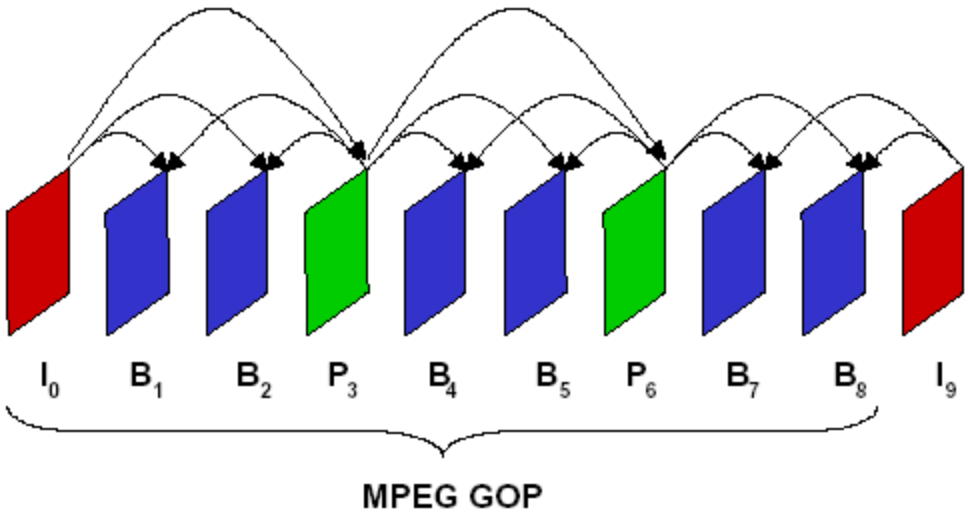


Figure 2-8 Temporal dependency in a MPEG-based coding stream

MPEG coding is based on both inter-frame and intra-frame coding, which produces three types of compressed frames, shown in Figure 2-8 as a group of pictures (GOP) in a frame-based MPEG coding stream:

- I-frame: Intra-coded frame, coded independently of all other frames.
- P-frame: Predictively coded frame, coded with reference to a previous I-frame or P-frame.
- B-frame: Bi-directionally predicted frame, coded with reference to both previous I-frame and future P-frame.

During the period of combing a group of blocks (GOB) into a medium unit, the packet loss caused by handoff makes the combination and decompression incomplete if the medium unit is split into different transmitted packets. If the information-lost part due to handoff belongs to I-frame I_0, and then the whole media sequence becomes erroneous, which results in a period of dummy video accompanied with some strange audio. Similarly, if the information-lost part due to handoff belongs to P-frame P_6, B-frames B_4, B_5, B_7, B_8 will be erroneously decoded. However, the damage of B-frames does not affect the decoding of any other-type frames in the GOP. Thus, the incomplete combination of a medium unit makes a multimedia presentation interrupted, asynchronous and discontinuous, and results in larger handoff latency.

State Information Migration

Typically, three types of state information should be reliably migrated for a successful handoff of a media stream:

1) The session description:

The session description is typically present at the beginning of the video stream as part of specific header information. It is possible to cache it on the old BS and transfer it from there to the new BS. Sometimes this information is provided as part of the control handshake, for example, using the Session Description Protocol (SDP).

2) The session parameters at the handoff decision point:

The session parameters at the handoff decision point include specifications from the clients request, the current position (offset) in the video stream, and some way of locating the video object, e.g., via a Uniform Resource Locator (URL). The media server should preferably have the ability to seek a specified stream offset.

3) The transcoding state information:

There are two types of state information associated with the transcoding process: reconstructible state information and dependent state information. Reconstructible state information can be recreated given an input stream. It consists of reference frame data (both original and down-sampled ones) and macroblock-level side information. Dependent state information includes data derived from the output stream. For example, the rate control module takes the number of bits consumed so far to evaluate the bit budget that can be allocated for the next coding unit. The volume of reconstructible state information is usually much larger than that of the dependent state information. In general, a transcoder needs to maintain and communicate at least dependent state information for a session handoff since the output stream is not shared between the transcoders in different BSs. Table 2-2 [10] gives the amount of data required for the transcoding state transfer, where CCIR refers to Consultative Committee for International Radiocommunication, CIF refers to Common Interleaved Frame, and QCIF refers to Quarter Common Interleaved Frame.

Table 2-2 The required amount of transcoding state information to be transferred

Source Format	Resolution (pixels)	Reference Frame Data (bytes)	Macroblock-level side information (bytes)	Total Transferred*	
				→CIF (bytes)	→QCIF (bytes)
CCIR 601	720 × 480	2,073,600	648,000	3,352,064	2,895,872
CIF	352 × 288	608,256	190,080	-	972,608
QCIF	176 × 144	152,064	47,520	-	-

* The total transferred includes fixed overheads.

The data forwarding has to wait until the migration of the above state information totally completes, which results in a larger handoff latency.

The handoff protocols in UMTS R99/Rel 4 are based on measuring the signal's quality to determine the time and place for initiating the handoff procedures, which do not consider the temporal dependence issue and the extra state information migration.

In [37], the concept of "glue-point" was proposed for resolving the temporal dependence issue. Handoff can only occur at glue-point so that the transcoding state information can be minimized. The glue-point delimits the boundary of two consecutive medium units (e.g., GOB in H.263 or Video Object Planes (VOP) in MPEG-4) or some type of boundary that is relatively less temporally dependent.

However, this solution will not work if the mobile terminal suddenly loses contact with the current base station before glue-point arrives, due to deep deterioration of the wireless channel condition and the high speed of the mobile user.

2.4.4 Media Synchronization

Media synchronization [5] refers to maintaining the temporal relationships within one data stream and between various media streams. There are three levels of synchronization, namely, intra-stream, inter-stream, and inter-object synchronization. The three levels of synchronization correspond to three semantic layers of multimedia data as follows [1].

1) Intra-stream synchronization

The unit of the compression layer of the MPEG encoder is a logical data unit such as a video/audio frame, which adheres to strict temporal constraints to ensure acceptable user perception at playback. Synchronization at this layer is referred to as intra-stream synchronization, which maintains the continuity of logical data units. Without intra-stream synchronization, the presentation of the stream may be interrupted by pauses or gaps.

2) Inter-stream synchronization

The unit of the synchronization layer is of the MPEG encoder a whole stream. Synchronization at this layer is referred to as inter-stream synchronization, which maintains temporal relationships among different continuous media. Without inter-stream synchronization, skew between the streams may become intolerable. For example, users could be annoyed if they notice that the movements of the lips of a speaker do not correspond to the presented audio (lip synchronazition).

3) Inter-object synchronization

In MPEG-4, the spatio-temporal location of audio-visual objects is defined by scene description using a tree-based structure in the Compress Layer. Synchronization according to this tree-based structure is referred to as inter-object synchronization. The objective of inter-object synchronization is to start and stop the presentation of the time-independent data within a tolerable time interval, if some previously defined points of the presentation of a time-dependent media object are reached. Without inter-object synchronization, for example, the audience of a slide show could be annoyed if the audio is commenting one slide while another slide is being presented.

The media stream interruption introduced due to handoff is typically unpredictable. The incurred delays and delay variations can disrupt intra-media, inter-media, and inter-object synchronization. Therefore, media synchronization mechanisms are required to ensure proper rendering of the multimedia presentation at the client.

For the purpose of media synchronization, the playout buffer in either BS or MS is deployed to eliminate the side effects that result from the wired network jitter or handoff latency. Due to the limitation of playout buffer, however, the buffered media units may be used up if the wired network jitter or handoff latency exceeds the expected value. The synchronization strategies thereby are proposed as follows when the expected media units are not available at the expected presentation time.

Nonblocking Strategy

If an expected video unit does not arrive at the expected time and the playout buffer in MS is not empty, the expected one is considered lost and thus ignored; continues to the next one. If an expected video unit does not arrive at the expected time and the playout buffer in MS is empty, the most recently displayed video unit is repeated until the video stream is available again.

Blocking Strategy

If an expected audio unit does not arrive at the expected time and the playout buffer in MS is not empty, the expected one is considered lost and thus ignored; continues to the next one. If an expected audio unit does not arrive at the expected time and the playout buffer in MS is empty, block the current presentation until the audio stream is available again.

2.4.5 Handoff Scalability

Typically, there are four possible handoff scenarios [55] [68] in GPRS/UMTS, as illustrated in Figure 2-9.

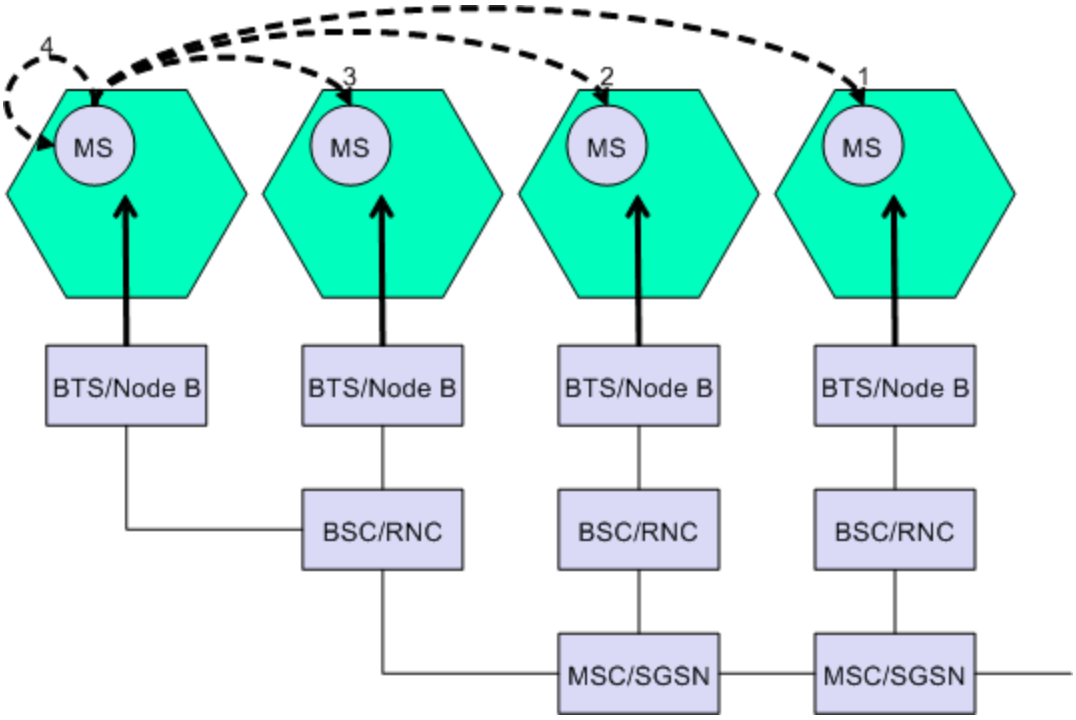


Figure 2-9 Types of handoff in GSM/GPRS/UMTS

1. Inter-RAN handoff: The calls are transferred between two cells belonging to different MSCs/SGSNs. Both MSCs/SGSNs perform the handover together.
2. Intra-RAN handoff: The calls are transferred between two cells belonging to different RNSs with the same MSC/SGSN. This handoff then has to be controlled by the MSC/SGSN.
3. Inter-cell, intra-RNS handoff: The calls are transferred between two cells but stays within the control of the same BSC/RNC. The BSC/RNC then performs a handoff, assigns a new radio channel in the new cell, and releases the old one.
4. Intra-cell handoff: The calls are transferred within the same cell. The BSC/RNC makes the handoff decision.

In UMTS, the types of handoff can also be classified into:

1. Inter-system handoff, between cells belonging to different radio access technologies (e.g., UMTS and GSM/EDGE) or different radio access modes (e.g., FDD/WCDMA and TDD/TD-CDMA).
2. Intra-system handoff, which can be further subdivided into:

- Intra-frequency handoff, between cells belonging to the same WCDMA carrier;
- Inter-frequency handoff, between cells operating on different WCDMA carriers.

In addition, UMTS supports both hard handoff and soft handoff. The soft handoff is fully performed within UTRAN, without involving the core network due to the existence of Iur interface. The hard handoff may be also performed within UTRAN or GERAN, or between GERAN and UTRAN, or the core network may be involved if the Iur or Iur-g interface between RNSs does not exist. Note that the Iur interface is only logical interface, which may be provided via a transport network. Thus, in the real world, the soft handoff latency may be worse due to the introduction of the transport network between RNCs.

Usually for streaming video services, video sessions are long-term sessions. During a long-term video presentation, mobile users may have to experience all or parts of the above scenarios. It is important to maintain relatively consistent handoff latency (delay jitter) in all the above scenarios so that there are no perceptible video or audio quality fluctuations during such a presentation. The handoff procedures are supposed to be specially designed to meet this challenge.

In summary, this chapter overviews the GPRS/UMTS mobile network architecture and the corresponding handoff protocols. The problems and requirements in UMTS associated with media streaming during handoff, such as bearer service QoS requirement and seamless handoff, transfer delay, handoff latency, media synchronization and handoff scalability are discussed in details.

3 System Model

Before the description of the proposed mobile system model and the protocol stack of network-aware end system, the concepts of layered coding and multiple description coding will be introduced first.

3.1 Layered Coding

From a video-coding point-of-view, *scalability* plays a crucial role in delivering the best possible video quality over unpredictable "best-effort" networks or time-varying wireless channels. From a networking point-of-view, scalability is needed to enable a large number of users to view any desired video stream, at anytime, and from anywhere. So far, *layered coding* with unequal error protection is the most popular and effective scheme for facilitating error resilience in a video transport system.

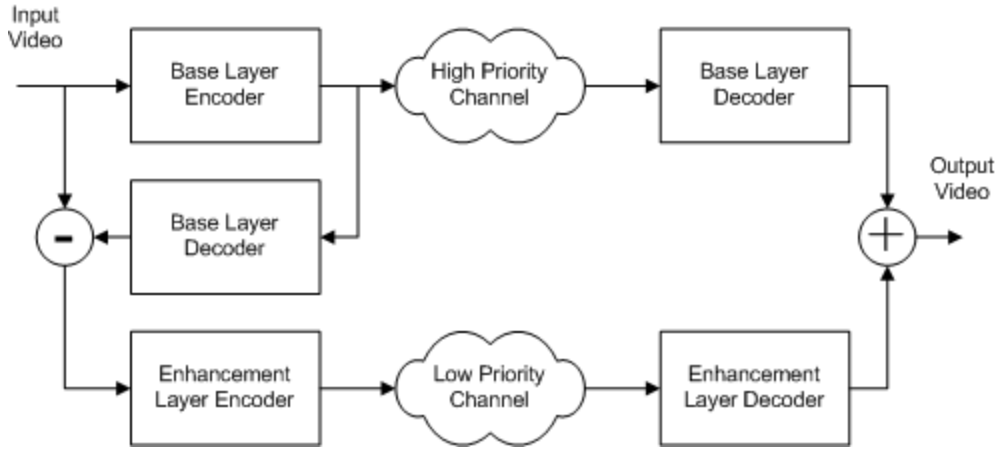


Figure 3-1 Block diagram of layered coding with transport prioritization

Principle of layered coding

Layered coding produces a hierarchy of bitstreams, where the first or *base layer* is coded independently at a coarser but acceptable level of quality, and subsequent *enhancement layers* are coded dependently. Each enhancement layer of the hierarchy can

increase the frequency, spatial, and temporal resolution over that of the previous layer and incrementally improve the quality. Figure 3-1 shows the block diagram of a generic two-layer coding and transport system.

Furthermore, layered coding has inherent error-resilience benefits, particularly when the base-layer bitstream can be transmitted with higher priority, guaranteeing a basic quality of service, and the enhancement-layer bitstreams can be transmitted with lower priorities, refining the quality of service. This approach is commonly referred to as layered coding with transport prioritization. By itself, layered coding is a way to enable users with different bandwidth capacity or decoding powers to access the same video at different quality levels. Therefore, layered coding is also called *Scalable Coding*. Unequal Error Protection will be discussed later.

Implementation mechanisms

Basically, there are four scalable mechanisms (*data partitioning*, *temporal scalability*, *SNR scalability*, and *spatial scalability*) depending on the way the video information is partitioned.

1) Data partitioning (Frequency domain partitioning)

In transform or subband based coding, the coder can include the low-frequency coefficients or low-frequency band subsignals in the base layer while leaving the high-frequency signal in the enhancement layer.

2) Temporal scalability (Spatial resolution refinement)

Temporal scalability is a technique to code a video sequence into two layers at the same spatial resolution, but different frame rates. The base layer is coded at a lower frame rate. The enhancement layer provides the missing frames to form a video with a higher frame rate. Coding efficiency of temporal scalability is high and very close to nonscalable coding.

3) Signal-to-noise ratio (SNR) scalability (Successive amplitude refinement)

SNR scalability is a technique to code a video sequence into two layers at the same frame rate and the same spatial resolution, but different quantization accuracy. The base layer can also encode the DCT coefficients of each block with a coarser quantizer, leaving the fine details (the error between the original and the coarsely quantized value) to be specified in the enhancement layer. A higher accuracy DCT coefficient is obtained

by adding the base-layer reconstructed DCT coefficient and the enhancement-layer DCT residue.

4) Spatial scalability (Spatial resolution refinement)

Spatial scalability is a technique to code a video sequence into two layers at the same frame rate, but different spatial resolutions. The base layer is coded at a lower spatial resolution and the enhancement layers contain additional information for obtaining higher spatial resolution. At the decoder, the reconstructed base-layer picture is up-sampled to form the prediction for the high-resolution picture in the enhancement layer.

Fine granularity scalability (FGS)

To provide more flexibility in meeting different demands of streaming (e.g., different access link bandwidths and different latency requirements), a new scalable coding mechanism, called fine granularity scalability (FGS), was proposed to MPEG-4 [6] [7] [8]. An FGS encoder also compresses a raw video sequence into two substreams, i.e., a base layer bit-stream and an enhancement bit-stream. Different from an SNR-scalable encoder, an FGS encoder uses bitplane coding to represent the enhancement stream. With bitplane coding, an FGS encoder is capable of achieving *continuous rate control* for the enhancement stream. This is because the enhancement bit stream can be truncated anywhere to achieve the target bit-rate.

A variation of FGS is *Progressive Fine Granularity Scalability* (PFGS) [17] which is developed by Microsoft Research. PFGS shares the good features of FGS, such as fine granularity bit-rate scalability and error resilience. The essential difference between FGS and PFGS is that FGS only uses the base layer as a reference for motion prediction while PFGS uses multiple layers as references to reduce the prediction error, resulting in a higher coding efficiency. PFGS is adopted in our SMDC approach.

Unequal Error Protection

To serve as an error resilient tool, layered coding must be paired with UEP in the transport system, so that the base layer is protected more strongly, e.g., by assigning a more reliable sub-channel, using stronger FEC codes [21] [53] , or allowing more retransmissions.

Different networks may implement transport prioritization using different means. In ATM networks, there is one bit (CLP) in the ATM cell header that signals the cell loss

priority. When traffic congestion occurs, a network node can choose to discard the cells having low priority first. Transport prioritization can also be implemented by using different levels of power to transmit the substreams in a wireless transmission environment. Also, DiffServ will be well suitable for prioritized transmission.

Advantages

1) It is highly adaptable to unpredictable bandwidth fluctuation due to dynamic changes in network conditions.
2) It provides the flexibility to combat both network heterogeneity and receiver heterogeneity. The scalable source coding leaves the media servers and proxies (or gateways) an opportunity to trim the video stream to appropriate bit rate before transmission or relay.
3) It is applicable to both unicast and multicast.
4) It provides error resilience through UEP, which is a nice match for future DiffServ networks.
5) There is no feedback channel requirement and therefore lower delay.
6) It is well suitable to combine with an encryption mechanism to support multiple levels of security for intellectual property protection [32].

Disadvantages

1) It will lead to a disastrous effect in the decoded visual quality or even break down if a loss is in the base layer or the channel of the base layer fails.

The current Internet can not support priority service and has to rely on the conventional way (e.g., FEC or retransmission) to realize error-free transmission of the base layer. However, FEC-based approaches often suffer from dynamic network conditions. Retransmission-based approaches are not applicable in streaming applications when a back-channel is not available or when the transmission delay is not acceptable. In the case of multicast or broadcast, too much feedback creates a problem, called "feedback implosion" [49]. Moreover, the sender can not afford to honor independent retransmission requests from each receiver.

2) The significant improvement performance of a layered coder over a single-layer coder in the presence of channel errors is at the cost of a coding overhead.

Generally, the four scalability modes in MPEG—namely, data partitioning, temporal scalability, SNR scalability, and spatial scalability—have increasingly better error robustness in that order, but also an increasing coding overhead. To be more precise, data partitioning requires the least number of bits (only 1% more bits), while the spatial scalability has a better reconstructed image when there exist significant losses in the enhancement layer. SNR scalability and temporal scalability is in the middle on both scales. Table 3-1 [20] summarizes the required ratio of the base layer to the total bit rate and the highest packet loss rate at which the video quality is still considered visually acceptable. These results are obtained by assuming that the base layer is always intact during the transmission.

Table 3-1 Comparison of different scalability modes in MPEG-2

Coding Mode	Required base layer to total bit rate ratio	The maximum sustainable packet loss rate
One layer (MP@ML)	100%	10^{-5}
Data partitioning	50%	10^{-4}
*Temporal scalability	<50%	10^{-4}~10^{-3}
SNR scalability	<20%	10^{-3}
Spatial scalability	<20%	10^{-3}~10^{-2}

* The data of temporal scalability came from theoretical analysis.

3.2 Multiple Description Coding

As described in Section 3.1, if a data network were able to provide a preferential treatment to the packets of the base layer and transmit them in an error-free channel, layered coding would be almost the best solution. The Internet, however, usually does not look inside packets and discriminate; packets are dropped at random (e.g., drop-tail or RED) when congestion occurs.

The typical way to handle lossless transmission in a lossy network is to invoke many retransmissions or add a lot of redundancy via strong FEC. In jitter-sensitive streaming applications, however, it may not be feasible or cost effective. A better alternative is to

make do with whatever arrives upon first transmission and combat the transmission error from the source side. Multiple Description Coding is well competent for this challenge due to its inherent diversity attribute.

Principle of multiple description coding

As with LC, multiple description coding also codes a source into several sub-streams, known as *descriptions*, but the decomposition is such that the resulting descriptions are correlated and have similar importance. Any single description should provide a basic level of quality, and more descriptions together will provide improved quality. Figure 3-2 shows the block diagram of a generic two-layer coding and transport system.

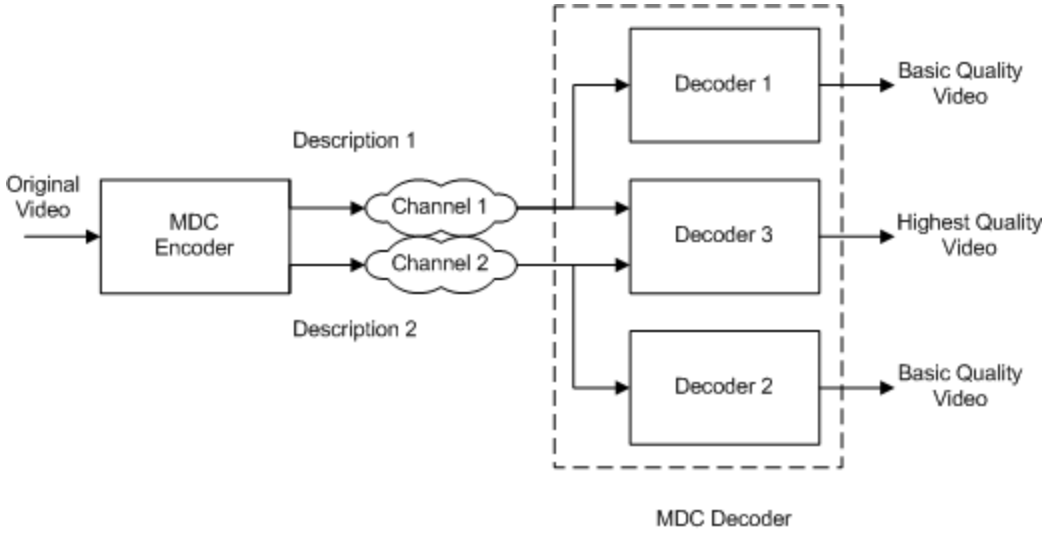


Figure 3-2 Block diagram of MDC coding and decoding

For each description to provide a certain degree of quality, all the descriptions must share some fundamental information about the source, and thus must be correlated. This correlation enables the decoder to estimate a missing description from a received one, and thus provide an acceptable quality level from any description. On the other hand, this correlation is also the source of redundancy in MDC.

An advantage of MDC over LC is that it does not require special provisions in the network to provide a reliable sub-channel. To accomplish their respective goals, LC uses a hierarchical, decorrelating decomposition, whereas MDC uses a non-hierarchical, correlating decomposition.

Implementation mechanisms

Some approaches [34] [49] that have been proposed for accomplishing such decomposition include *Multiple Description quantization*, *MDC with correlation transform*, *transform domain subsampling*, *spatial domain subsampling*, *temporal domain subsampling* (e.g., Multiple State Recovery (MSR) [46]), and *interleaved spatial-temporal sampling*. The last approach is known as *video redundancy coding* (VRC) in H.263+ [18]. MSR has excellent capability of error recovery in the presence of packet loss on all descriptions and is adopted in our SMDC approach.

Advantages

1) Robustness to losses and bit errors
    a. Path diversity

MDC directly attacks the problem of communicating the continuous-valued source. MD coders can be designed with concern for every combination of received descriptions with an appreciable probability. MDC assumes that there are several parallel channels between the source and the destination and that each channel may be temporarily down or suffering from long burst errors. Furthermore, the error events of different channels are independent, so that the probability that all channels simultaneously experience losses is small.

Diversity techniques have been studied for many years in the context of wireless communication, e.g., frequency, time, and spatial diversity. However, the problem of path diversity over a packet network has been largely unexplored. The recent work [26] adds justification to our approach for path diversity: in comparing the performance of the default path between two hosts on the Internet to that of alternative paths between those two hosts, it is found that "in 30-80% of the cases, there is an alternate path with significantly superior quality". The quality is measured in terms of round-trip-time, loss rate, and bandwidth. Therefore, diversity would also appear to be beneficial for communication over the Internet. There are several ways to set up multiple paths or links for a single virtual connection in a wireless network. In a single-hop wireless network, a station would need to establish channels to multiple base stations instead of one. This is already done in "soft" hand-off systems, during the hand-off phase. In a multiple-hop adhoc networks, each station has router-like functionality to establish multiple disjoint

paths with another wireless station. The Internet Engineering Task Force (IETF) MANET Working Group has been the main forum for research in this area. Most of the proposed ad hoc routing protocols have the ability to discover multiple routes. In a CDMA system, the multiple antenna technique can be employed with the MDC. Each description should use one transceiver and one antenna in the BS and the MS.

b. Error recovery

Some MDC algorithms have an excellent capability of error recovery in the presence of packet loss on all descriptions, such as multiple states and state recovery [46].

Due to robustness to losses and bit errors, it is well suitable for the imperfect and unpredictable channels that have relatively high loss or failure rates, such as streaming media channels over the current Internet or a wireless network. Also, it should be forward compatible with the potential DiffServ network.

2) Enhanced quality

If a receiver receives multiple descriptions, it can combine them together to produce a better reconstruction than that produced from any one of them.

3) Distributed storage

Distributed storage [49] of streaming media matches the MD framework well. Consider a database of images stored at several locations with MD encoding. A typical user would have fast access to the local image copies; for higher quality, one or more remote copies could be retrieved and combined with the local copy.

Distributed storage is common in the use of edge servers for popular content. In current implementations, identical data is stored at the servers, so there is no advantage in receiving multiple copies. Storage can also be distributed to make the reliability of each device less important; lowering reliability requirements can decrease costs.

4) No feedback channel requirement and therefore lower delay.

Disadvantages

1) Low coding efficiency

To guarantee an acceptable quality with a single description, each description must carry sufficient information of the original signal. This implies that there will be overlap in the information contained in different descriptions. Obviously, this will reduce the coding efficiency.

The core issue in designing MD coder is the tradeoff between coding efficiency and redundancy among the descriptions, such that the degradation in quality in the event of failures is graceful. A mechanism [19] is required to adapt the amount of redundancy added to the channel condition.

2) Unbalanced MD operation

The characteristics of each path in a packet network are different and time-varying, therefore the available bandwidth in each path may differ. This results in the requirement of unbalanced MD operation [26], where the bit rate of each description is adapted based on the available bandwidth along its path. We will discuss this issue in the next section.

3.3 Distributed Multimedia Delivery Mobile Network

3.3.1 Concept of Content Delivery Network

Content Delivery Network (CDN) [43] was developed to overcome performance problems, such as network congestion and server overload in the star-type network topology, which arise when many users access popular content. CDN that is distributed via a WAN generally consists of the origin server containing the content and a set of edge servers. Each edge server is located closer to users and stores a subset of the content or caches popular content. CDN provides a number of advantages.

1) It enhances server scalability and helps prevent server overload and network congestion.

Conventionally, content is delivered by the central content server to the entire network. The central server become a bottleneck of the networks and is lack of scalability under the limitation of bandwidth, storage and computational complexity in the content server. In the case of media multicast or broadcast, too much feedback creates a problem, called “feedback implosion”.

The numerical investigations in [12] indicated that, for typical scenarios, the revenue rate increases logarithmically with the cache space and linearly with the link bandwidth connecting the cache space to the central server. Thus, it is beneficial to establish the edge servers for caching before increasing the link bandwidth of the central server. With

the concept of distributed storage, CDN helps to prevent server overload and network congestion, since the replicated content can be delivered to users from edge servers.

2) It reduces the content delivery time for services, the probability of packet loss, and the total network resource occupation

As content is delivered from the closest edge server and not from the origin server, the content is sent over a shorter network path, thus reducing the content delivery time, the probability of packet loss, and the total network resource occupation.

3.3.2 Proposed Mobile Network Model

While CDN was originally intended for static web content, it can be extended to our approach for the delivery of streaming media as well. With the comprehensive consideration and integration of concepts of CDN and SMDC, a distributed multimedia delivery mobile network is proposed for the introduction of media streaming services in GSM/GPRS/UMTS. The proposed network model in the scenario of intra-RAN handoff and inter-RAN handoff are illustrated in Figure 3-3 and Figure 3-4 respectively, where DiffServ is employed to provision UEP for scalable video coding. The media streaming services are pushed to the edge of CN so that the content delivery time for services is significantly reduced by the media delivery network.

In each media delivery network, a set of complementary distributed *Media Description Servers* (MDSs) interact and collaborate with each SGSN or MSC for media delivery to mobile stations in the radio access network. Each MDS should keep one complementary description of media streams that were originally downloaded from the service provider during the streaming service publication.

In a real world of networks, such as GSM, the high-level streaming service control functions (e.g., service registration, publication and discovery, service subscription and binding, subscriber authentication and service allowance verification) can be peeled off from the MSC and be implemented in VLRs. It is cost-effective because the MSC becomes more efficient, does not waste cycles in processing new services, and simplifies new service development.

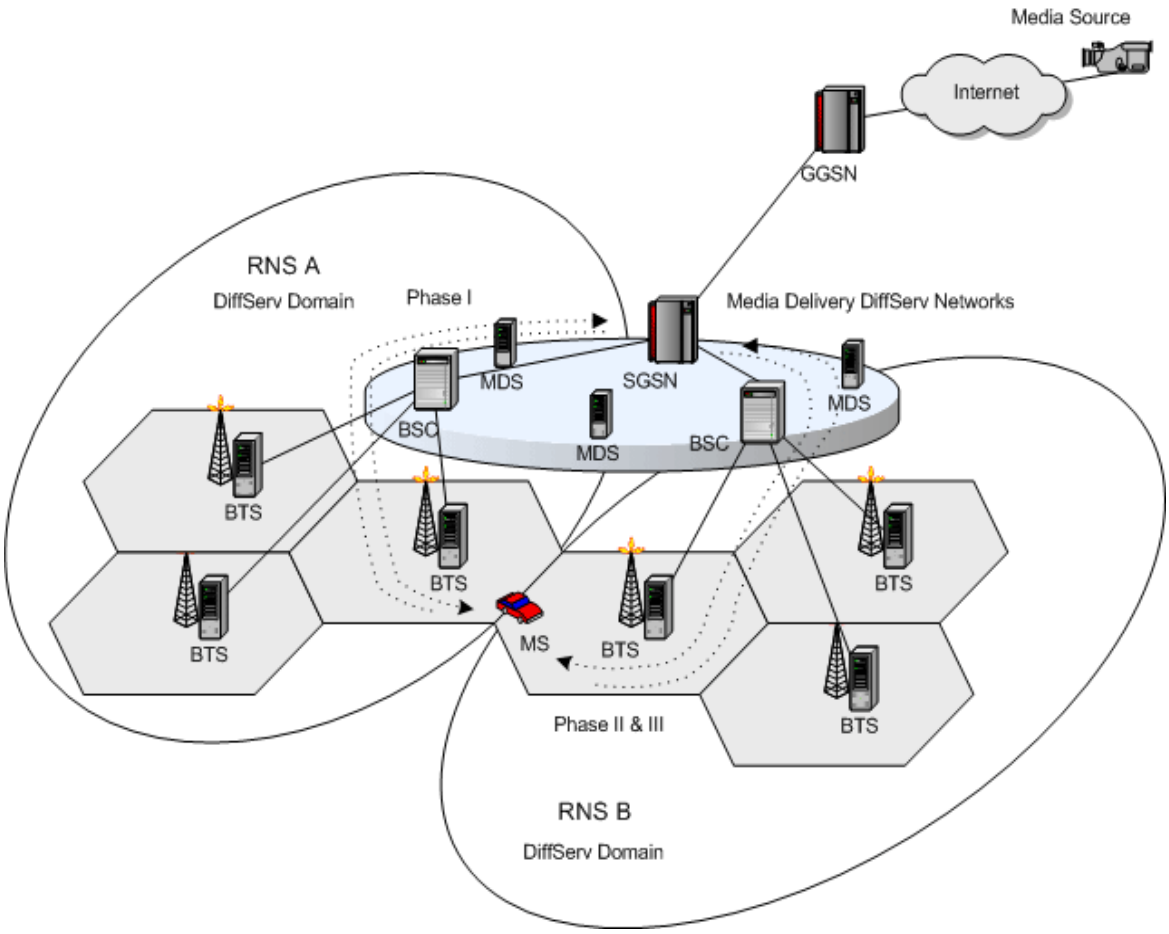


Figure 3-3 Proposed network model of intra-RAN handoff (Data plane)

The proposed distributed multimedia delivery mobile network helps to deal with the following problems:

1) Network congestion and server overload problems in the star-type network topology.
2) The media streaming handoff problems in the mobile networks.

The streaming media is delivered from the closest edge server and not from the origin server, the streaming media is sent over a shorter network path, thus reducing the media service delivery time (end-to-end delay), the probability of packet loss, and the total network resource occupation.

3) The high requirement of storage, reliability and load balancing among the distributed media edge servers and thus high cost of network components in the conventional CDN.

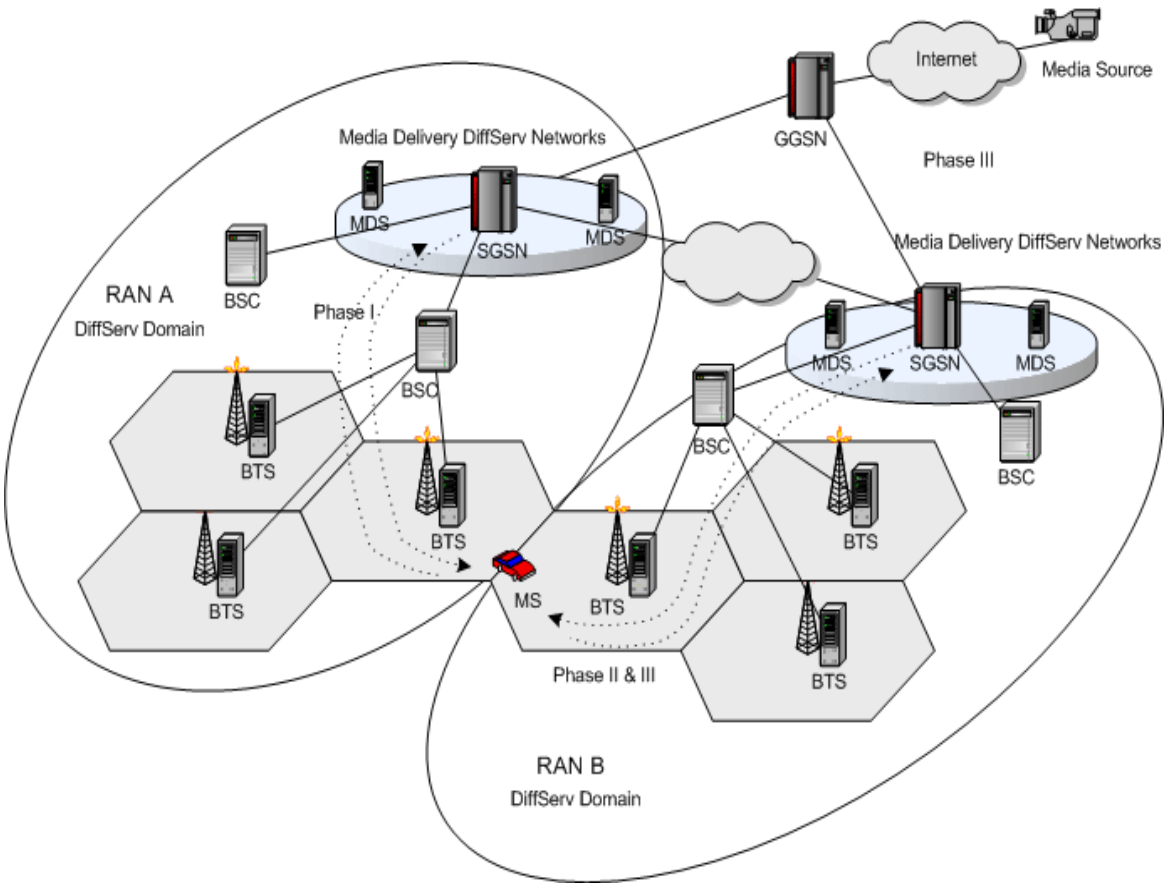


Figure 3-4 Proposed network model of inter-RAN handoff (Data plane)

One difference between CDN and D-MDMN which should be paid attention to is that load balancing is even more important for media edge servers than for web servers, since the resource commitment is typically larger and lasts longer.

Figure 3-5 shows a comparison between the single description (SD) and multiple description (MD) approach.

During the construction phase of conventional CDN where the SD approach is used, a redirection server (RS) should choose a set of media edge servers (ESs) that can achieve the best end-user performance for each edge router (ER). The ER acts as an entry of a wired access network. In Figure 3-5, ES 1 and ES 2 are supposed to be the best set for the client.

Usually, each ER has one RS to monitor and balance the traffic load among the set of ESs. Each edge server should keep one copy (SD) of media streams. Suppose ES 1 is overloaded during a media delivery, then RS should direct ES 2 which is underloaded to

continue the delivery of the rest part of that media stream. This incurs the server-side handoff problem [11] due to load balancing.

Similarly, during the phase of D-MDMN construction, each SGSN or MSC should choose a set of MDS that can get the best end-user performance through the simulation of system configuration. Each MDS should keep one description of media streams. In Figure 3-5, MDS 1 and MDS 2 are supposed to be the best set for the client (SGSN or MSC) and keep D1 and D2 of media streams, respectively.

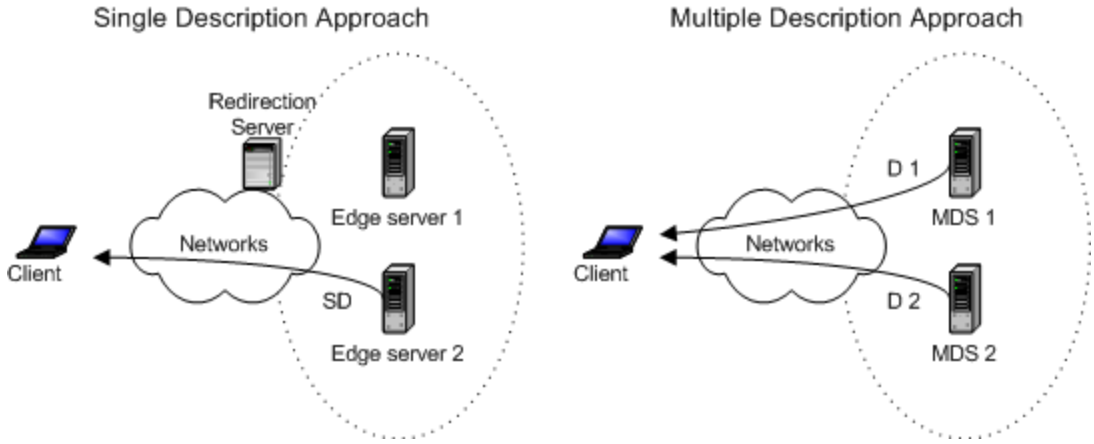


Figure 3-5 Single Description versus Multiple Description

With the employment of the MD approach, however, there is no or extremely less requirement of load balancing, especially when the proposed SMDC is used to combat the problem of unbalanced MD operation. MDS 1 and MDS 2 provide the delivery of different description of a media stream simultaneously. If MDS 1 is overloaded during media delivery, it can discard the appropriate enhancement layers to lower its load. If MDS 1 remains overloaded even after layer dropping, D2 still can be presented to the client at a tolerable quality.

MD approach also lowers the reliability requirement of MDS due to its loss tolerance capability. Suppose the reliability of a single description is 90%, then that of two descriptions will be 1-(1-90%)(1-90%) = 99%. Also, each SMDC description is almost only 56~60 % of the amount of a SD copy [47], which decreases the storage requirement of MDS. In addition, SMDC has a very strong capability of error recovery during media transmission.

A summary of comparison between the SD and MD approaches is given in Table 3-2.

Table 3-2 Single Description versus Multiple Description

	Single Description	Multiple Description
Storage	100 %	56~60 %
Reliability	90 %	99 %
Error recovery	Weak	Strong
Load balancing	Required	Not required

3.3.3 MPEG-4 Video Streaming over IP DiffServ

UMTS QoS Classes and IP DiffServ Classes

The UMTS specifications define in [61] four QoS classes: conversational, streaming, interactive, and background. The main distinguishing factor among these classes is delay sensitivity. The conversational class is the most sensitive, while background is the least sensitive. Conversational and streaming classes are intended for real-time traffic. They both preserve time relation (variation) between information elements of the stream, but conversational has stricter delay requirements. Example applications are IP telephony for the former and streaming video for the latter. For the interactive and background classes, transfer delay is not the major factor. Instead, they both preserve the payload content. The interactive class follows a request-response pattern and defines three priorities to differentiate bearer qualities, while it does not provide explicit quality guarantees. The main characteristic of the background class is that the destination does not expect the data within a certain time. Example applications are FTP or Web traffic for interactive and download of emails for background.

On the other hand, the IETF has also defined DiffServ mechanisms in RFC 2475 [45] for IP based networks aiming at QoS provisioning by means of Class of Service (CoS) approaches, which is well suitable for unequal error protection of video layered coding.

DiffServ architecture defines a simple forwarding mechanism, i.e., per-hop behavior (PHB) [39], at interior network nodes while pushes most of the complexity to network boundaries. Differentiated services are realized in RFC 3290 [54] by the use of particular packet classification and traffic conditioning mechanisms at boundaries coupled with the concatenation of per-hop behaviors along the transit path of the traffic.

The traffic conditioner is decoupled from the network interior and consists of marker, meter, shaper and policer (i.e., dropper). Marking is performed at the source host or the first-hop router administrative domain by means of mapping the DiffServ codepoint (DSCP) contained in the IP packet header to a PHB. A replacement header field, called the DS field, is defined in RFC [44], which is intended to supersede the existing definitions of the IPv4 TOS octet and the IPv6 Traffic Class octet. Six bits of the DS field are used as a DSCP to select the PHB that a packet experiences at each node.

In the packet forwarding path, per-hop behaviors are defined to permit a reasonably granular means of allocating buffer and bandwidth resources at each node among competing traffic streams. PHBs are expected to be implemented by employing queue management and scheduling on a network node's output interface queue. In DiffServ, three main PHB have been defined:

- Expedited Forwarding

Expedited Forwarding (EF) [41] provides a low delay, a low loss and an assured bandwidth similarly to CBR in ATM.

- Assured Forwarding

Assured Forwarding (AF) PHB group [39], similarly to nrt-VBR/ABR/GFR in ATM, is a means for a provider DS domain (i.e., Media Delivery DiffServ Network in Figure 3-4) to offer different levels of forwarding assurances for IP packets received from a customer DS domain (e.g., a RAN A or B in Figure 3-4). *N* independent AF classes are defined, where each AF class in each DS node is allocated a certain amount of forwarding resources (buffer space and bandwidth). IP packets that wish to use the services provided by the AF PHB group are assigned by the video provider in the Media Delivery DiffServ Network into one or more of these AF classes according to the services that the customer has subscribed to. All packets belonging to an AF class are admitted into one AF queue to avoid out of order delivery. Within each AF class, IP packets are

marked (again by the video provider DS domain) with one of *M* different levels of drop precedence. In case of congestion, the drop precedence of a packet determines the relative importance of the packet within the AF class. A congested DS node tries to protect packets with a lower drop precedence value from being lost by preferably discarding packets with a higher drop precedence value.

An IP packet that belongs to an AF class *i* and has drop precedence *j* is marked with the AF codepoint *AFij*, where $1 \leq i \leq N$ and $1 \leq j \leq M$. Currently, four classes ($N = 4$) with three levels of drop precedence in each class ($M = 3$) are defined in RFC 2597 for general use. More AF classes or levels of drop precedence may be defined for local use.

The queue management and scheduling mechanisms of AF PHBs are illustrated in Figure 3-6. The drop preferences within each class should be considered in the potential approaches of queue management, such as WRED. The drop precedence queue management can be implemented, for example, by using a leaky bucket traffic policer with one token rate and two bucket size, which can be decided according to the service level agreement (SLA).

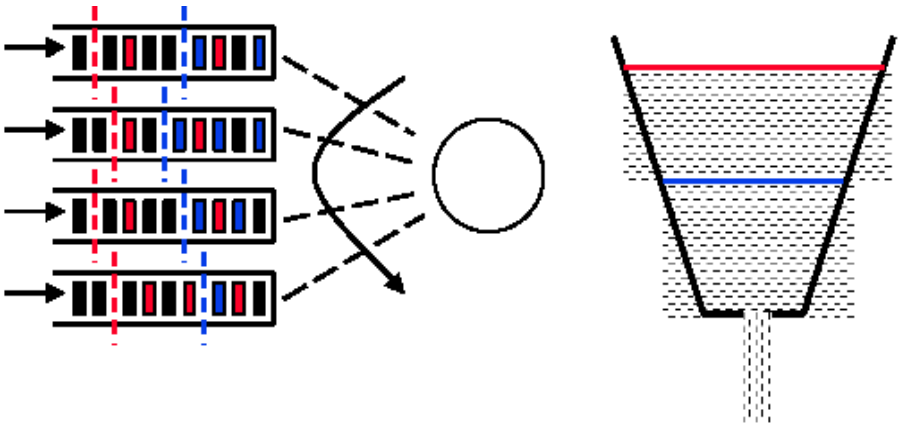


Figure 3-6 Queue management and scheduling mechanisms of AF PHBs

- Best Effort

Best Effort (BE) provides no QoS guarantee.

It is challenging to provide QoS attribute translation and mapping between the IP world and the UMTS world and to implement the IP differentiated services for the traffic encapsulated and isolated by tunneling in UMTS. This issue will be addressed in Section 4.2.1 and Section 4.2.2.

IP DiffServ and GTP/FP Tunnels in UMTS

The protocol stack in the IP transport mode for non-video services in UMTS is shown in Figure 3-7 (a) [47] [56] [62] [67] [69], while that for video services in the proposed D-MDMN is shown in Figure 3-7 (b), where the Uu, Iub, Iu, Gn and Gi interfaces are defined in 3GPP TS 25.401 V5.5.0 [74].

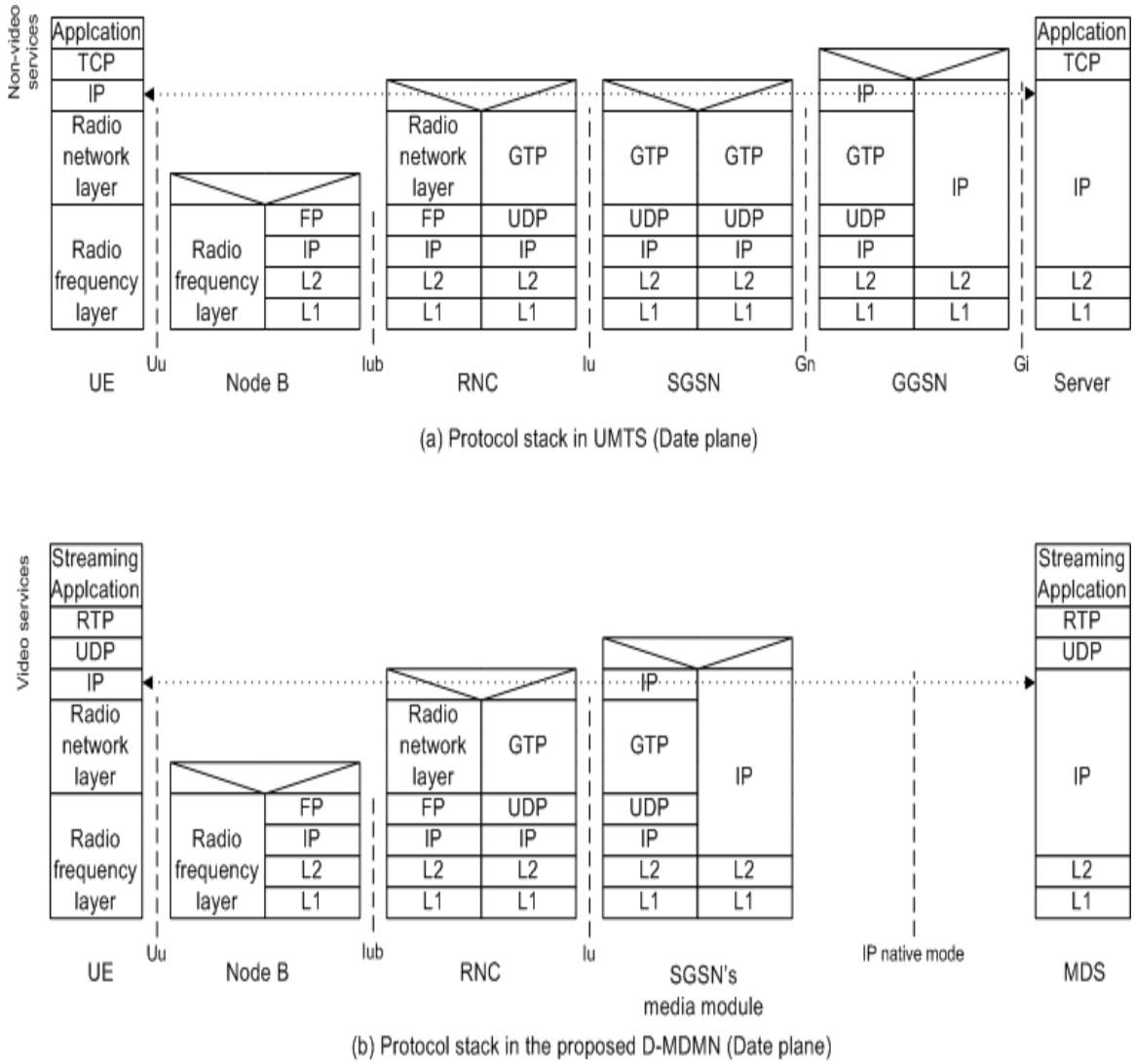


Figure 3-7 Protocol stack in UMTS and the Proposed D-MDMN (Data plane)

For non-video services, The uplink user-level packets are segmented by the user equipment (UE) into *Radio Network Layer* (RNL) frames, called transport blocks. These are carried over the *Radio Frequency Layer* (RFL), using W-CDMA access and modulation techniques, to the Node Bs within reach of the mobile. Each Node B

encapsulates a set of transport blocks into a single frame of the RNL Frame relay transport Protocol (FP) and forwards the frame to its RNC over the Iub interface.

The FP frames can be exchanged between the drifting and serving RNCs over the Iur interface. The serving RNC of the host is responsible for frame selection among the multiple received copies of the same transport block, processing the other sublayers of the RNL, and finally reassembling the user-level packet.

To deploy the IP transport mode on the Iub interface, the FP frames are encapsulated into IP packets. The destination addresses of these packets refer to the network components (i.e., RNC or Node B) and not the user's IP address. In fact, the host's IP address is never used for forwarding purposes in the UTRAN, the decisions being made on the basis of RNL specific protocols. The Mobile Wireless Internet Forum has specified further details [50] concerning the implementation of IP in the UTRAN in the transport mode.

In order to communicate with the data network, the mobile host needs to register with the CN by performing a GPRS attach operation. This results in the creation of two *GPRS Tunneling Protocol* (GTP) sessions, specific to that host: between the RNC and the SGSN on the Iu interface, and between the SGSN and the GGSN on the Gn interface. The user-level multi-protocol packets are allowed to be encapsulated into GTP frames and be forwarded between the RNC and the GGSN. The GTP protocol is implemented only by SGSNs and GGSNs. No other systems need to be aware of the GTP's presence. Mobile hosts are connected to an SGSN without being aware of GTP.

Upon the GPRS attachment, a mapping is created at the RNC between the host identity and the GTP session between the RNC and the SGSN. In addition, a record is created at the GGSN, which contains the mapping between the host's IP address and the GTP session with the corresponding SGSN.

To deploy the IP transport mode on the Iu and Gn interfaces, the GTP frames are encapsulated into IP packets. The destination addresses of these packets refer to the network components (i.e., RNC, SGSN, or GGSN) and not the host's IP address. Forwarding decisions are based on the GTP mapping tables in those nodes.

For video services, the only UMTS network component needed to be enhanced is the SGSN where a media module should be added to interconnect with IP-based media delivery network in the IP native mode. The reason not adopting the IP transport mode is

based on the overhead of GTP, UDP and another IP layer which has to be inserted between L2 and the original IP layer in all MDSs. The IP transport mode is kept, however, for non-video services. The deployment of the IP native mode between SGSN's media module and MDS leverages the evolution of UMTS towards its final all IP-based phase. The protocol stack of UDP/RTP will be discussed in details in Section 3.4.

As discussed above, the video IP packets generated from the video provider are supposed to pass two different tunnels in UMTS. One is GTP tunnel existing between GGSN and RNC in the CN. The other one is FP tunnel started from RNC and terminated at Node B in the RAN. As the video IP packets pass through the tunnel, there are additional headers (i.e., GTP or FP tunnel header) inserted between the two IP headers. The inner IP header is that of the original traffic with differentiated services; an outer IP header is attached and detached at tunnel endpoints without differentiated processing. In general, intermediate network nodes between tunnel endpoints operate solely on the outer IP header, and hence DiffServ-capable intermediate nodes access and modify only the DSCP field in the outer IP header. Thus, it is a challenging design issue [40] to implement DiffServ in the mobile network for the traffic which is encapsulated by the tunnels. This issue will be addressed in Section 4.2.3.

3.4 Protocol Stack of End Systems

One main design goal of the network-aware end systems is to extend the applications of real time streaming protocols in the Internet to wireless networks. The protocol stack of streaming server and client is shown in Figure 3-8.

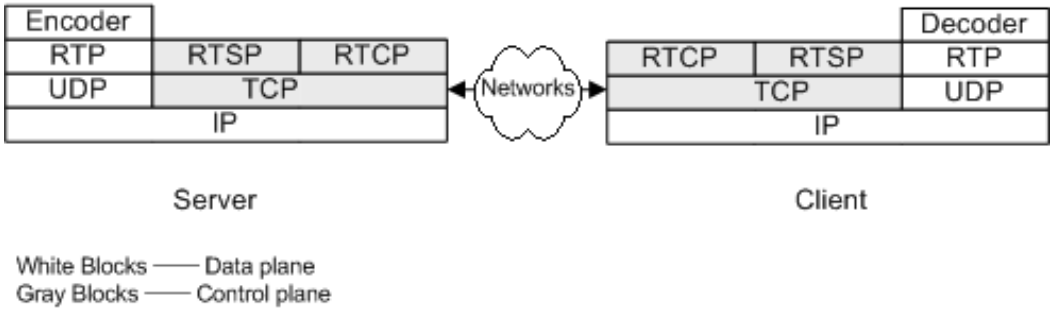


Figure 3-8 Protocol stack of network-aware end systems

There are two communication planes required by remote media access: a Data Plane for media transport, and a Control Plane for media session control. The MPEG-4 specification adopts out-of-band signaling, so that the Data and Control Planes can use different transport protocols, i.e., User Datagram Protocol (UDP) and TCP respectively in our approach. The process of transporting of MPEG-4 content over the Internet or wireless networks can therefore be split into two parts, the MPEG-4 Media Transport and MPEG-4 Media Control.

3.4.1 MPEG-4 Media Transport

The transport protocol family for media streaming includes UDP, TCP, Real Time Protocol (RTP), and Real Time Control Protocol (RTCP) protocols. UDP and TCP provide basic transport functions while RTP and RTCP run on top of UDP/TCP.

UDP and TCP

UDP/TCP protocols support such functions as multiplexing, error control, congestion control, or flow control. Since TCP retransmission introduces delays that are not acceptable for streaming applications with stringent delay requirements, UDP is typically employed as the transport protocol for video streams, while TCP is for streaming control. In addition, since UDP does not guarantee packet delivery, the receiver needs to rely on the upper layer (i.e., RTP) to detect packet loss.

RTP and RTCP

The *Real Time Protocol* [15] is an Internet standard protocol designed to provide end-to-end transport functions for supporting real-time applications. The *Real Time Control Protocol* [15] is a companion protocol with RTP and is designed to provide QoS feedback to the participants of an RTP session. In other words, RTP is a data transfer protocol while RTCP is a control protocol.

RTP does not guarantee QoS or reliable delivery, but rather, provides the following functions in support of media streaming: time-stamping, sequence numbering, payload type identification, and source identification.

RTCP is the control protocol designed to work in conjunction with RTP. In an RTP session, participants periodically send RTCP packets to convey feedback on quality of

data delivery and information of membership. Basically, RTCP provides the following services:

1) QoS feedback: This is the primary function of RTCP. RTCP provides feedback to an application regarding the quality of data distribution. The feedback is in the form of sender reports (sent by the source) and receiver reports (sent by the receiver). The reports can contain information on the quality of reception such as:

a) fraction of the lost RTP packets, since the last report;

b) cumulative number of lost packets, since the beginning of reception;

c) packet interarrival jitter;

d) delay since receiving the last sender's report. The control information is useful to the senders, the receivers, and third-party monitors.

Based on the feedback, the sender can adjust its transmission rate; the receivers can determine whether congestion is local, regional, or global; and network managers can evaluate the network performance for multicast distribution.

2) Participant identification

3) Control packets scaling: To scale the RTCP control packet transmission with the number of participants, a control mechanism is designed as follows. The control mechanism keeps the total control packets to 5% of the total session bandwidth. Among the control packets, 25% are allocated to the sender reports and 75% to the receiver reports. To prevent control packet starvation, at least one control packet is sent within 5 s at the sender or receiver.

4) Inter-media synchronization

5) Minimal session control information

3.4.2 MPEG-4 Media Control

In the media control part, specific session control protocol should be used to define the messages and procedures to control the delivery of the multimedia data during an established session. The Real Time Streaming Protocol (RTSP) [14] is such a session control protocol which has been recommended by 3GPP for packet-switched streaming

service (PSS) [9] and International Telecommunication Union (ITU) H.323 for multimedia teleconferencing services [43].

One of the main functions of RTSP is to support video-cassette-recorder-like control operations such as stop, pause/resume, fast forward, and fast backward. In addition, RTSP also provides means for choosing delivery channels (e.g., UDP, multicast UDP, or TCP), and delivery mechanisms based upon RTP. RTSP works for multicast as well as unicast. Another main function of RTSP is to establish and control streams of continuous audio and video media between the media servers and the clients.

However, RTSP is specially designed for the Internet. For the wireless applications, it should reply on other's mobility mechanisms, such as the GTP-based link-level mobility mechanism in GPRS and UMTS or SIP-based application level mobility mechanism. In our approach, an enhanced GTP-based handoff procedure is proposed for handling user mobility in media streaming.

In this chapter, the concepts of layered coding and multiple description coding are introduced in order to solve the bandwidth fluctuation, packet loss and heterogeneity problem in the wireless environment. The system model is proposed as a distributed multimedia delivery mobile network, followed by the discussion of the video streaming over IP DiffServ. The protocol stacks of the proposed D-MDMN and the network-aware end system are presented. In the next chapter, we propose solutions for video mobility under this system model.

4 Proposed Solutions for Video Mobility

The problem of handoff in a wireless network is well-know; however, it is largely unexplored in the applications of streaming media.

For the purpose of solving the handoff problems in media streaming as discussed in Chapter 2, a combination of scalable multiple description coding with distributed video storage in the DiffServ mobile network to support streaming video handoff is proposed. It leverages the distributed multimedia delivery mobile network to provide path diversity to combat outage due to handoff. If a feedback channel is available, receiver reports from both the base station and mobile host will be employed to split the wired and wireless domain, such that the wireless channel condition (e.g., packet loss) can be known by the sender.

In this chapter, the join design of layered coding and multiple description coding and the proposed scalable multiple description coding will be presented. And then, the MPEG-4 video streaming issues over the IP DiffServ mobile network and the proposed handoff procedures will be discussed.

4.1 Joint Design of MDC and LC

There are almost no referenced works on the joint design of layered coding and multiple description coding to complement their drawbacks. In our approach, a Scalable Multiple Description Coding framework is proposed to leverage the distributed multimedia delivery mobile network to provide path diversity to combat outage due to handoff. The coded video stream consists of *MDC components* and *LC components*. In the proposed multimedia delivery mobile network, MDC components enhance the robustness to losses and bit errors of LC components through path diversity and error recovery. MDC components also reduce the storage, reliability and load balancing requirement among distributed media edge servers. At the same time, LC components not only deal with the unbalanced MD operation at the server end, but also combat the bandwidth frustrations of the time-varying wireless channels.

4.1.1 Architecture of Proposed SMDC Framework

The architecture of the proposed SMDC framework is depicted in Figure 4-1. It is an object-based coding which jointly employs the PFGS and Multiple State Recovery (MSR) on the condition that it is compatible with MPEG-4.

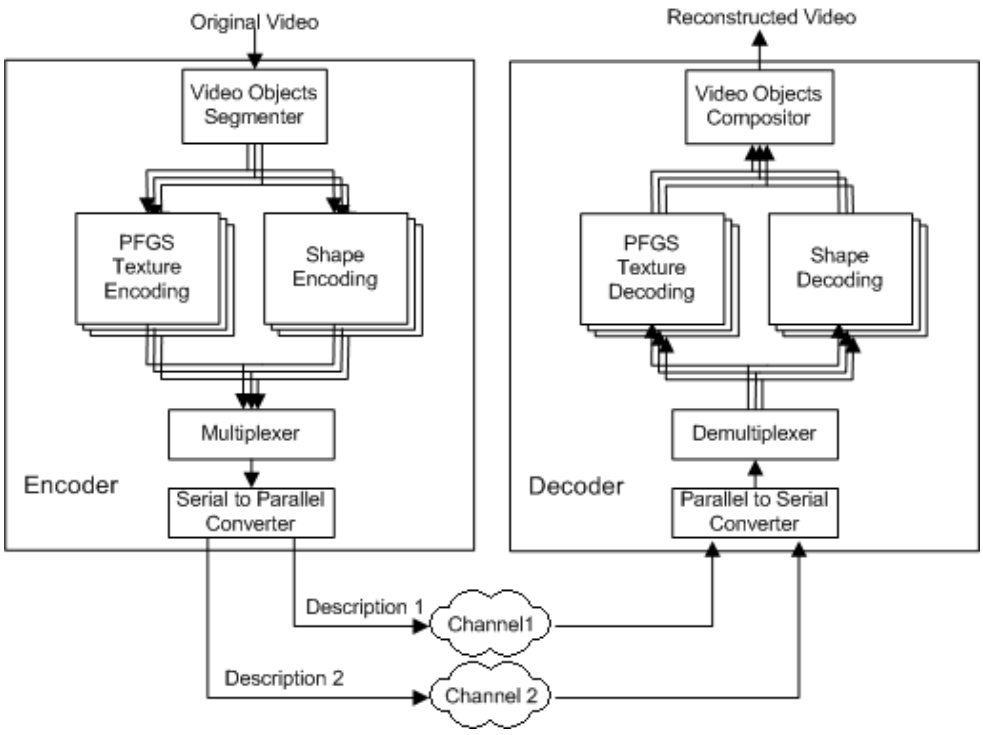


Figure 4-1 Proposed SMDC architecture

Similar to the human visual system mechanism, the smallest entity in the SMDC is each object in a picture with its associated shape, texture in the interior of the shape, and motion. The original video input to the encoder is segmented into a set of individual video objects (VOs). Each VO then is separately compressed through shape encoding and PFGS texture encoding, such that the shape and texture information can be split into four different VOPs. For support of two descriptions, the encoder should store the last two previously coded frames (instead of just the last one) and choose which previously coded frame to be used as the reference for the current prediction. After multiplexing, the four different VOPs converge into one video stream with four different layers. This video

stream is further partitioned into two subsequences of frames: odd numbered video frames (Description 1) and even numbered video frames (Description 2).

The different descriptions should be transmitted over different channels undergoing independent error effects to minimize the chance that both video streams are corrupted at the same time. As a matter of fact, the video stream can be partitioned into *N* complementary frame subsequences if there are *N* different channels between the encoder and the decoder. However, it also adds complexity of MD assembling at the decoder. It is an open issue to determine how many descriptions should be used for encoding. For presentation simplification, two descriptions and two corresponding channels are chosen in the following discussion.

At the decoder, the processing procedures reverse in accordance. Similarly, the decoder should alternate which previous erroneously decoded frame it uses as the reference for the next prediction. Both MPEG-4 version 2 (NEWPRED mode) [8] [29] [30] and H.263+ (RPS mode) [31] [21] support switching prediction among reference frames. If both descriptions are received erroneously after parallel to serial converter (MD assembler) and demultiplexing, then the shape and texture information of VOs are restored from shape and PFGS texture decoder for final composition into reconstructed video. If there is an error in a stream, the error propagation will happen in that stream due to motion compensation and differential encoding.

The SMDC framework can employ any shape coding [13] [42], e.g., binary shape coding or grayscale coding [63]. The texture coding techniques are still DCT-based coding for arbitrary shaped objects. The concept of object based representation makes it possible to exploit the content redundancy in addition to the data redundancy and improve the coding efficiency for the very low bit-rate transmission.

For an illustration of the capability of error recovery, an MD approach is compared with a conventional single description approach in Figure 4-2 [47]. For simplicity, B-frames are not illustrated in the figure. There is an error when decoding P-frame 4 in SD or P-frame 5 correspondingly in MD which is forward predicted by P-frame 3. In the SD approach, P frame 4 is lost and the decoder has to freeze P-frame 3 (or perform other error concealment) until I-frame 10. In the MD approach, however, P-frame 5 may be recovered or concealed by using information from its previous P-frame 4 and future P-

frame 6 which are correctly decoded in the other description. Errors on both descriptions in decoding P-frames 8 and 11 are recovered or concealed in the same manner as long as both descriptions are not simultaneously lost.

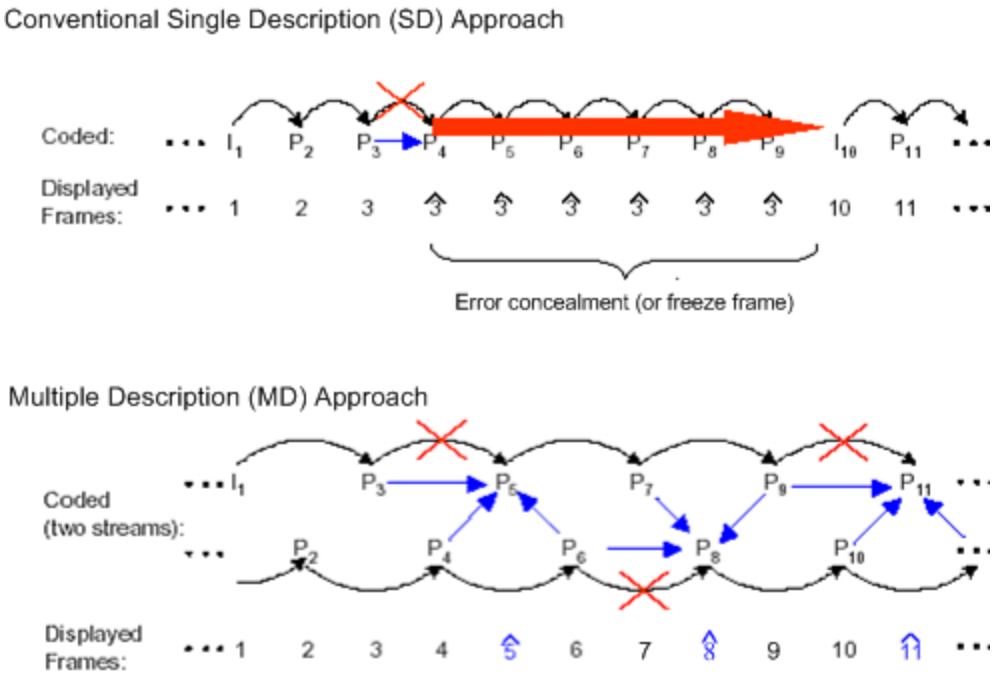


Figure 4-2 Single Description versus Multiple Description

Note that, the MD approach should add extra bits for MD property; and the SD approach should also add extra bits (e.g., I-frame 10) for intra coding or FEC. It has been shown that the MD approach requires an increase of 12~20 % transmission bit rate as compared with the SD approach but will result in much stronger capability of error recovery than its counterpart [47].

4.1.2 Scalability Structure of SMDC Framework

The proposed SMDC scalability structure (shown in Figure 4-3) is as follows:

1) The Shape Base Layer that consists of shape information of VOs in the intra-coded plane (I-VOP) or shape and motion information of VOs in the predictively coded plane (P-VOP);
2) The Texture Base Layer that consists of basic texture information of VOs contoured by the Shape Base Layer;

3) The Texture PFGS Layer that consists of texture information of SNR scalable enhancement for the Texture Base Layer;
4) The Texture PFGST Layer that consists of motion-compensated residual frames (i.e., motion vectors and bitplane-DCT residual signals) predicted from the Texture Base Layer for temporal scalable enhancement. In comparison, the motion-compensated PFGST frames in SMDC take the place of B-frames in the multilayer FGS-temporal scalability structure presented in [6].

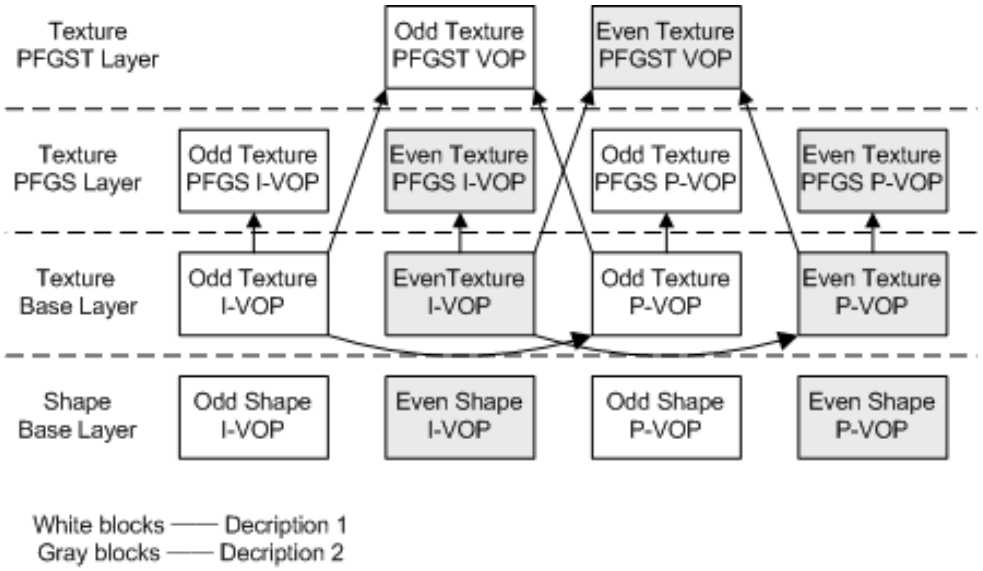


Figure 4-3 Proposed SMDC scalability structure

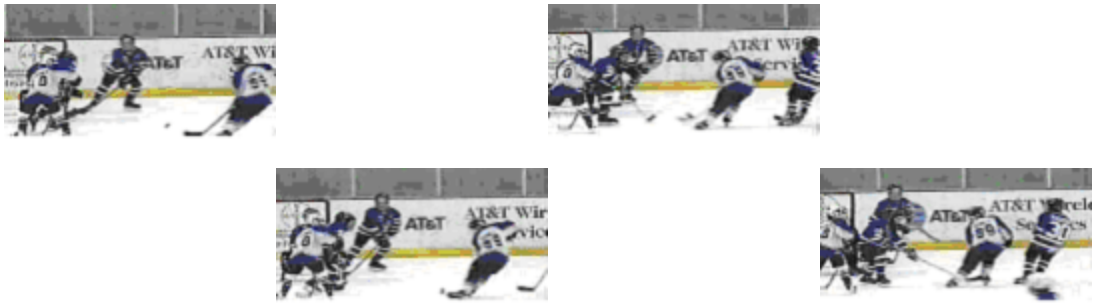
For illustration sake, shown in Figure 4-4, suppose the first three layers are implemented and the Texture PFGST Layer is left as an option. Thus, playing only one description with only the Shape Base Layer gets a black and white (or grayscale) video at the half frame rate, shown in Figure 4-4 (a). Playing only one description with the Shape Base Layer and the Texture Base Layer gets a color video in a basic quality at the half frame rate, shown in Figure 4-4 (b). Playing only one description with all three layers yields a color video in a better quality at the half frame rate, shown in Figure 4-4 (c). In the same way, if both two descriptions with all the three layers can be decoded correctly, it yields a color video in the best quality at the full frame rate, shown in Figure 4-4 (d).

However, the layering in SMDC is more flexible than that of illustration. The Texture PFGS Layer needs not be discarded as a whole. The enhancement bit stream can

be truncated anywhere to achieve the target bit-rate. This benefit of achieving continuous rate control comes from the bitplane coding in the PFGS encoder [17] for the enhancement stream.

(a) One description with only the Shape Base Layer

(b) One description with the Shape Base Layer and the Texture Base Layer



(c) One description with all three layers

(d) Both two descriptions with all the three layers

Figure 4-4 Video illustration of SMDC layered structure

4.1.3 Advantages of Proposed LC Component

1. Wired Domain

1) Unbalanced MD Operation

As mentioned in Section 3.2, the requirement of unbalanced MD coding [26] is well-known but largely unexplored in MD video coding. The characteristics of each path in a packet network are different and time-varying, therefore the available bandwidth in each path may differ. This results in the requirement of unbalanced MD operation, where the bit rate of each description is adapted based on the available bandwidth along its path. The proposed SMDC description is scalable along its path without any close-loop feedback delay, which is well suitable for unbalanced MD operation.

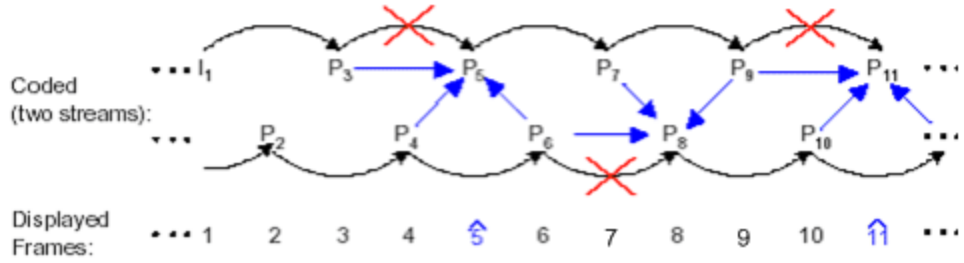
The proposed SMDC approach is naturally balanced in both streams (i.e., descriptions) assuming that the even and odd frames have equal complexity. To achieve unbalanced operation one can adapt the *quantization*, *spatial resolution* or *frame rate.* However it is important to preserve approximately equal quality in each stream to prevent an observer from perceiving a quality variation (flicker) at half the original frame rate (particularly important for the case of no losses).

Rate control via coarser quantization may be used for small rate changes (e.g., 10% rate reduction at a cost of 0.5 dB [26]); however, it may not be appropriate for large rate changes. The potential flicker also suggests that changes in spatial resolution (i.e., spatial subsampling) may be inappropriate.

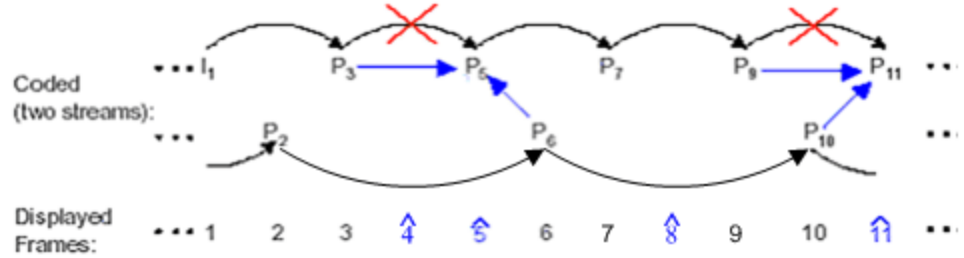
Adapting the frame rate (i.e., temporal subsampling) [26] is a simple and effective mechanism for reducing the required bit rate while preserving the quality per frame and largely preserving the error recovery capability, as illustrated in Figure 4-5. However, if the frame rate of one stream is decreased too much, the quality variation of that stream can not be approximately preserved. Also, the unbalanced MD operation will fail if the bit rate ratio of these two streams is larger than 2:1, which is illustrated in Figure 4-5.

Suppose that the bit rate of the upper stream is bigger than that of the lower stream in Figure 4-5, where P_x denotes the P-frame X. The balanced MD operation is shown in Figure 4-5 (a), where the damaged P-frame 5 can be recovered or concealed from P-frames 4 and 6, and P-frame 11 is recovered or concealed from P-frames 10 and 12. In Figure 4-5 (b), the frame rate of the lower stream has to be decreased by 50% for a bit rate ratio of 2:1. That is,

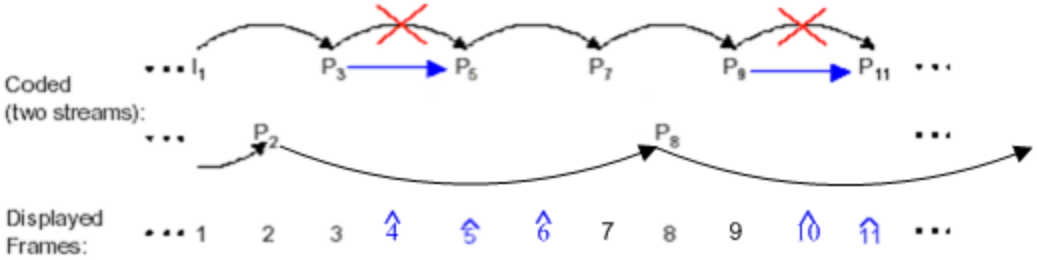
P-frames 4 and 8 have to be discarded. The damaged P-frame 5 can be recovered but only from P-frame 6, and P-frame 11 can be recovered but only from P-frame 10. It is straightforward that the error recovery capability of 2:1 unbalanced MD operation, illustrated in Figure 4-5 (b), is lower that that of the balanced MD operation illustrated in Figure 4-5 (a). In Figure 4-5 (c), the frame rate of the lower stream has to be decreased by 75% for a bit rate ratio of 3:1. However, the damaged P-frames 5 and 11 in the high bit rate stream can not be recovered from the low bit rate stream because their adjacent previous P-frames 4 and 10, and their adjacent future P-frames 6 and 12 have to be discarded.



(a) Balanced MD Operation for Bit Rate Ratio of 1:1



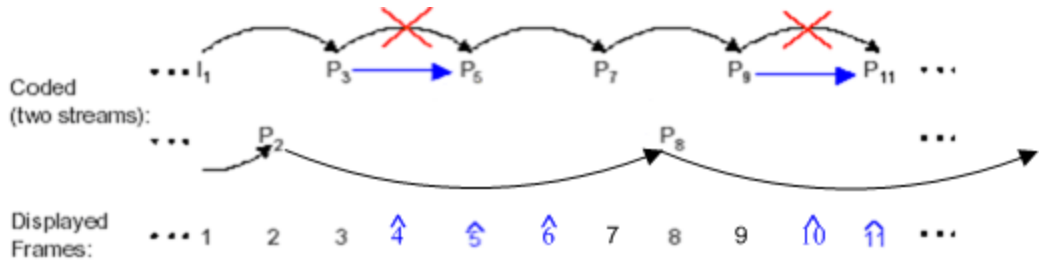
(b) Unbalanced MD Operation for Bit Rate Ratio of 2:1 using temporal subsampling



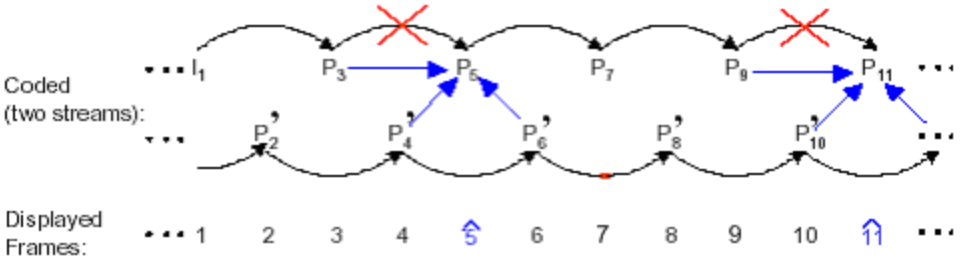
(c) Unbalanced MD Operation for Bit Rate Ratio of 3:1 using temporal subsampling

Figure 4-5 Balanced and Unbalanced MD Operation

In order to address the unbalanced problem, the concept of layered coding is proposed for the MD video coding. The capability of error recovery of MDC and SMDC are compared under the bit rate ratio of 3:1 in Figure 4-6.



(a) Unbalanced MD Operation for Bit Rate Ratio of 3:1 using temporal subsampling



(b) Unbalanced MD Operation for Bit Rate Ratio of 3:1 using Layered Coding

Figure 4-6 Comparison of Unbalanced MDC with Unbalanced SMDC

As discussed above, the errors occurred in the high bit rate stream can not be recovered or concealed from the low bit rate stream in the case that the bit rate ratio is larger than 2:1. Suppose that the bit rate of the upper stream in Figure 4-6 is three units and that of the lower stream is only one unit. Instead of temporal subsampling illustrated in Figure 4-6 (a), the layered coding is introduced in Figure 4-6 (b) for the unbalanced MD operation. As to the path of low bandwidth, part of the enhancement layers can be dropped so that the original frame rate can be preserved. In Figure 4-6, P_x' denotes the rest past of P-frame X after layer-dropping in order to adapt to the bandwidth limitation. Thus, the damaged P-frames 5 and 11 in the upper stream still can be recovered from the frame of the lower stream, shown in Figure 4-6 (b), in the same manner as the balanced MD operation shown in Figure 4-5 (a). In other words, the unbalanced MD operation using LC does not affect the error recovery capability of SMDC.

Table 4-1 summaries the methods of adapting bit rate of each description. LC technique is suitable for the unbalanced MD operation, especially when the bit rate ratio of both descriptions is larger than 2:1.

Table 4-1 Methods of adapting bit rate of each description

	Performance without loss	Performance with loss
Quantization	Good for small changes (0-10 %), above which possible flickers exist	Good for small changes (0-10%), above which error recovery capability reduces
Spatial Subsampling	Potential flicker	Potentially reduced
Temporal Subsampling	Good for middle range of bit rate changes (bit rate ratio less than 2:1)	Generally good recovery (bit rate ratio less than 2:1)
Layered Coding in SMDC	Good for large range of bit rate changes (bit rate ratio larger than 2:1)	Generally good recovery (bit rate ratio larger than 2:1)

2) Error Resilience Enhancement

To explore the error resilience and concealment tools in MPEG-4, there is a clear advantage to distinguish not only different kinds of frames (referred to as Video Object Plane or VOP in MPEG-4), but also different types of information, such as shape, motion, and texture, within the same frame. Therefore, the shape, motion, and texture information in the bit-stream of an object-based video is re-organized into different layers in the proposed SMDC scheme to support the classification and priority assignment in the DiffServ network, which is discussed in details in Section 3.3.3.

2. Wireless Domain

1) Open-loop Rate Control

Design issue 1: RTCP-based rate control is specifically designed for the Internet

A number of papers have considered how to control the transmission rate of non-TCP flows. TCP-friendly model-based [66] and probe-based [36] rate control mechanisms calculate their maximum transmission rate using a TCP throughput formula [24] [22] or mimicking TCP behavior. To determine the transmission rate, these mechanisms require feedback from the receiver to obtain packet loss rate and round trip time (RTT)

information. Some rate control mechanisms [36] utilize the RTCP to obtain feedback information from receivers . The receiver of an RTP media stream sends back RTCP receiver reports, which include packet loss and jitter information, so that the sender can identify network congestion condition and control its transmission rate accordingly. Most of the rate control mechanisms mentioned above are designed specially for the Internet and assumed that packet loss, delay, and jitter are caused by network congestion.

In mobile networks, however, packet loss and jitter may also be caused by radio link errors. Since radio links have much higher bit error rates, packets are frequently discarded due to the presence of bit errors. When conventional rate control mechanisms are applied to mobile networks, a sender cannot identify the network congestion condition correctly, and this leads to inappropriate rate control. A typical symptom is that a sender reduces its transmission rate even if the network is not congested [75].

Design issue 2: The relatively long RTCP transmission interval.

To scale the RTCP control packet transmission with the number of participants, a control mechanism is designed as follows in RFC 1889 [15]. The control mechanism keeps the total control packets to 5% of the total session bandwidth. Among the control packets, 25% are allocated to the sender reports and 75% to the receiver reports. To prevent control packet starvation, at least one control packet is sent within 5 seconds at the sender or receiver.

RTCP makes no provision for timely feedback that would allow a sender to repair the media stream immediately: through retransmissions [65], reactive FEC, or media-specific mechanisms such as reference picture selection for some video codecs. Typically, the feedback interval is constrained on the order of tens to hundreds of milliseconds [48].

The QoS maintenance or guarantees to multimedia streams using RTCP-based reports is still under investigation, especially in the wireless environment. As a result, instead of close-loop rate control mechanism, the concept of layered coding is adopted in our approach to implement open-loop adaptive rate control mechanism (e.g., DiffServ-based rate filtering) in the BS and other wired intermediate nodes to combat traffic congestion.

2) Resource Reservation and Heterogeneity

The concept of layered coding provides a very flexible and efficient solution to the problem of resource reservation and receiver heterogeneity in the wireless domain.

First, there is no need to reserve bandwidth for the complete stream since typically only the base layer needs QoS guarantee. As a result, CAC can be based only on the requirement of the base layer and resources are reserved only for the base layer. Second, the enhancement layers of one connection can share the leftover bandwidth with the enhancement layers of other connections. The CAC algorithms for wireless channels are out of the scope of the thesis.

In addition, the technique of the layered coder is a better alternative of trancoder in the base stations to combat the network heterogeneity or receiver heterogeneity so that there are no migration of transcoding state information required, which will be discussed in Section 4.3.3.

4.2 MPEG-4 Video Streaming over IP DiffServ

4.2.1 QoS Mapping between UMTS and IP DiffServ

Table 4-2 Mapping of UMTS QoS classes to DiffServ PHB classes

	UMTS QoS classes	Conversational Class	Streaming Class	Interactive Class	Background Class
Features	Application	VoIP/video conferencing	Streaming audio/video	Web browsing	Background download /E-mails
	Delay jitter	Stringent and low	Bounded	Tolerable	Unbounded
	BER	Tolerable	Tolerable	Low	Low
Defined attributes	Maximum bit rate	2Mbps	2Mbps	2Mbps	2Mbps
	Guaranteed bit rate	2Mbps	2Mbps	-	-
	Transfer delay	≤100ms	≤280ms	-	-
Mapping	DiffServ PHB classes	DSCP = EF	DSCP = AF	DSCP = BE	DSCP = BE

To provide an end-to-end QoS for IP based traffic over a UMTS network, one of the most difficult issues [72] [73] is to provide QoS attribute translation and mapping between the IP world and the UMTS world. The features and the attributes of UMTS QoS classes defined in [61] are summarized in Table 4-2. Our mapping scheme between the DiffServ and UMTS QoS classes is also presented in Table 4-2.

4.2.2 IP DiffServ MPEG-4 Video Marking Algorithm

Existing Internet video dissemination schemes usually do not support classification of varying video information beyond the simple distinguishment of different types of frames (I, P and B frames) [16]. Encoded video data are placed in the bit-stream according to the temporal and spatial positions, i.e., block by block, macroblock by macroblock, and frame by frame. Different types of information, such as shape, motion, and texture, are interleaved together, although they have different levels of importance during decoding. For example, the shape and motion information is more important than the texture for a P frame in MPEG-4 [30]. If the shape and motion information is lost during transmission, it is hard for the decoder to reconstruct the P frame successfully. However, if partial texture information is lost, it is still possible to reconstruct the P frame with somewhat acceptable quality using error concealment algorithms.

Aiming at robust transmission, MPEG-4 supports a set of error resilience and concealment tools, such as video packet based resynchronization approach which provides a flexible self-contained decoding unit, and data partitioning mode which separates the shape, motion and texture data in VPs using DC Markers or Motion Markers. Such tools are quite suitable for wireless transmissions where most errors are at the bit-level, but they are not sufficient for Internet transmission where most errors are caused by packet loss. This is because a video packet (or several video packets) is usually encapsulated as packet payload directly, and consequently the information is still interleaved together within an IP packet.

To explore the error resilience [28] [29] [30] and concealment [34] tools in MPEG-4, there is a clear advantage to distinguish not only different kinds of frames (i.e., VOP in MPEG-4), but also different types of information within the same frame. Usually, IP

video traffic is classified as one AF class with three different levels of drop precedence. For example, an IP DiffServ video marking algorithm (DVMA) is proposed in [71] [72] as follows:

If stream is "video stream" then

If "base layer video stream" then

'Level 1 = minimum QoS'

DSCP= AF Low Drop Precedence (e.g., AF21)

If "enhanced layer video stream 1" then

'Level 2 = medium QoS'

DSCP= AF Medium Drop Precedence (e.g., AF22)

If "enhanced layer video stream 2" then

'Level 3 = maximum QoS'

DSCP= AF High Drop Precedence (e.g., AF23)

Typically, (W)RED [4] or similar active queue management approach has to be adopted to combine stochastic dropping of packets with IP Precedence. The WRED gateway calculates the average queue size *AvQ*, using a low-pass filter with an exponential weighted moving average. Given the minimum threshold TH_{min}, a maximum threshold TH_{max}, the exponential weight factor *ef* and the mark probability denominator *p*, the WRED algorithm is described in the following:

For each packet arrival, calculate the average queue size

$$AvQ = \begin{cases} (1-W_q)\times AvQ + W_q \times Q & if \quad Q \neq 0 \\ (1-W_q)^m \times AvQ & if \quad Q = 0 \end{cases}$$

If $TH_{min} \leq AvQ < TH_{max}$

Drop the arriving packet with probability *P*

Else if $TH_{max} \leq AvQ$

Drop the arriving packet

where

$W_q = (1/2)^{ef}$,

$P = (1/p)\times(AvQ - TH_{min}/TH_{max} - TH_{min})$,

m = queue_idle_time / transmission_time.

(W)RED takes advantage of the TCP retransmission mechanism. However, as discussed in Section 3.4, UDP is more suitable for video streaming. All the random dropping packets will be considered as packet loss and will not be retransmitted. Because of error propagation of streaming video, the effect of packet loss gets worse. Since MPEG-4 video is predictive inter-frame coded and layered coded, artifacts due to random packet dropping can persist for many frames or layers. For example, consider a 30 frame/s MPEG video sequence with one I frame every 15 frames. If an error occurs while transmitting the I frame, the effect persists for 15 frames, or 500 ms, which is quite noticeable to a viewer. Jill and Gaglianello analyzed and presented the results of the relationship between the packet loss rate and the frame error rate [35], shown in Figure 4-7, from a study of streaming MPEG compressed video over the public Internet, using the RTP and UDP transport protocols. Similarly, if an error occurs while transmitting the base layer, its enhancement layers have to be discarded. It means that stochastically isolated single packet loss or bit error is converted to burst packet loss or bit errors. Therefore, early random packet dropping before congestion is not suitable for video or audio streaming.

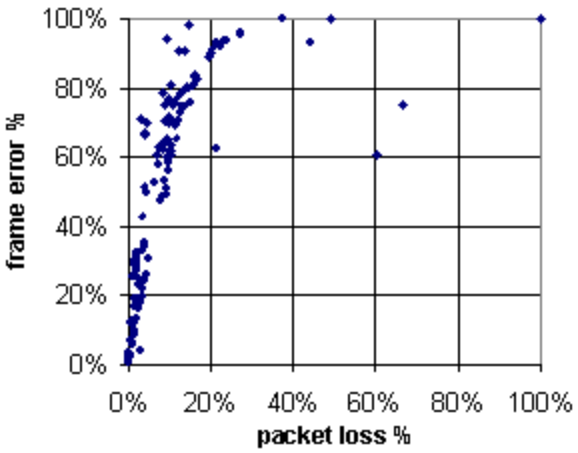


Figure 4-7 Packet loss effect on frame error rate

A novel marking algorithm for MPEG-4 video encoded by SMDC is proposed in Table 4-3 to support the priority assignment after the classification given in Table 4-2. In the proposed SMDC scheme, we re-organize different types of information, such as shape, motion, and texture, in the bit-stream of an object-based video into four different layers. Moreover, different layers of information are packetized into three different

classes of Application Level Packets (ALP) with various priorities, so that more important compressed information can be put into higher priority packets and less important information into lower priority ones. In comparison with the DVMA solution where there is only one queue with three different levels of precedence for video stream, each AF class in the proposed algorithm has one separated queue. This algorithm can be implemented by class-based Weighted Fair Queuing (WFQ). Details of the WFQ algorithm can be found in [76]. WRED should be disabled in each class.

Table 4-3 Proposed IP DiffServ MPEG-4 video marking algorithm

	Control Information	Shape Base Layer	Texture Base Layer	Texture PFGS Layer	Texture PFGST Layer
OD & BIFS	DSCP = *EF* or *AF11*	-	-	-	-
I-VOP	-	DSCP = *AF11* (Class I stream)	DSCP = *AF11* (Class I stream)	DSCP = *AF21* (Class II stream)	-
P-VOP	-	DSCP= *AF11* (Class I stream)	DSCP= *AF21* (Class II stream)	DSCP = *AF31* (Class III stream)	-
PFGST VOP	-	-	-	-	DSCP = *AF31* (Class III stream)

In addition, MPEG-4 introduces extra data control stream, such as the object descriptor (OD) and scene description (BIFS). These signalling streams are very loss- and jitter-sensitive and need to be protected and marked as *EF* or *AF11* if EF PHB is not available.

4.2.3 Evolution of System Model for Native IP DiffServ

Evolution of Mobile Network Model

In order to support DiffServ in UMTS, we propose to copy the DSCP value in the inner IP header to the outer IP header at encapsulation and copy the outer header's DSCP

value to the inner IP header at decapsulation. This mechanism allows GTP/FP tunnels to be configured without regard to DiffServ domain boundaries.

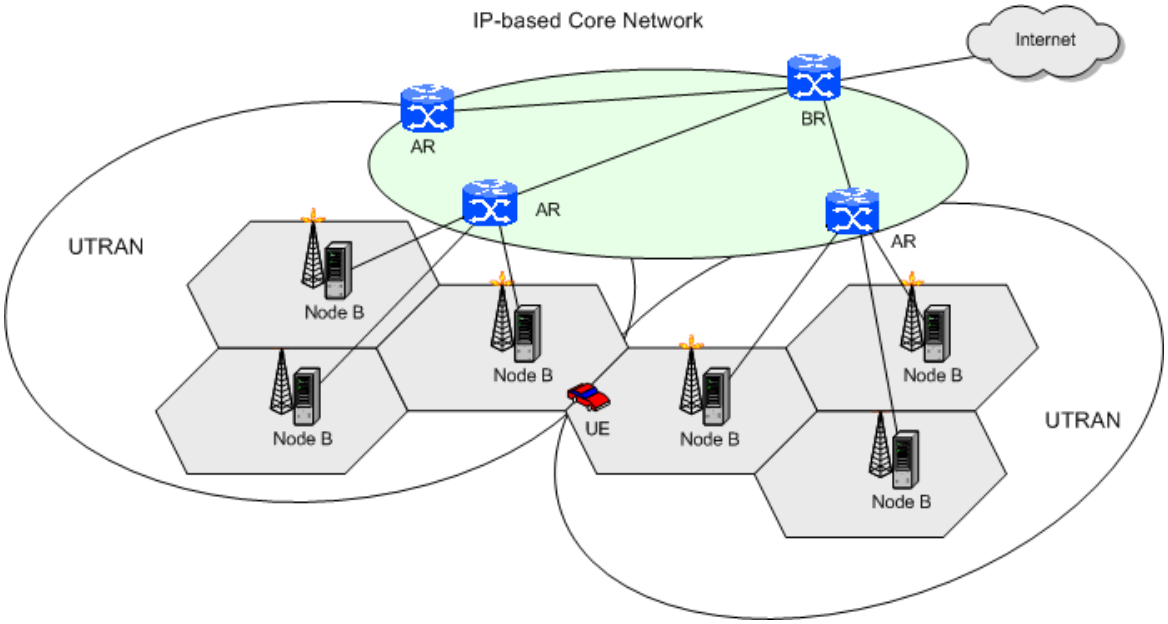


Figure 4-8 UMTS network model in IP native mode with 3G AR

A more efficient alternative, however, is to keep the native DiffServ processing procedure in the mobile network rather than the modified one as above (i.e., copying DSCP value). This requirement leverages the evolution of UMTS towards its final all IP-based phase, which is depicted in Figure 4-8. Note that the IP-based CN has enlarged to the edge of UTRANs compared with the network model shown in Figure 2-2. The functions of the GGSN, SGSN and RNC are further combined and implemented at one node, *Access Router* (AR). The GGSN and SGSN functions within the 3G AR provide all the UMTS-specific accounting and security features. The rest of the CN consists of regular routers and switches that forward packets on the basis of the user-level IP addresses. The *Border Router* (BR) denotes the functionality to avoid unwanted traffic between GPRS CN and the Internet. One or more BRs are served as gateways to the public Internet.

This network architecture provides a solution to implement the IP native mode forwarding in a larger portion of the operator's domain independent of any given access technology, and hence can be used by the operator to support heterogeneous access

networks. As the coverage of the IP native mode increases, the wireless-specific protocols are pushed farther toward the access segment. The operator may share the domain with other access techniques by just using a specialized AR. For example, an IEEE 802.11 AR may coexist with a 3G AR, using the same CN.

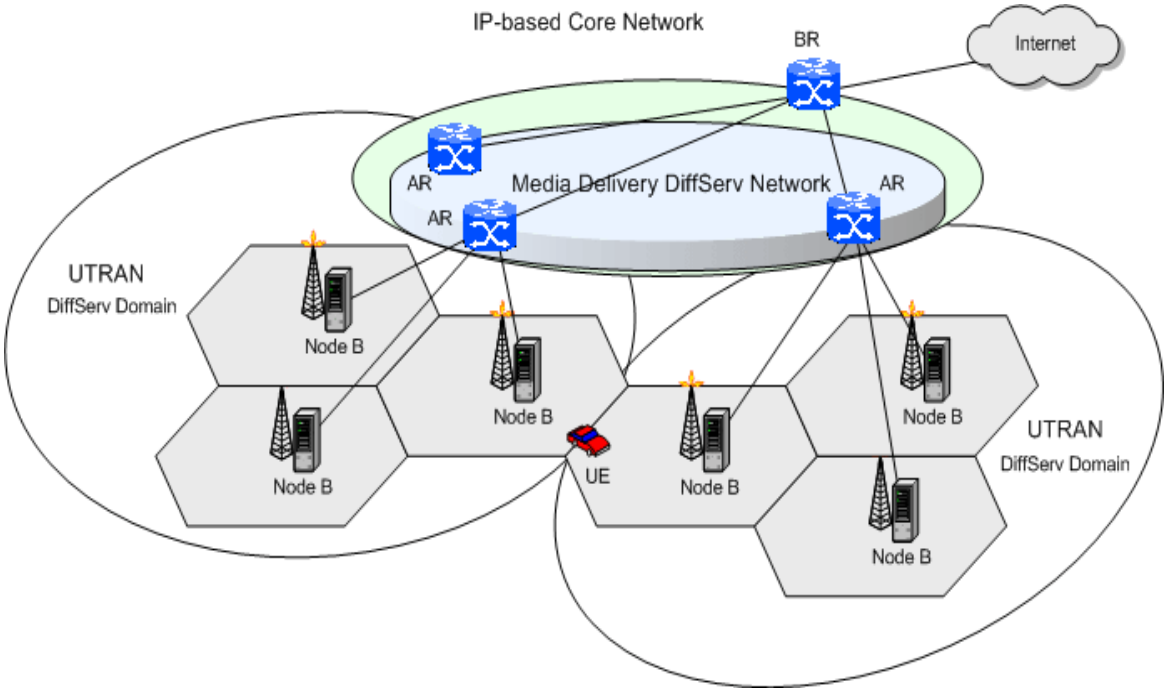


Figure 4-9 D-MDMN network model in IP native mode with 3G AR

Along with the enlargement of the IP-based CN, the introduction of D-MDMN in the 3GPP network is depicted in Figure 4-9. The IP-based Media Delivery Network can also be enlarged around ARs towards the whole IP-based CN, so that the media streaming services can be pushed further to the edge of UTRANs.

Conceivably, the IP native mode coverage can be extended into the UTRAN by implementing an AR with collocated Node B, RNC, SGSN and GGSN functions. Some equipment vendors are adopting this approach by building what are known as intelligent base stations with varying combined functionalities.

Evolution of Protocol Stack

In order to further analyse the implementation of DifferServ in Figure 4-9, the protocol stack in the IP native mode for non-video services is shown in Figure 4-10 (a), while that for video services is shown in Figure 4-10 (b).

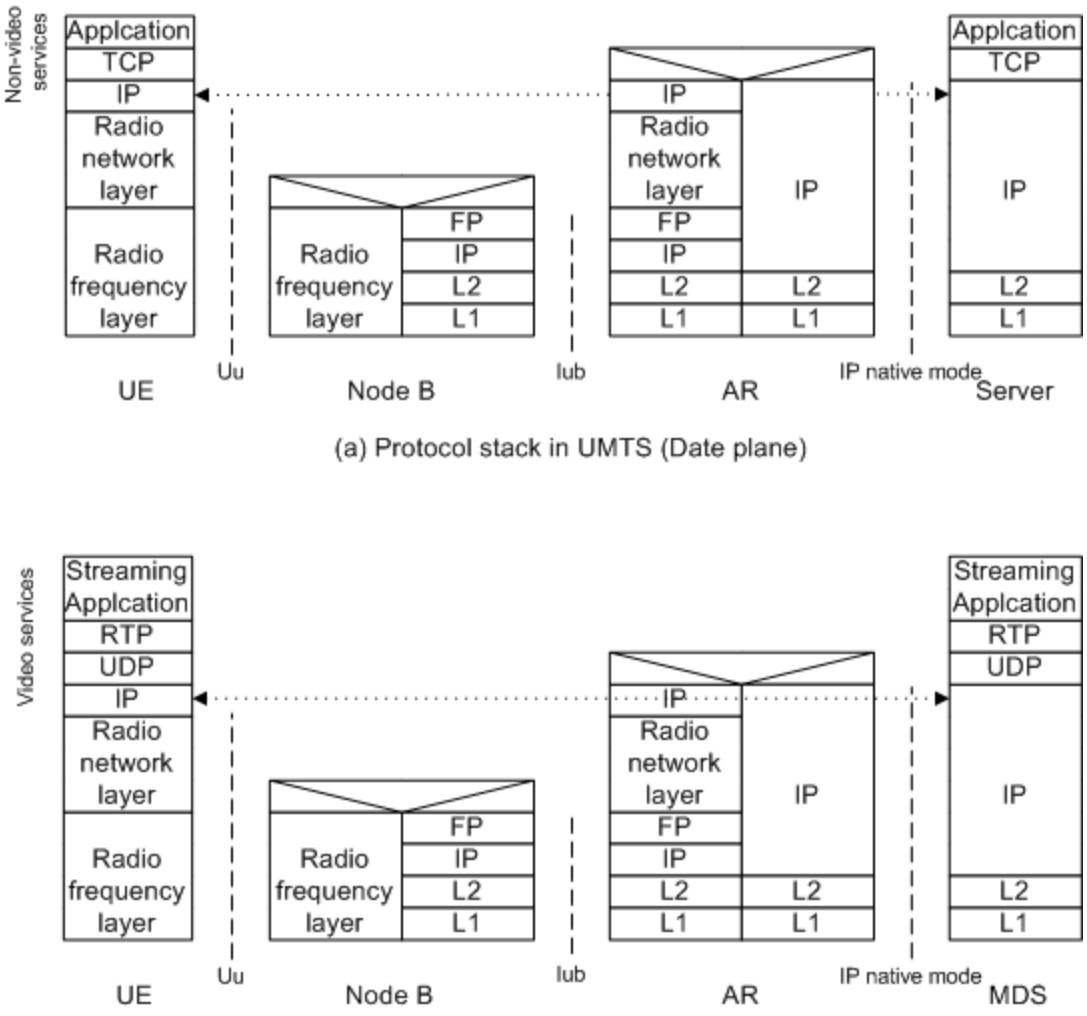


Figure 4-10 Evolution of Protocol stack (Data plane)

As the coverage of the IP native mode increases, the stack becomes more efficient, and the whole CN uses regular IP forwarding based on the end-user's IP address instead of the tunnel ID. FP frames are transported to and from the ARs over an IP network in the transport mode. If the ARs are the next hop of Node Bs, the tunnels are short enough that the ingresses (i.e., ARs) can execute DiffServ PHB based-on the inner IP header before the it is encapsulated by the addition of the outer FP tunnel header, and the egresses can also perform DiffServ PHB based-on the inner IP header after it is decapsulated by the removal of the outer FP tunnel header. Thus, the native DiffServ processing procedure in the mobile network is implemented rather than copying DSCP value from the inner IP header to the outer IP header.

4.3 Proposed Handoff Procedures for Video Streaming

4.3.1 Proposed intra-RAN Handoff Procedure

Under the proposed intra-RAN network model of the distributed multimedia delivery mobile network, shown in Figure 3-3, and the same assumptions given in Section 2.3.2, the intra-RAN handoff procedure for media streaming are proposed as follows, which also consists of three phases. The control plane of handoff procedure is shown in Figure 4-11, while the data plane of these three phases is shown in Figure 3-3. Only the differences are described in comparison with that in UMTS Release 4.

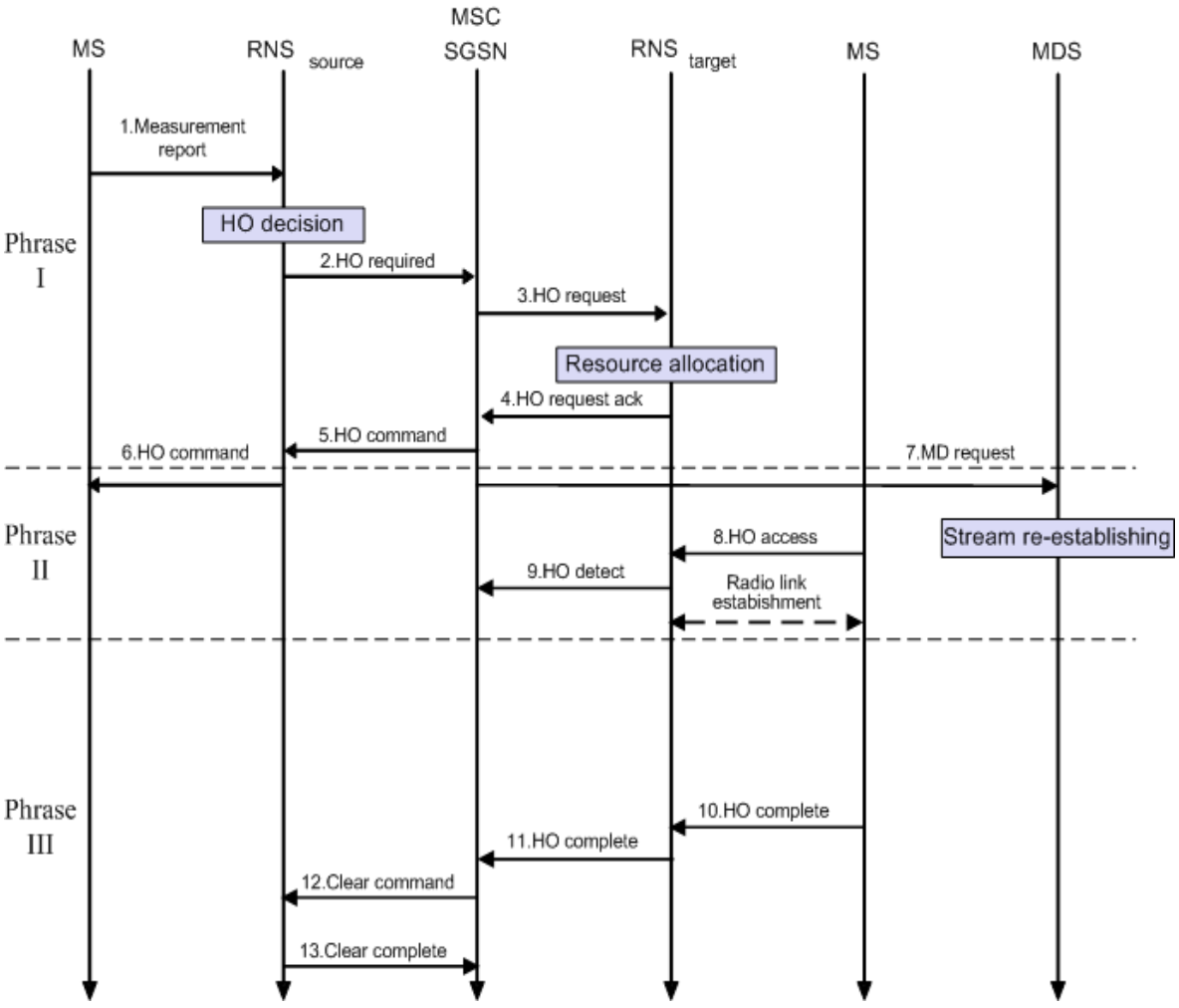


Figure 4-11 Propose intra-RAN handoff procedure (Control plane)

- **Phase I: Preparation of RNS handoff and resource allocation**

The control plane of the proposed handoff phase I is almost the same as that in UMTS Release 4, shown in Figure 2-4, except that the current position (offset) of the received media stream should go along with measurement report given by the MS. It is the only state information required for session migration which is small enough to be hidden inside the handoff signalling and be relayed to the SGSN.

On the data plane of handoff phase I, at the end of the preparation phase, the sRNS stops transmitting downlink data to the MS but will not store all downlink data which continue to arrive from the SGSN to the sRNC as no data forwarding is required.

- **Phase II: Moving the Serving RNS role to target RNS**

The most important difference between the proposed handoff phase II in the control plane and that in UMTS Release 4 in Figure 2-4, is that the *Stream Re-establishing* takes the place of the *Media Stream Forwarding*. There are no buffered data required to be forwarded. As soon as the GTP tunnel is created between the tRNS and the SGSN, the SGSN initiates the MD-request message (signal # 7) and the uplink flow is switched from the old path to the new path. Upon receiving the MD-request, the set of MDSs surrounding the SGSN starts the downlink media delivery from the offset point of the same stream at the handoff decision according to subscriber-service bindings in VLR (i.e., how many descriptions and layers the MS subscribes). In other words, the media stream is re-established. The MD-request message contains the offset information at the handoff decision point.

- **Phase III: Releasing resource reservation in the old path**

The most important difference between the proposed handoff phase III and that in UMTS Release 4, shown in Figure 2-4, is that there is no buffer requirement in the BSs for *data forwarding* and *resequencing*. Only a smaller buffer is needed in the BSs for absorbing the delay jitter of a video stream and for re-ordering due to changes in routing paths. The functionality of multiple description assembly is implemented in the MSs.

4.3.2 Proposed inter-RAN Handoff Procedure

Under the proposed inter-RAN network model of the distributed multimedia delivery mobile network, shown in Figure 3-4, and the same assumptions given in

Section 2.3.2, the proposed inter-RAN handoff procedure for media streaming also consists of three phases. Here the control plane of handoff procedure is briefly presented in Figure 4-12, while the data plane of these three phases is shown in Figure 3-4.

- **Phase I: Preparation of RNS handoff and resource allocation**

Note that there is no GTP tunnel required between SGSNs compared to UMTS Release 4, since no data forwarding is required.

- **Phase II: Moving the Serving RNS role to target RNS**
- **Phase III: Releasing resource reservation in the old path**

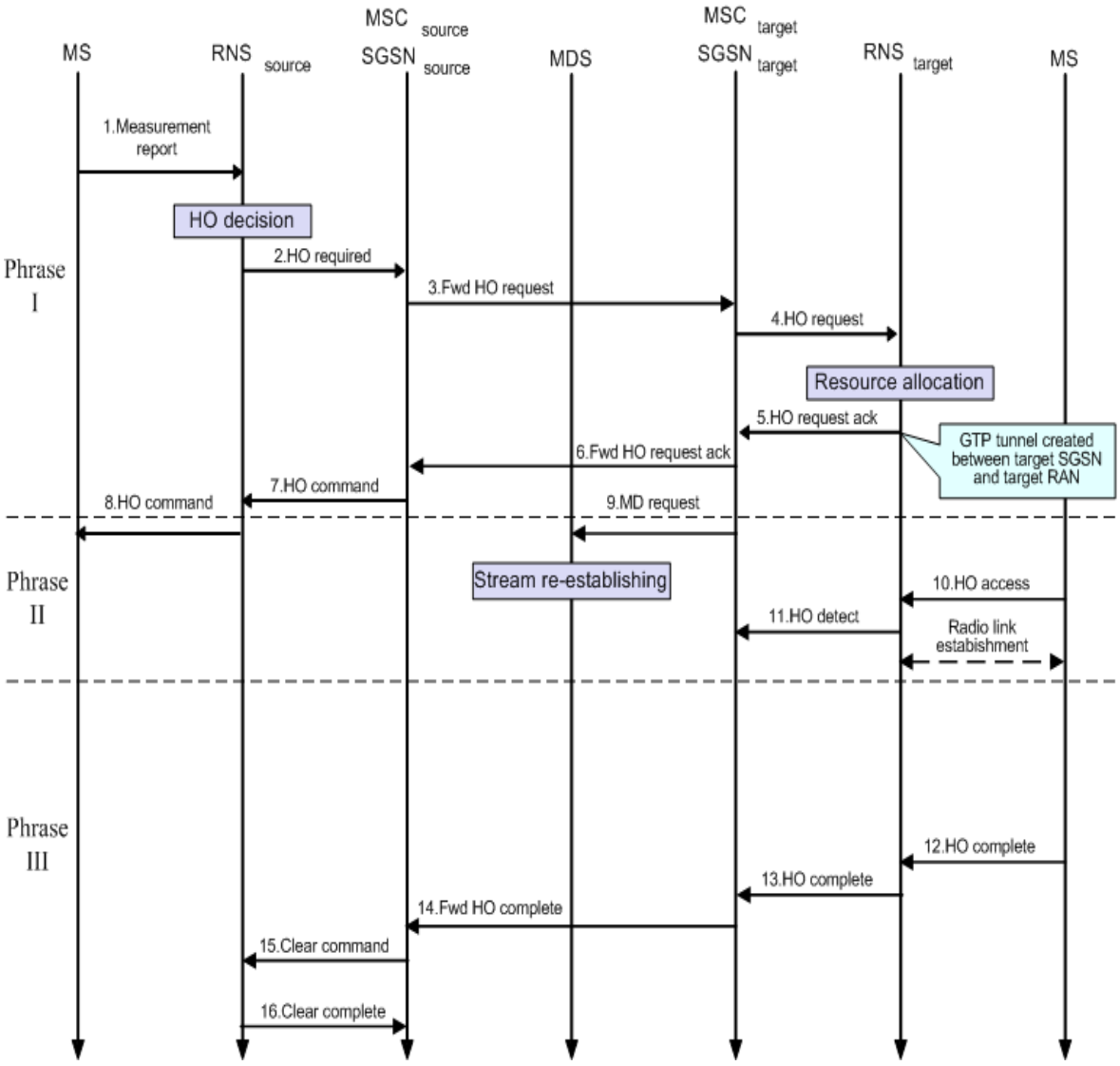


Figure 4-12 Proposed inter-RAN handoff procedure (Control plane)

4.3.3 Handoff Enhancement for Streaming Services

The advantages of the proposed handoff approach for media streaming are summarized as follows:

1. Due to the replacement of *stream re-establishing* with *data forwarding*, the handoff latency can be reduced, which also reduce the buffer size in the BSs.

In the UMTS Rel 4 handoff procedures, for downlink media streams, there are two possible situations when media stream gap or overlapping may happen:

(a) The media stream overlap/gap may be introduced when the tRNS takes the Serving RNS role and starts to produce the downlink data from forwarded GTP-PDUs. In this case the estimated gap/overlap for hard handoff is equal to the delay of the GTP tunnel used for data forwarding. This first instance of media stream overlap coincides with radio hard handover.

(b) The additional media stream gap may be introduced when the CN transport is optimized. In this case the gap will exist only if the delay via the optimized route is larger than the delay via the forwarding route.

In comparison, for downlink media streams during the proposed handoff procedures, there are only one possible situation when media stream gap (but no overlapping) may happen. That is, the media stream overlap/gap may be introduced when the tRNS takes the serving RNS role and starts to request a new set of MDSs for the rest of video delivery. In this case the estimated gap for hard handoff is equal to the delay of stream re-establishment for media delivery. This coincides with radio hard handoff.

If the transport bearer delay difference is smaller than the air interface Transmission Time Interval (TTI) (10, 20, 40 or 80 ms depending on the service), the amount of gap is most likely not existent.

In addition, as discussed previously, there is no buffer requirement in the BSs for *data forwarding* and *resequencing* in case of *stream re-establishig.* Only a smaller buffer is needed in the BSs for absorbing the delay jitter of a video stream and for re-ording due to changes in routing paths. The relatively small queue in the BSs also reduces the handoff latency.

2. It has relatively low packet loss (or frame error), and end-to-end delay.

Since the media streaming services are pushed to the edge of core network and the streaming media can be delivered over a shorter network path, the transfer delay and delay jitter of media service delivery, the probability of packet loss, and the total network resource occupation will be reduced.

Furthermore, with the employment of *stream re-establishig*, the relatively small queue in the BSs reduces the end-to-end delay further.

3. There is no extra handoff latency introduced due to session migration.

Since the technique of layered coding, e.g., SMDC, is employed as an alternative of trancoding to combat the network heterogeneity or receiver heterogeneity, there are no migration of transcoding state information required. Also, the migration of session description is not necessary because of the principle of stream re-establishing instead of data forwarding. Thus, only the migration of session parameters at the handoff decision point is required. The amount of state information is thereby small enough to be hidden inside the handoff signalling, so that there are no extra handoff latency introduced due to session migration.

4. It has relatively consistent QoS in all scenarios (Handoff scalability enhancement).

In UMTS Release 4, the values of handoff latency (i.e., delay jitter) vary with the lengths of data-forwarding path in different handoff scenarios. Also, the end-to-end delay varies with different delivery paths and different locations of the media providers, which is outside the CN and far from the mobile hosts.

However, due to the introduction of the distributed multimedia delivery mobile network, the values of handoff latency and end-to-end delay in different handoff scenarios depend mainly on the length of media delivery path from MDs to SGSN/MSC, and then to MS. Usually, the SGSN/MSC and the MDs are neighbor nodes. It is straightforward that the length of media delivery path from the MDs to the MS is relatively consistent in different handoff scenarios.

5. The amount of signalling traffic is slightly reduced during inter-RAN handoff.

A brief summary of above comparison between our proposal and that of UMTS is given in Table 4-4.

The proposed procedure in the scenario of inter-cell, intra-RNS handoff can be found in Appendix. Note that, in the scenario of intra-cell handoff, it is not necessary to re-establish media stream and the corresponding handoff procedures in UMTS R99 may be extended to support media streaming services.

Table 4-4 Summary of handoff solution comparison

		UMTS	D-MDMN
Principle		Data forwarding	Stream re-establishing
Queue size in the BSs		Large	Small
Packet loss (Frame error)		High	Low
End-to-end delay		High	Low
Handoff latency (gap or overlapping)		High (For downlink, two instances of stream gap/ overlapping may occur)	Low (For uplink, only one instance of stream gap may occur)
Consistency of QoS in all scenarios (i.e., Handoff scalability issue)		Poor	Good
Signalling traffic	intra-RAN handoff	12	13
	inter-RAN handoff	18	16

This chapter presents the details of the proposed solutions for video mobility under the system model defined in Chapter 3. In Section 4.1, the Scalable Multiple Description Coding framework is proposed to explore the joint design of layered coding and multiple description coding. Section 4.2 describes a novel IP DiffServ video marking algorithm to support the UEP of SMDC, which re-organizes the shape, motion, and texture information of video stream into different layers in order to implement the DiffServ in UMTS. Finally, the corresponding intra-RAN handoff and inter-RAN handoff procedures in D-MDMN are studied in Section 4.3 with the employment of the principle of video stream re-establishing for seamless handoff.

5 Simulations

5.1 Simulation Models

The simulation model of UMTS intra-RAN and inter-RAN handoff are shown in Figure 5-1 and Figure 5-3, respectively. Correspondingly, the simulation model of D-MDMN intra-RAN and inter-RAN handoff are proposed as illustrated in Figure 5-2 and Figure 5-4, respectively. The parameters and configuration attributes of the simulation model can be chosen for different simulation scenarios.

The difference between UMTS and MDMN model is that the central Video Provider outside the CN in the UMTS is distributed and pushed to the edge of the RAN in the MDMN. For the sake of simplicity, one practical topology of multimedia delivery networks in the real world is that each distributed media server is simplified as a multimedia database into each SGSN. IP DiffServ is implemented in each node within both UMTS and MDMN.

Note that since the bandwidth fluctuations and limitation of the wireless channels are what we are more concerned with, we set up the system such that no congestion happens at wired nodes, except for the BSs.

A brief description of the network node and link models and their roles is presented below.

Radio Access Network

RAN is modeled as RNS and Wireless AP in our simulation. RNS node model is shown in Figure 5-5. It consists of RNS (data plane), RNS (control plane), AN router and IP-based AN. The radio functionality of Base Station (Node B) is implemented in the Wireless AP. The data functionalities of BS and RNC are implemented in the RNS (data plane); the control functionalities of BS and RNC are in the RNS (control plane). The protocol stack is illustrated in Figure 3-7. The scale of the radio access network is dependent on the attributes (e.g., packet latency and packet loss ratio) of IP-based AN

cloud model. In our simulation, the packet latency of IP-based AN is configured as exponential distributed with mean value of 15 ms; the packet loss ratio is zero.

The BS acts as the extension of mobile clients. It handles the difference between wireless and wired networks. Each BS is generally responsible for wireless connection setup, handoff support, and medium access control in its service area. For streaming video applications, it also has the responsibility of QoS control, such as rate filtering, scheduling and ARQ, for media streaming.

The Packet Analyzer is used for OPNET ACE Tools to capture packet traces, which will not affect the simulation results.

SGSN/GGSN and Video Provider

SGSN/GGSN node model in UMTS is shown in Figure 5-6. It is composed of SGSN/GGSN (data plane), SGSN/GGSN (control plane) and Access Router. Figure 5-7 shows the SGSN/GGSN node model in MDMN, where Video Providers are distributed as multimedia databases. The protocol stacks of SGSN/GGSN and Video Provider are illustrated in Figure 3-7.

IP backbone core network

The scale of the CN is dependent on the attributes (e.g., packet latency and packet loss ratio) of IP backbone CN cloud model. The packet latency of IP backbone CN is configured as exponential distribution with mean value of 20 ms; the packet loss ratio is zero.

Mobile Station

MS node model is shown in Figure 5-8, which is implemented according to the protocol stack illustrated in Figure 3-7. The RFL, RNL, ip, ip_encap, tcp, udp, rtp and application layer processors are taken from OPNET Modeler's library. Each MS supports three types of services: video streaming service, voice service and web service. The voice and web services are configured as background traffic which are established between IP Phone User and Voice User, Web Server and Web User, respectively.

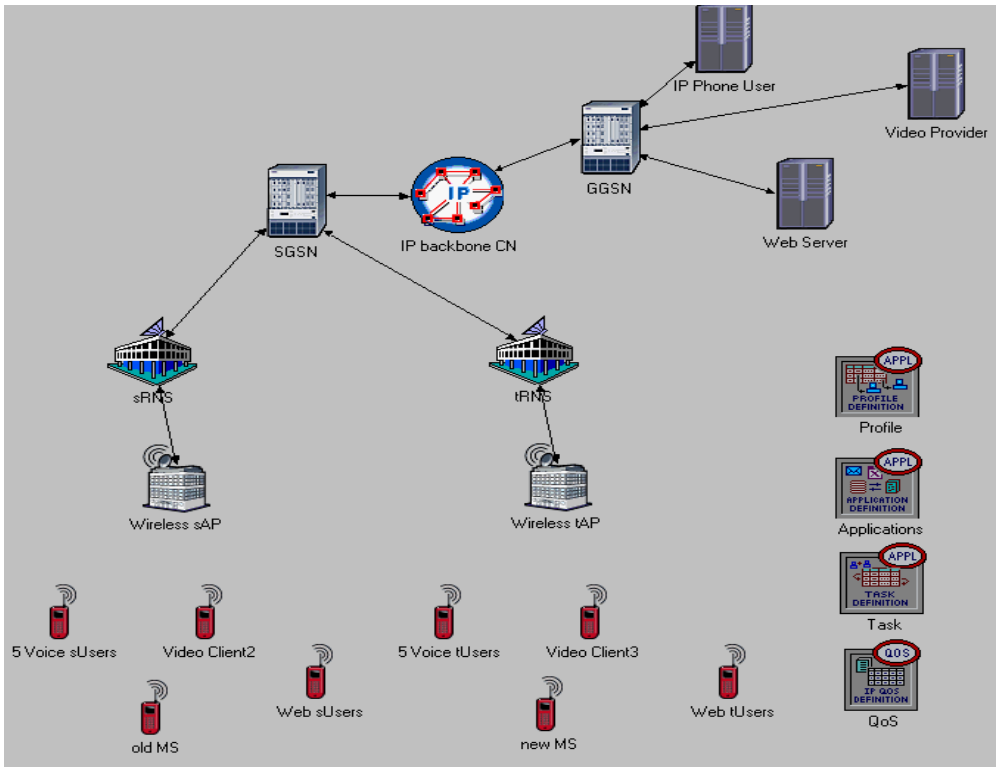


Figure 5-1 UMTS simulation model (Intra-RAN Handoff)

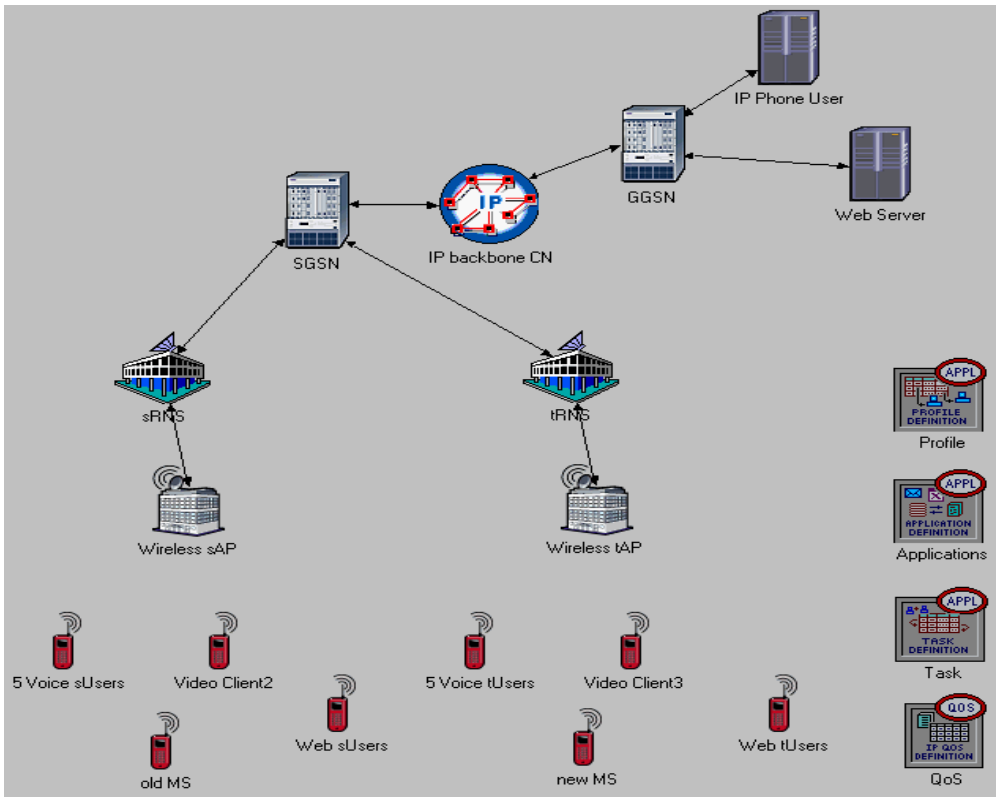


Figure 5-2 MDMN simulation model (Intra-RAN Handoff)

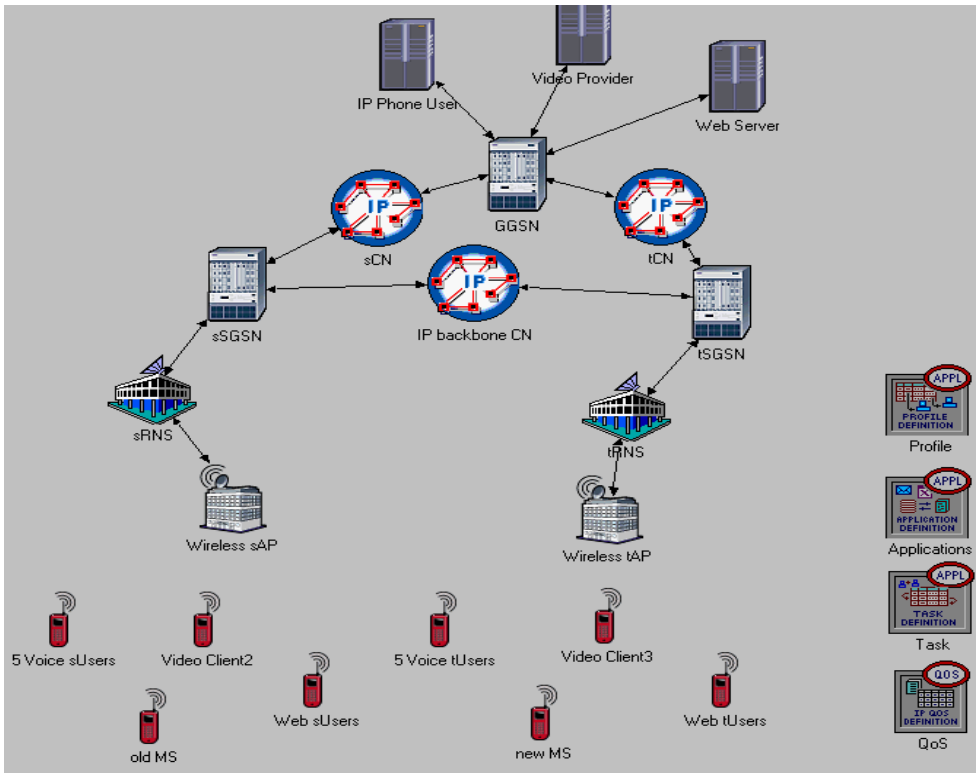


Figure 5-3 UMTS simulation model (Inter-RAN Handoff)

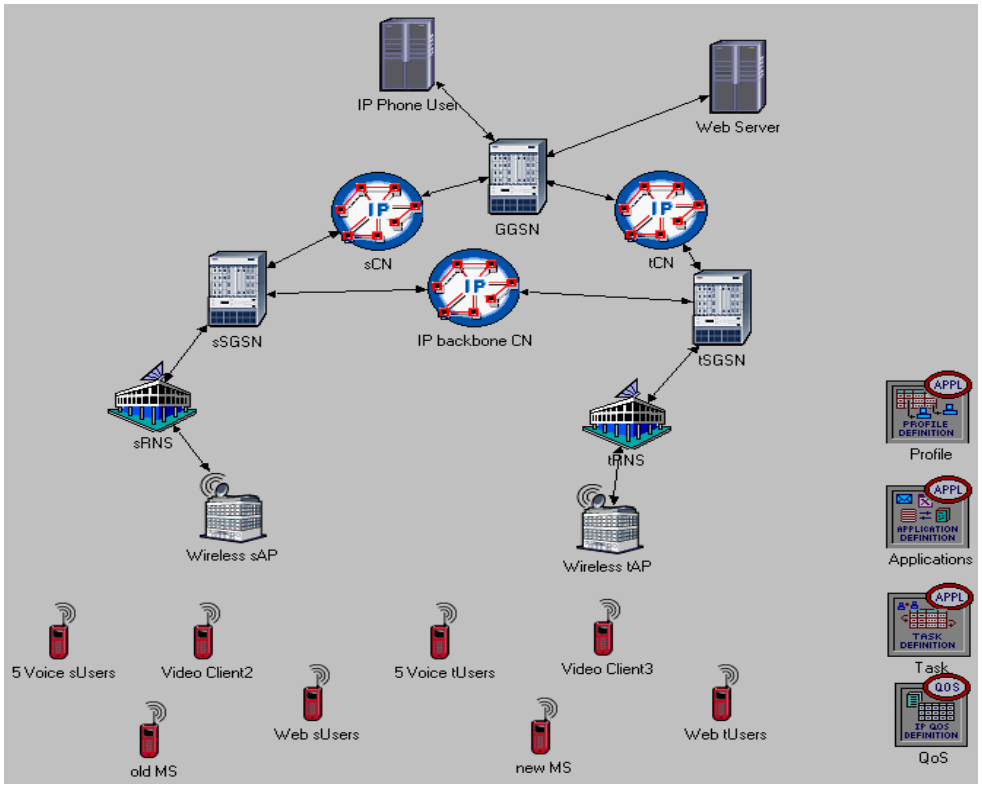


Figure 5-4 MDMN simulation model (Inter-RAN Handoff)

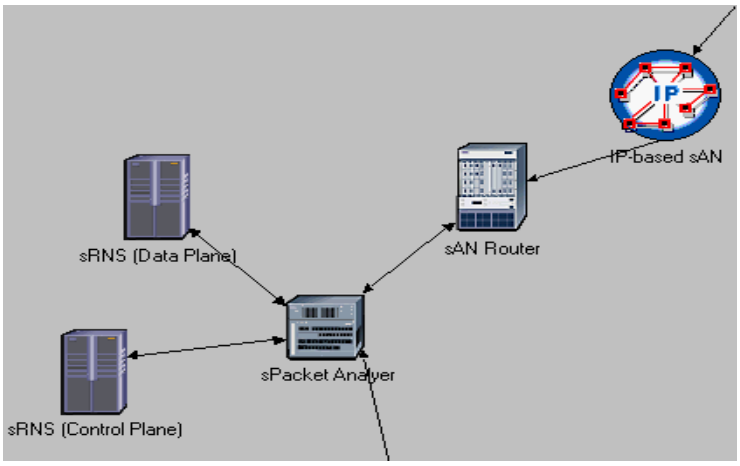


Figure 5-5 Node Model of RNS

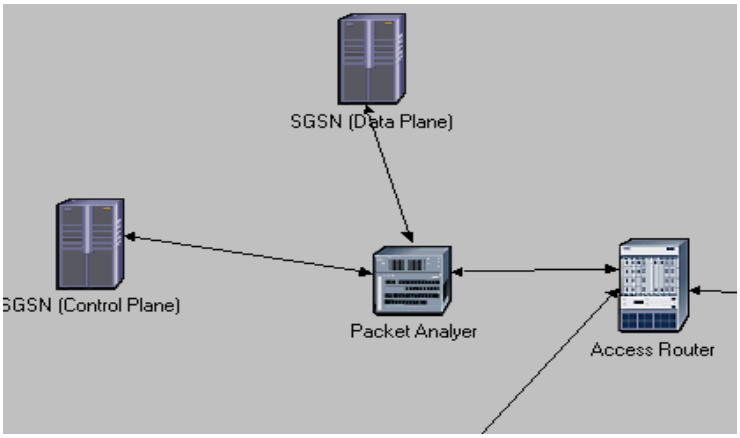


Figure 5-6 Node Model of SGSN (or GGSN)

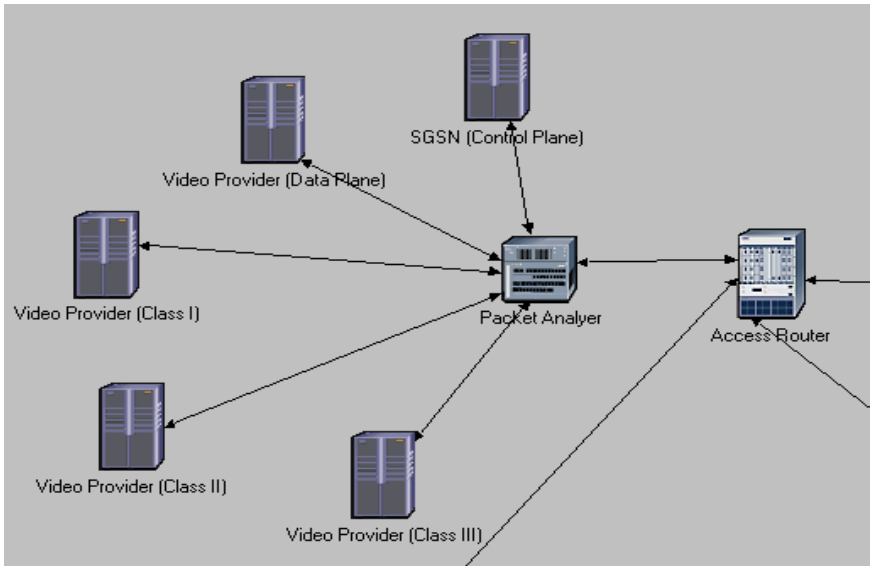


Figure 5-7 Node Model of SGSN in the proposed MDMN

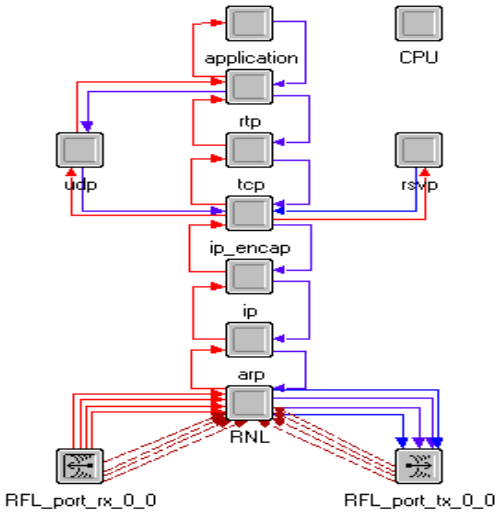


Figure 5-8 Node Model of MS

Wired Links

The PPP point-to-point links at OC3 data rate (155Mbps) are used between nodes with serial interfaces (e.g., routers with PPP ports) within the RAN and CN. The video provider, IP phone user or web server is connected to a mobile network access point (e.g., GGSN) by using a 100BaseT duplex link at 100 Mbps.

Radio Links

Unlike point-to-point links, radio links are not statically represented; that is, they cannot be seen in the network model. Instead, radio links are dynamically established during simulation. Radio links exist between any radio transmitter-receiver channel pair, but establishing a link depends on many physical characteristics of the components involved, as well as time-varying parameters. During simulation, parameters such as frequency band, modulation type, transmitter power, distance, and antenna directionality are common factors that determine whether a radio link exists at a particular time or can ever exist.

OPNET uses the radio transceiver pipeline model, shown in Figure 5-9, to model wireless transmission of packets. It consists of thirteen stages. The attributes of each stage are configured as shown in Figure 5-10. The stage 10 to stage 13 in the radio pipeline model which are related to wireless BER will be discussed in Section 5.2. For more details of each stage in the radio pipeline model, please refer to the OPNET Modeler 9.0 documents [77].

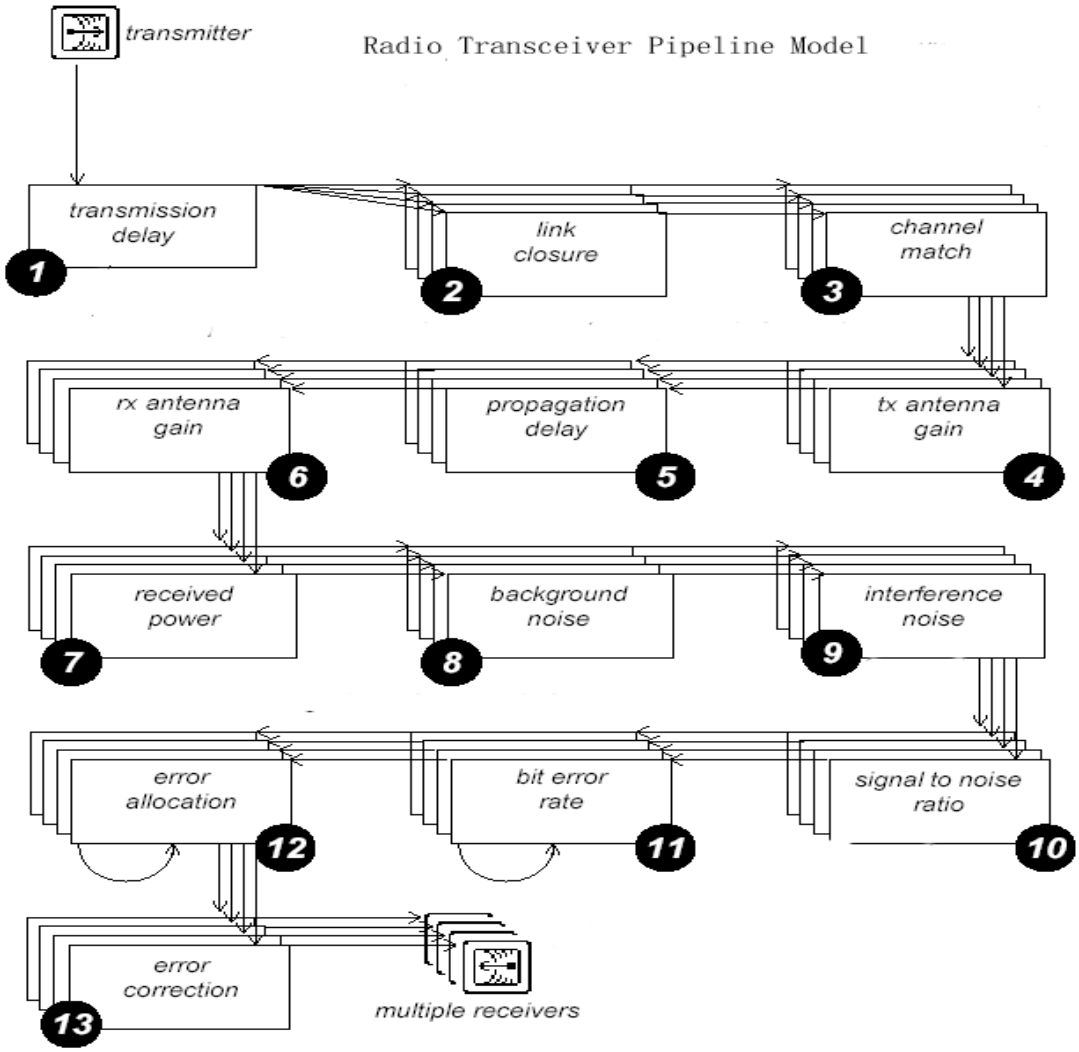


Figure 5-9 Radio Transceiver Pipeline Model

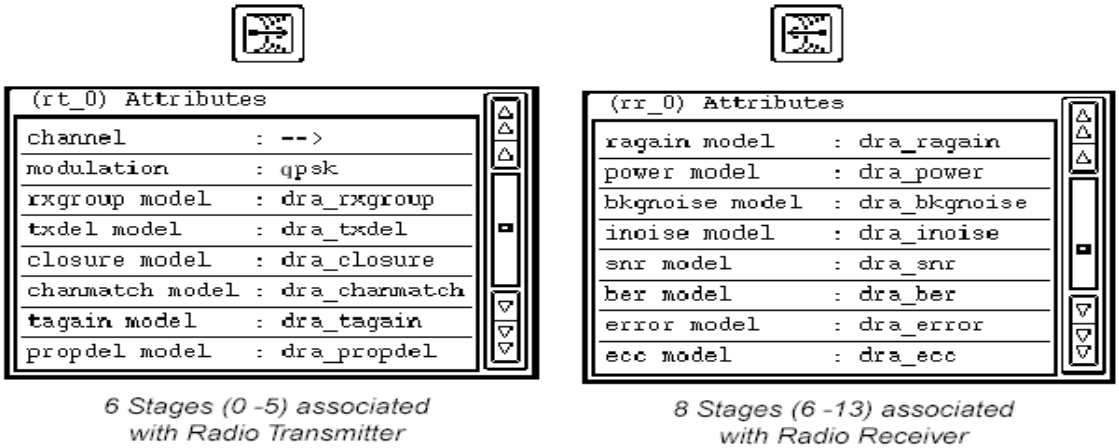


Figure 5-10 Radio Transceiver Attributes for Specifying Pipeline Stages

5.2 System Setup and Test Conditions

5.2.1 Rate Shaping (Filtering) and Packetization

The purpose of rate shaper (filter) [1] [2] [3] [23] includes: optimization of bandwidth usage, adoption of filter for handling client heterogeneity, network heterogeneity, and optimizations in the retrieval of stored media.

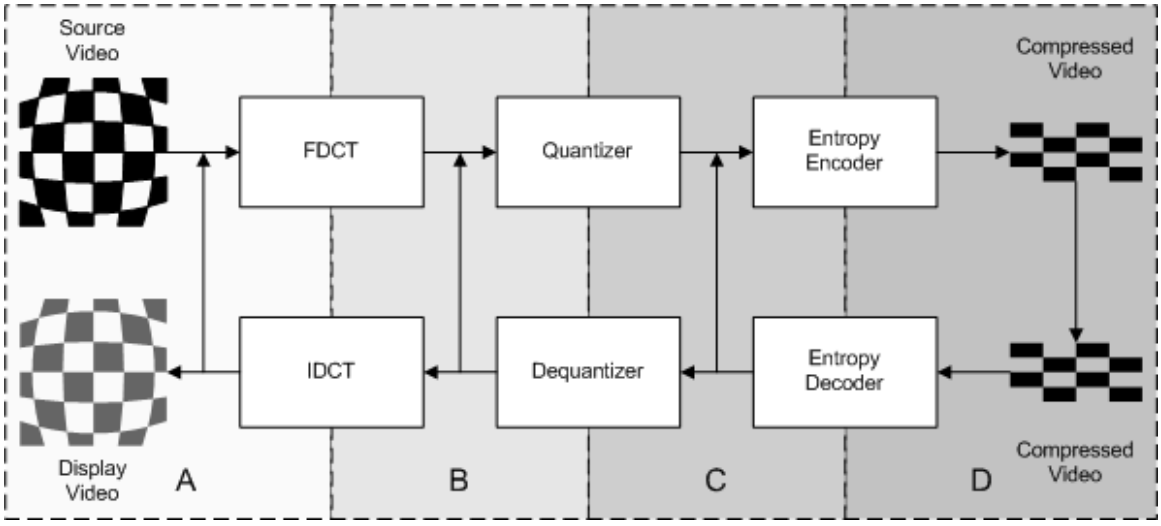


Figure 5-11 Level of rate shaping

In Figure 5-11, the points at which rate shaping can be performed on the compressed bit-stream are illustrated.

Region A is the uncompressed raw video where rate shaping can be preformed relatively simply, e.g., resizing/stretching, but the amount of data that has to be processed is very large due to its uncompressed nature.

In region B, the data is the same size as the raw data, but if the bit-stream is being decompressed in order to accomplish a particular rate shaping and then recompressed (e.g., re-quantization filtering), performing filtering at this point saves completing the computationally intensively functions of forward-DCT (FDCT) and inverse-DCT (IDCT).

In region C, the many zeros produced by the FDCT have been removed and the data is considerably smaller, the functions of frequency filtering that are feasible at this point

include: low-pass filtering, color-reduction filtering, color to monochrome conversion and simple coder-conversions. The current transcoder which is designed for speed therefore operates in this region.

Compared with region A, B and C, rate shaping (e.g., frame dropping, layer dropping) on the fully compressed data in region D is standard specific and relatively simple. We employ the techniques of rate shaping in the sender directly to video distribution services as network filtering used in the BSs. For simplicity of simulation, we packetize each frame of every layer into one packet in order to mimic the principle of a frame dropper in the BSs.

5.2.2 BER over Rayleigh Fading Channels

As mentioned previous, the BER on wireless channels is computed at the BER stage (stage 11) in the radio pipeline model. In general, the bit error rate provided by this stage is a function of the type of modulation used for the transmitted signal. This stage evaluates BER based on the previously computed average SNR and also accounts for processing gain at the receiver.

The SNR (in dB) is given by

$$SNR = 10 \log [P_r / (N_b + N_i)]$$

where P_r = received power (Watts), which is computed in stage 7;

N_b = background noise power (Watts), which is computed in stage 8;

N_i = interference noise power (Watts), which is computed in stage 9.

The SNR value is added to the processing gain (also in dB) to obtain the effective SNR. This effective SNR is also written as E_b / N_0 where E_b is the received energy per bit (in Joules) and N_0 is the noise power spectral density (in Watts/hertz). The bit error rate is derived from the effective SNR based on the QPSK (downlink) / BPSK (uplink) modulation curve assigned to the receiver in a Rayleigh fading channel. The probability of bit error P_b (i.e., BER) is given by [52]

$$P_b(E_b/N_0) = \frac{1}{2}\left[1 - \sqrt{\frac{E_b/N_0}{1 + E_b/N_0}}\right]$$

Figure 5-12 shows the BER curves for QPSK (downlink) and BPSK (uplink) in a Rayleigh fading channel and an AWGN channel.

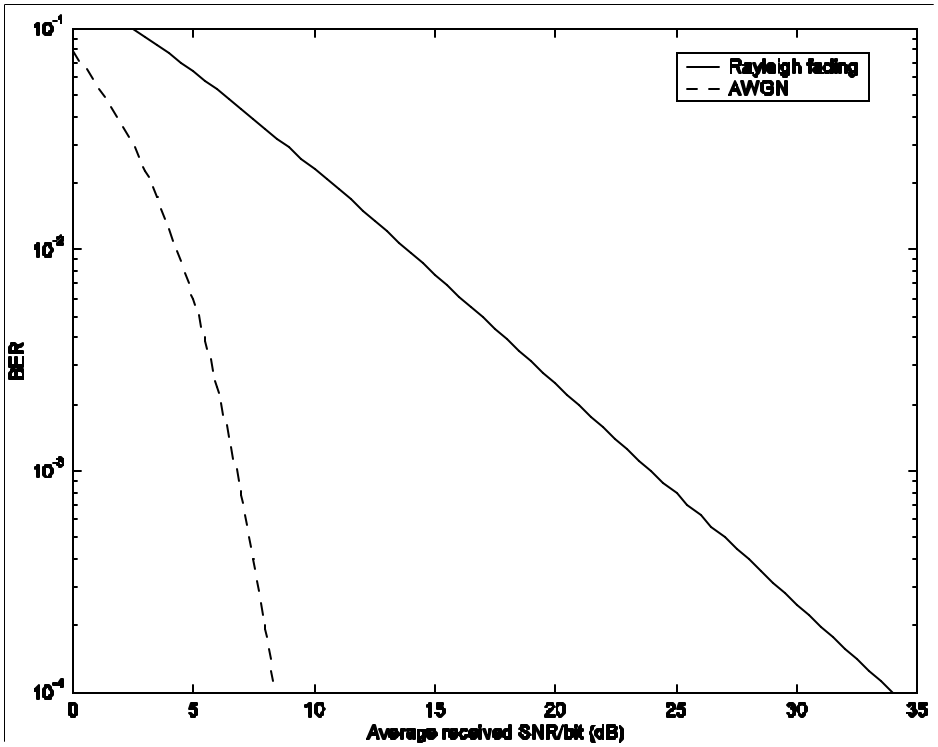


Figure 5-12 BER performance over flat Rayleigh fading channels

Following the BER stage, the error allocation stage (stage 12) translates the given BER computed in stage 11 into an actual set of bit errors for each valid packet which is received. The approach taken in stage 12 is to avoid sequencing through all the bits in the packet and, therefore, not to generate positional information about bit errors, but to still accurately compute the overall number of bit errors. This should be done with the minimum number of computations for simulation efficiency considerations.

The error allocation algorithm is based on the expression for the probability of exactly k bit errors occurring in a packet segment of length N. This probability is denoted as P_k. Under the assumption of independent bit errors, given the wireless channel bit error probability P_b (i.e., BER), P_k can be expressed by

$$P_k = P_b^{\,k}(1-P_b)^{N-k}\binom{N}{k}$$

$$\sum_{k=0}^{N} P_k \geq r$$

The algorithm first generates a random number r between 0 and 1, to provide a value against which the probability of occurrences of different numbers of bit errors will be

tested. Next the algorithm begins iterations. First the probability that 0 bit error occurs is computed according to the formula above, and compared to *r*. If *r* is lower than this probability, then 0 is the number of bit errors in the packet. Else, the probability of occurrence of 1 or fewer bit errors is computed. If *r* is less than this probability, yet higher than the probability of occurrence of the previous number of bit errors, then 1 is the number of bit errors allocated to the packet. The algorithm continues iterating in this manner until a value *k* is found for which the probability of *k* or fewer errors occurring is greater than *r*. Then the number of errors assigned to the packet is *k*.

The purpose of error collection stage (stage 13) is to determine whether or not the arriving packet can be accepted in the destination node. This is usually dependent upon whether the packet has experienced collisions, the result computed in the error allocation stage, and the capability of the receiver to correct the errors affecting the packet. In our simulation, there are no error collection techniques implemented except for ARQ.

The following three cases in Table 5-1 are tested for the evaluation of the relationship between BER and frame error rate.

Table 5-1 Test Cases of BER and FER in wireless channels

Test Case	Channel Model	Modulation	Multiple Access	BER
a	Rayleigh Fading	QPSK (DL), BPSK(UL)	DS-CDMA	$[10^{-5}, 2 \times 10^{-5}]$
b	Rayleigh Fading	QPSK (DL), BPSK(UL)	DS-CDMA	$[10^{-4}, 2 \times 10^{-4}]$
g	Rayleigh Fading	QPSK (DL), BPSK(UL)	DS-CDMA	$[.6 \times 10^{-3}, 10^{-3}]$

5.2.3 Delay-constrained ARQ

To enhance the video quality in the presence of unavoidable packet loss or bit errors, error control mechanisms have been proposed. Basically, error control approaches can be broadly categorized as open-loop error control (e.g., SMDC, error resilience tools in MPEG-4, error concealment) and close-loop error control, (e.g., delay-constrained ARQ). However, because of the complexity of open-loop error control, multiple description coding, error resilience tools in MPEG-4 and error concealment can not be emulated in this simulation. Instead, delay-constrained ARQ is under consideration.

According to the decision-making points, delay-constrained retransmission can be sender-based, receiver-driven, or hybrid. Since the computational complexity is limited at the mobile handset, the sender-based control at the BS is chosen to suppress retransmission of packets that will miss their display time at the mobile user.

Given the maximum transfer delay D_{max} in UMTS, the wired link delay D_{wired} and the wireless link $RTT_{wireless}$, the maximum number of retransmissions N_{ARQ} allowed for a video packet is approximately given by

$$N_{ARQ} = \begin{cases} \left\lfloor \dfrac{D_{\max} - D_{wired}}{RTT_{wireless}} \right\rfloor, & if \quad D_{\max} > D_{wired} \\ 0, & if \quad D_{\max} \le D_{wired} \end{cases}$$

Typically, the wireless link $RTT_{wireless}$ is about several milliseconds. The BS can calculates the wired link delay D_{wired} from Video Provider to the BS by recording the time T when a RTCP sender report (SR) is received, and then subtracting the value of NTP timestamp field in SR to obtain $D_{wired} = (T - NTP)$.

5.2.4 Traffic Profile

In order to evaluate the effectiveness of the proposed IP DiffServ MPEG-4 video marking algorithm and the streaming video handoff procedures under MDMN, two groups of traffic parameters have been set up and are listed in Table 5-2 and Table 5-4, respectively.

Traffic Profile for Evaluation of Video Marking Algorithm

The video traffic, shown in Figure 5-14, is made up of Class I, Class II and Class III layer-coded video streams, shown in Figure 5-13, which are classified through the proposed IP DiffServ MPEG-4 video marking algorithm. The video traffic which is generated by OPNET has to be subjected to the requirements of UMTS bearer service attributes of streaming class [61], listed in Table 5-3.

The background traffic in each cell is composed of one video client, 25 voice users, and web users, shown in Figure 5-15 ~ Figure 5-17, respectively. Under the condition of background traffic, the serious congestion will occur in the BSs due to the limitation and fluctuation of wireless bandwidth.

Table 5-2 Traffic profile of each cell for evaluation of video marking algorithm

Traffic Type	User Number	Sending Rate (mean)	Packet Length (mean)	Standard
Video Stream*	2	240 kbps per client	1 Kbytes	MPEG-4
Voice Stream	25	16 kbps per user	200 Bytes	G.728
Web Traffic	-	400 kbps	1 Kbytes	HTTP

* Frame Rate (mean) = 20 frame/s, Frame Length (mean) = 1 Kbytes.

Table 5-3 UMTS bearer service attributes of streaming class

Maximum bit rate (outdoor)	384 kbps
Maximum SDU size	1500 Bytes
SDU error ratio (i.e., Frame error rate)	$\leq 10\,\%$
Transfer delay	≤ 280 ms
Delay jitter	≤ 50 ms (Frame Rate = 20 frame/s)

Traffic Profile for Evaluation of Streaming Video Handoff

Table 5-4 Traffic profile of each cell for evaluation of streaming video handoff

Traffic Type	User Number	Sending Rate (mean)	Packet Length (mean)	Standard
Video Stream*	2	240 kbps per client	1 Kbytes	MPEG-4
Voice Stream	5	16 kbps per user	200 Bytes	G.728
Web Traffic	-	400 kbps	1 Kbytes	HTTP

* Frame Rate (mean) = 20 frame/s, Frame Length (mean) = 1 Kbytes.

The video traffic, shown in Figure 5-18, is also made up of Class I, Class II and Class III layer-coded video streams and has to be subjected to the requirements of UMTS bearer service attributes of streaming class [61], listed in Table 5-3. The hard handoff will occur 28 times during the testing period of 5 minutes.

The background traffic in each cell is composed of one video client, 5 voice users, and web users, shown in Figure 5-19 ~ Figure 5-21, respectively. With the decrease of background traffic, no congestion happens in the BSs during the handoff tests.

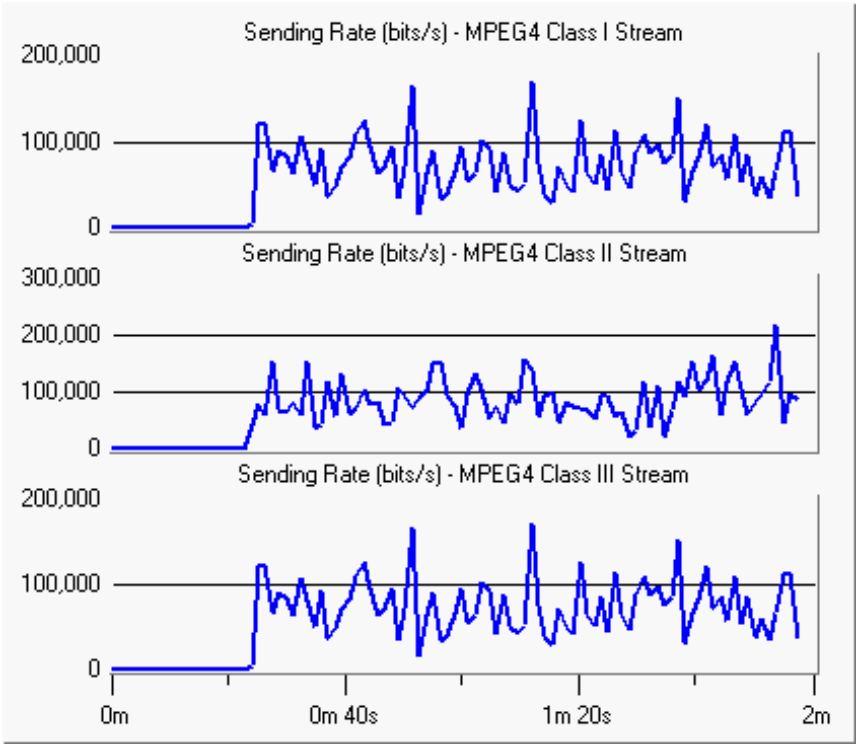


Figure 5-13 Layered Video Traffic (Video Client1)

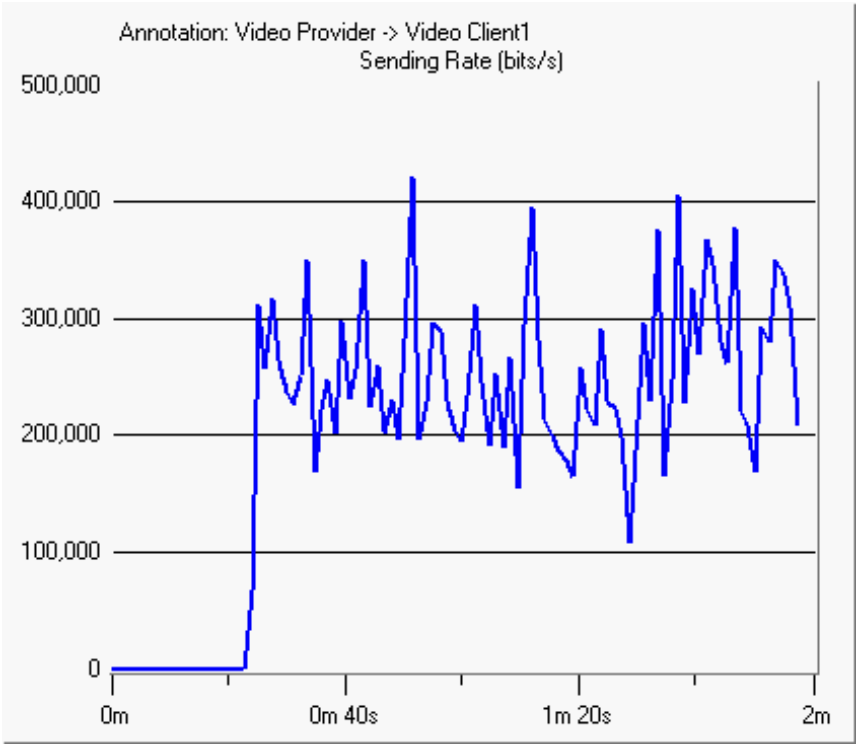


Figure 5-14 Video Traffic (Video Client1)

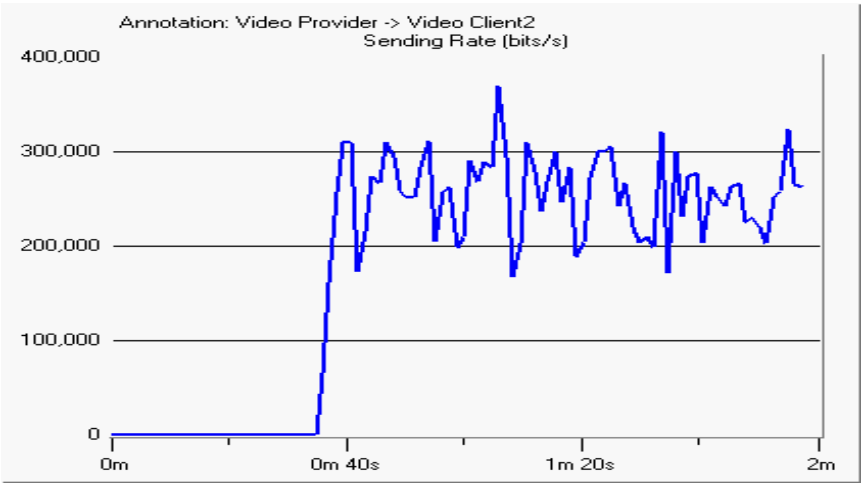


Figure 5-15 Background Traffic (Video Client2)

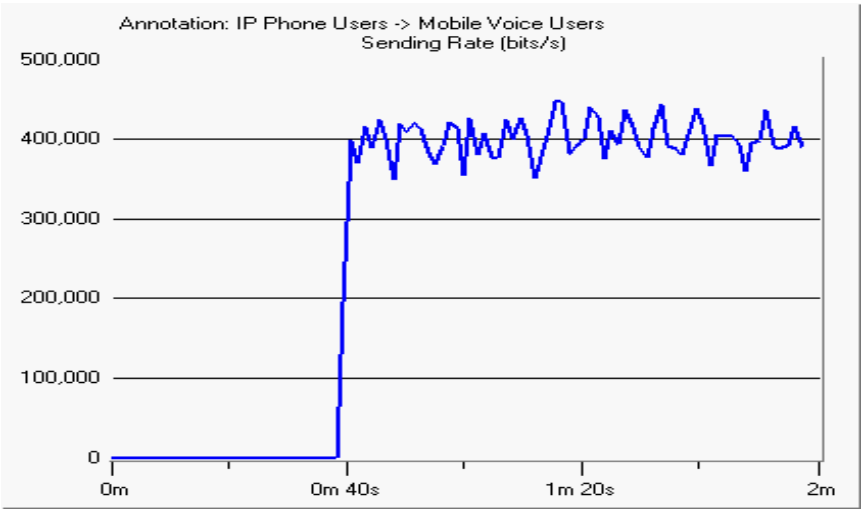


Figure 5-16 Background Traffic (25 Voice Users)

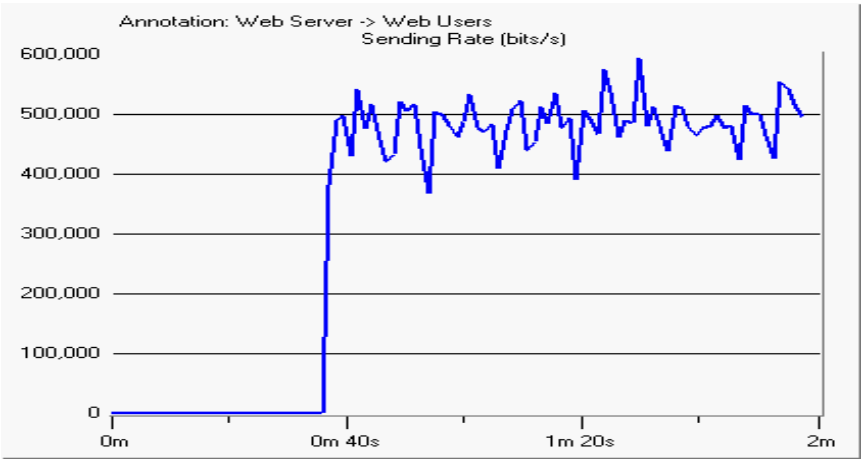


Figure 5-17 Background Traffic (Web Users)

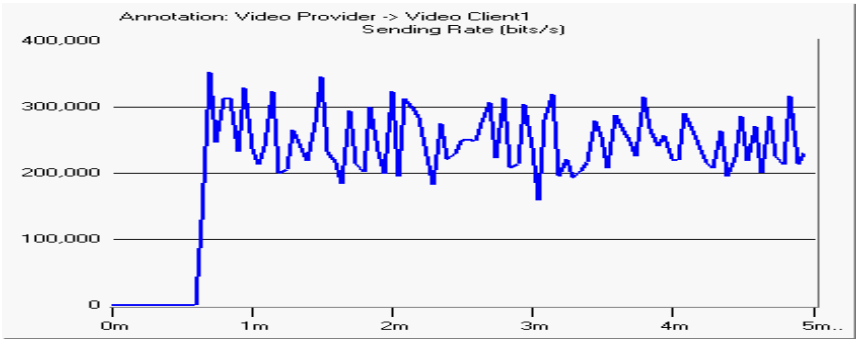


Figure 5-18 Video Traffic (Handoff occurs 28 times)

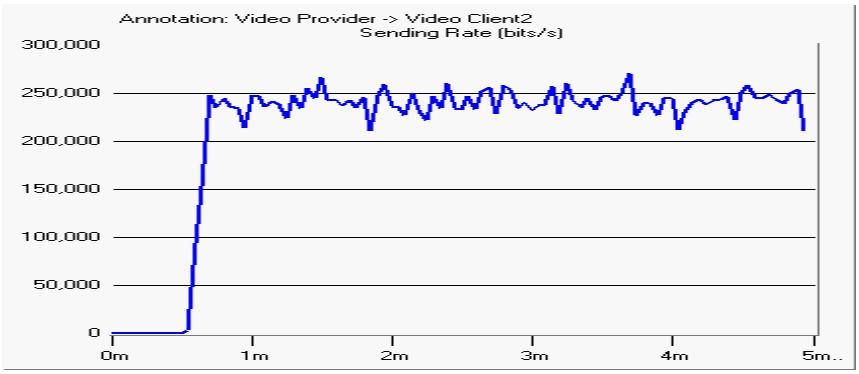


Figure 5-19 Background Traffic (Video Clients)

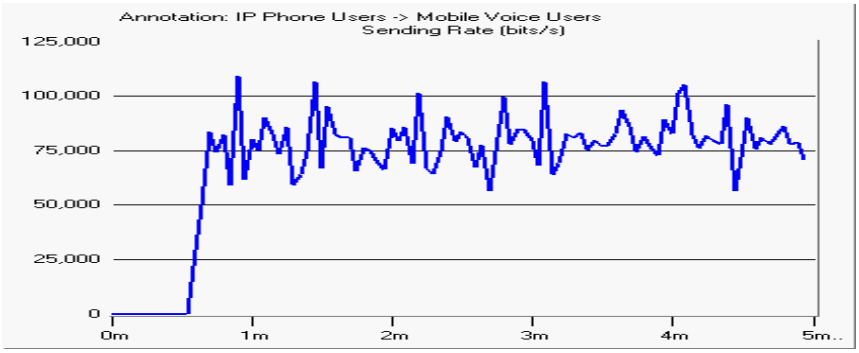


Figure 5-20 Background Traffic (5 Voice Users)

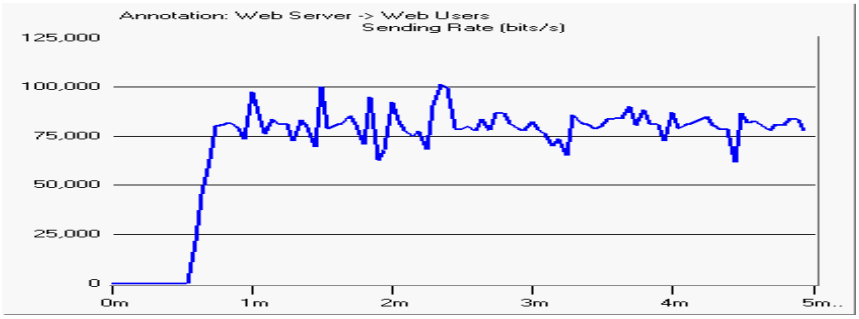


Figure 5-21 Background Traffic (Web Users)

5.2.5 Test Cases

AF Queue Size

To implement the proposed IP DiffServ video marking algorithm in each node, we use WFQ discipline to classify and schedule the incoming packet into and out of EF, AF1, AF2, AF3 or BE queue. The WFQ profile for the proposed IP DiffServ video marking algorithm in each node is listed in Table 5-5. The selection of buffer size is crucial to video over IP network. An optimum buffer size has to be found which balances both end-to-end delay and packet loss ratio to tolerable levels. If the buffer is set too low, some packets may be lost; if set too high, higher delays result.

Table 5-5 WFQ profile for proposed IP DiffServ video marking algorithm

Queue Scheduling	Queue Classification	Queue Size (1 unit = 100 ms)	Normalized Bandwidth
WFQ	BE Queue	15	4.5 %
	AF3 Queue	1~9	9.1 %
	AF2 Queue	1~9	13.7 %
	AF1 Queue	1~9	22.7 %
	EF Queue	15	50 %

There are eighteen cases, listed in Table 5-6, under test in order to evaluate the effect of AF queue size. The effect of AF queue size is analyzed in Section 5.2.5. After the evaluation of the effect of AF queue size, we select 500 ms as the optimum queue size of the AF1, AF2 and AF3 queue.

Video Marking Algorithm

In order to compare the proposed IP DiffServ video marking algorithm with DVMA, three scenarios are experimented. The first one is the best effort model using a drop-tail BE queue in each node. The second one is the DiffServ model where DVMA is used with WRED AF queue. The third one is the DiffServ model where the proposed IP DiffServ video marking algorithm is used with the drop-tail AF1, AF2 and AF3 queues. Table **5-7** lists the configuration of queue scheduling in these three scenarios. Table 5-8 gives the

WRED profile for AF Queue in the DVMA scenario. The test cases of video marking algorithm are listed in Table 5-9.

Streaming Video Handoff

The performance of the proposed streaming video handoff in D-MDMN is examined and compared with that in UMTS model under the scenario of the proposed IP DiffServ video marking algorithm. The test cases of streaming video handoff are listed in Table 5-10.

Table 5-6 Test Cases of Effect of Queue Size

Test Case	Physical Characteristics	Rayleigh BER	QoS Mechanism	AF Queue Size
1	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	100 ms
2	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	200 ms
3	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	300 ms
4	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	400 ms
5	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	500 ms
6	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	600 ms
7	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	700 ms
8	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	800 ms
9	QPFK(DL),DS-CDMA	10^{-5}	DiffServ, WFQ	900 ms
(1)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	100 ms
(2)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	200 ms
(3)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	300 ms
(4)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	400 ms
(5)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	500 ms
(6)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	600 ms
(7)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	700 ms
(8)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	800 ms
(9)	QPFK(DL),DS-CDMA	10^{-4}	DiffServ, WFQ	900 ms

Table 5-7 Queue scheduling configuration

Solution	Queue Scheduling	Queue Classification	Queue Size (1 unit = 100ms)	Normalized Bandwidth
Best Effort	FIFO	BE Queue	45	100 %
DVMA	WRED+CBQ	BE Queue	15	4.5 %
		AF Queue	15	45.5 %
		EF Queue	15	50 %
Proposed	WFQ	BE Queue	15	4.5 %
		AF3 Queue	5	9.1 %
		AF2 Queue	5	13.7 %
		AF1 Queue	5	22.7 %
		EF Queue	15	50 %

Table 5-8 WRED profile for AF queue in DVMA

Video Stream Class	Exponential Weight Factor	Min Threshold (1 unit = 100 ms)	Max Threshold (1 unit = 100 ms)	Mark Probability Denominator
AF31	9	1	5	5
AF21	9	2	5	5
AF11	9	3	5	5

Table 5-9 Test Cases of Video Marking Algorithm

Test Case	Physical Characteristics	Rayleigh BER	QoS Mechanism	Queue Scheduling	Marking Algorithm
A	QPFK(DL), DS-CDMA	10^{-5}	Best Effort IP	FIFO	BE
B	QPFK(DL), DS-CDMA	10^{-5}	IP DiffServ	WRED	DVMA
C	QPFK(DL), DS-CDMA	10^{-5}	IP DiffServ	WFQ	Proposed
a	QPFK(DL), DS-CDMA	10^{-4}	Best Effort IP	FIFO	BE
b	QPFK(DL), DS-CDMA	10^{-4}	IP DiffServ	WRED	DVMA
c	QPFK(DL), DS-CDMA	10^{-4}	IP DiffServ	WFQ	Proposed

Table 5-10 Test Cases of Handoff Procedures

Test Case	Modulation	Rayleigh BER	Network Model	QoS Mechanism	Handoff
I	QPSK(DL), BPSK(UL)	10^{-5}	UMTS	DiffServ, WFQ	Intra-RAN
II	QPSK(DL), BPSK(UL)	10^{-5}	MDMN	DiffServ, WFQ	Intra-RAN
III	QPSK(DL), BPSK(UL)	10^{-5}	UMTS	DiffServ, WFQ	Inter-RAN
IV	QPSK(DL), BPSK(UL)	10^{-5}	MDMN	DiffServ, WFQ	Inter-RAN
i	QPSK(DL), BPSK(UL)	10^{-4}	UMTS	DiffServ, WFQ	Intra-RAN
ii	QPSK(DL), BPSK(UL)	10^{-4}	MDMN	DiffServ, WFQ	Intra-RAN
iii	QPSK(DL), BPSK(UL)	10^{-4}	UMTS	DiffServ, WFQ	Inter-RAN
iv	QPSK(DL), BPSK(UL)	10^{-4}	MDMN	DiffServ, WFQ	Inter-RAN

5.3 Simulation Results and Analysis

5.3.1 Effect of BER on FER over Wireless Channels

Figure 5-22 ~ Figure 5-27 show the effect of BER on FER over the Rayleigh fading channel. The FER measure indicates the difficulty of sending video stream over the wireless channel. A small BER translates into a much higher FER. For example, the BER mean value of 1.36×10^{-4} in Figure 5-23 can translate into the FER mean value of 3.16% in Figure 5-26. In the test case *a* with a BER mean value of 1.57×10^{-5}, there are no frame errors showed in Figure 5-25. That is because we use the delay-constrained ARQ to combat Rayleigh BER.

Given that the maximum frame error rate allowed in UMTS is 10%, the video delivery in UMTS can tolerate Rayleigh BER up to 10^{-4} with delay-constrained ARQ. That is reason that we choose BER in the range of $[10^{-5}, 2\times10^{-5}]$ and $[10^{-4}, 2\times10^{-4}]$ as our test conditions.

However, from Table 1-1, the BER of wireless video can be as high as 10^{-3}. It is desirable to employ the proposed scalable multiple description coding as a diversity technique to combat the high BER in the wireless channel.

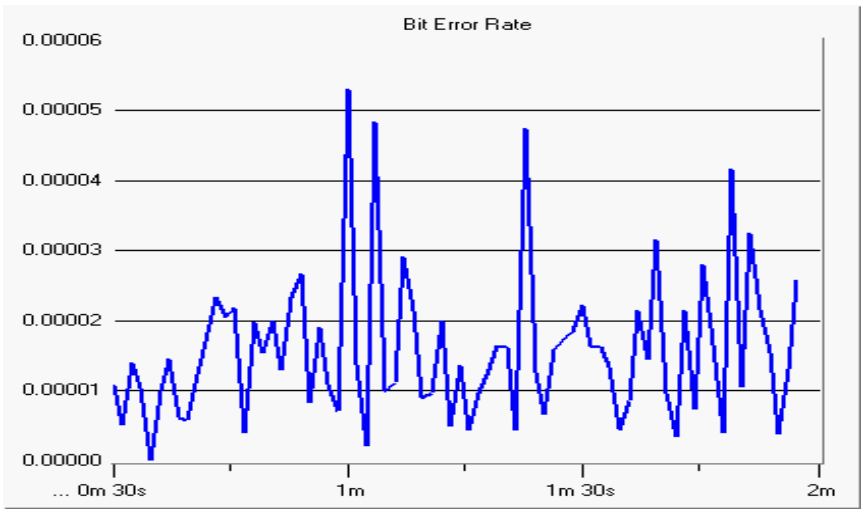


Figure 5-22 Bit Error Rate over the Rayleigh fading channel (Test Case *a*)

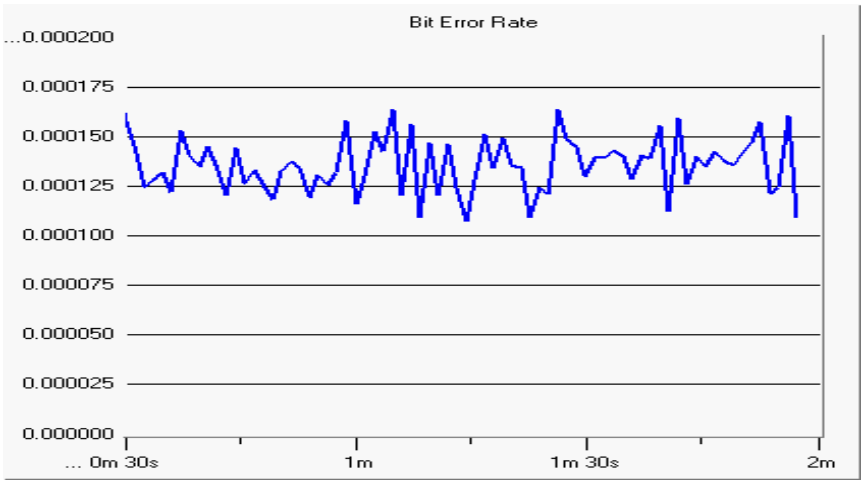


Figure 5-23 Bit Error Rate over the Rayleigh fading channel (Test Case *b*)

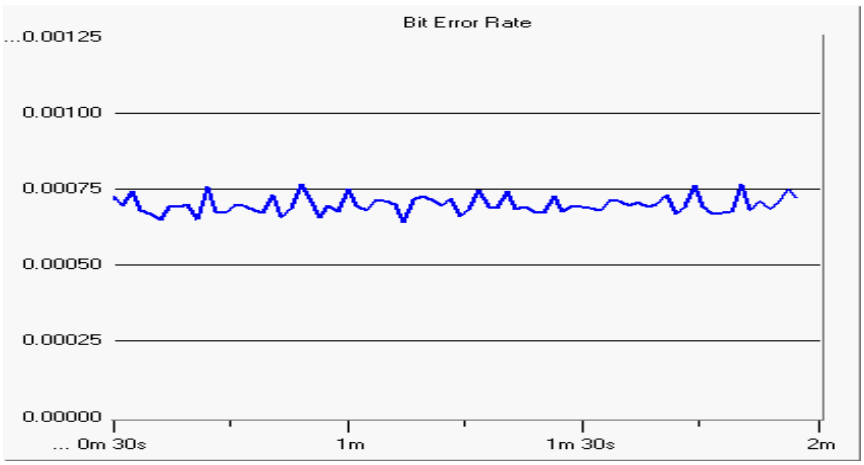


Figure 5-24 Bit Error Rate over the Rayleigh fading channel (Test Case *g*)

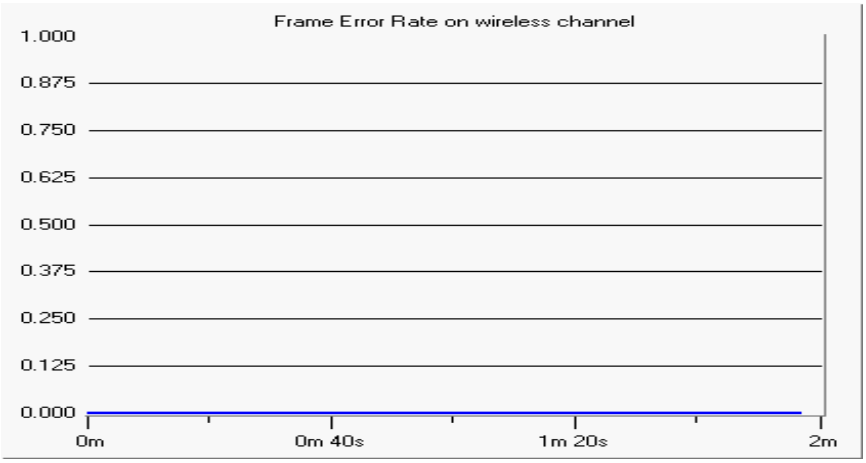


Figure 5-25 FER over the Rayleigh fading channel (Test Case *a*)

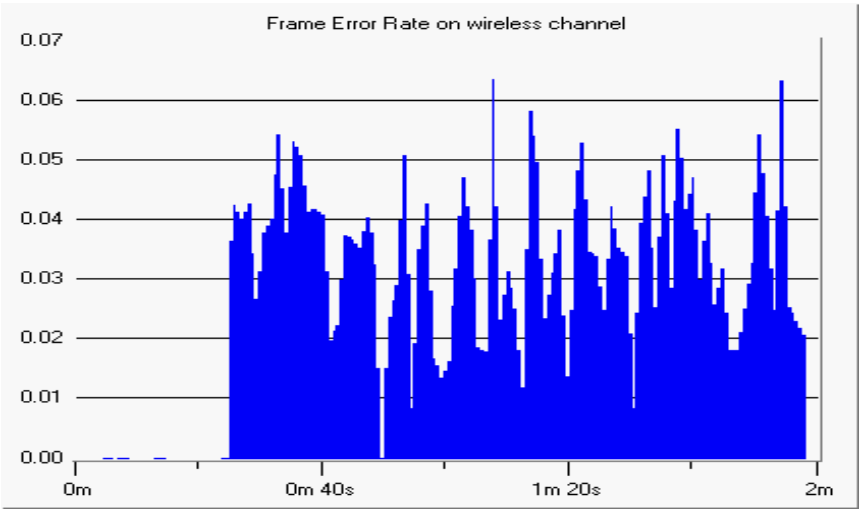


Figure 5-26 FER over the Rayleigh fading channel (Test Case *b*)

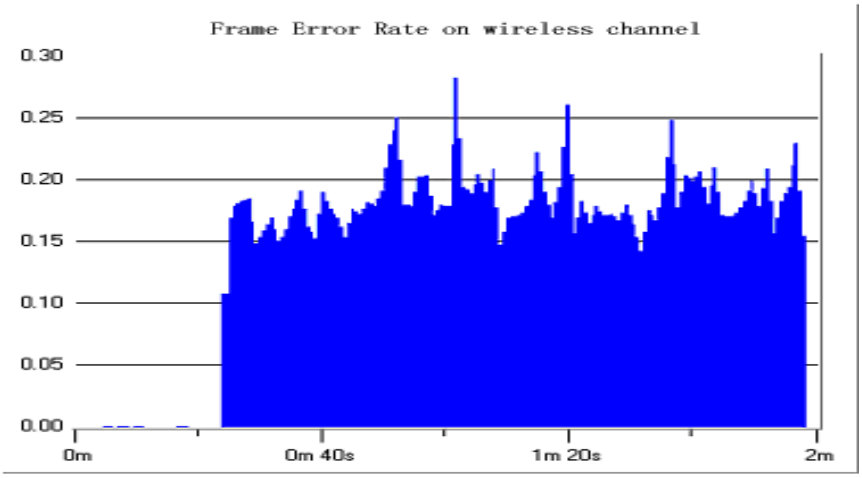


Figure 5-27 FER over the Rayleigh fading channel (Test Case *g*)

5.3.2 Effect of AF Queue Size

Figure 5-28 ~ Figure 5-31 illustrate the effect of the AF queue size on packet end-to-end delay and packet loss ratio of AF queue at the BSs. The video traffic begins at time 25 s and the background traffic starts at time 36 sec. From time 25 s to 36 s, the packet arriving rate at BS is lower than the queue service rate and there is no congestion in the AF queues. At time 36 s, the packet arriving rate at BS suddenly increases due to the background traffic and exceeds the queue service rate. The AF queues continue growing until time 50 s and start to drop frames.

As the AF queue size increases, the packet end-to-end delay increases, but the packet loss ratio of AF queue decreases. Compared the results of total nine test cases, we choose queue size 500 ms as the optimum value which balances both end-to-end delay and packet loss ratio to tolerable levels for further simulation setting.

5.3.3 Effect of Video Marking Algorithm

Performance of Frame Error Rate

MPEG-4 video stream in UMTS can tolerance SDU error rate (i.e., FER) up to 10%. The effects of different video marking algorithms on FER are shown in Figure 5-32 ~ Figure 5-37.

In the scenario of BER = 10^{-5}, the FERs are caused by the packet loss at the BSs due to congestion. In the scenario of BER = 10^{-4}, the FERs are caused by both the packet loss at the BSs due to congestion and the wireless channel bit errors. Due to the employment of WRED for proactive packet-dropping in DVMA, the packet loss in DVMA begins earlier than that in the proposed solution. However, the packet loss in the tail-dropping BE solution occurrs even earlier than that in DVMA. That is because the background traffic and the video traffic enter into the same queue, and cause the congestion happened earlier. Note that the background traffic and the video traffic are separated in different queues in DiffServ-based solutions.

Since we re-organize the shape, motion, and texture video information into different layers, UEP can be introduced and results in differentiated service. If a bit error or a

packet loss occurs in the MPEG-4 Class I stream, the corresponding bits or packets in the MPEG-4 Class II and Class III streams have to be considered erroneous or lost. Similarly, if a bit error or a packet loss occurs in the MPEG-4 Class II stream, the corresponding bits or packets in the MPEG-4 Class III stream also have to be considered erroneous or lost. Some packets may arrive late and will also be considered lost. If the higher priority traffic is protected, less packet loss (i.e., FER) will occur.

Compared with BE and DVMA, the protection of both voice stream and MPEG-4 Class I stream in the proposed solution is the best. This also results in the least FER among all the scenarios. In addition, the protection of EF traffic in DVMA is better than that in BE. However, this is at the cost of high FER of AF traffic in DVMA.

A brief comparison of QoS (e.g., FER) guarantee in the three solutions can be summarized in Table 5-11. Note that: "Unacceptable QoS" means that the performance of FER can not meet the QoS requirement all the time; "Unpredictable QoS" is a typical characteristic of best-effort traffic; "Unwarranted QoS" means that the performance of FER meets the QoS requirement sometimes and is predictable; "Guaranteed QoS" means that the performance of FER meets the QoS requirement all the time and is predictable.

Table 5-11 Comparison of QoS (e.g., FER) guarantee in three solutions

Traffic	BE	DVMA	Proposed
Voice Stream	Unacceptable QoS	Unwarranted QoS	Guaranteed QoS
MPEG-4 Class I	Unpredictable QoS	Unwarranted QoS	Guaranteed QoS
MPEG-4 Class II	Unpredictable QoS	Unwarranted QoS	Unwarranted QoS
MPEG-4 Class III	Unpredictable QoS	Unwarranted QoS	Unwarranted QoS

Performance of End-to-end Delay and Delay Jitter

The maximum end-to-end delay and delay jitter of video streaming allowed in the simulation are 280 ms and 50 ms, respectively, under the test conditions. The effects of different video marking algorithms on end-to-end delay and delay jitter are presented in Figure 5-38 ~ Figure 5-49.

In the BE solution and the DVMA solution, the end-to-end delay and delay jitter of video traffic in different classes can not be differentiated. That is because different

classes video streams go through the same queue (e.g., BE queue or AF queue). As we expect in the proposed solution, the performance of MPEG-4 Class I stream is better than Class II; and Class II is better than Class III.

In the scenario of BER = 10^{-4}, with a sudden increase of the background traffic at time 36 s, the end-to-end delay and delay jitter of video streams jump sharply up to a higher level. The proposed solution delays 10 s the start point of performance degradation compared with BE and DVMA. After the BSs begin to drop packets, the performance of end-to-end delay and delay jitter turns better. In DVMA, the video delay jitters of all three classes are not acceptable, though Class III and Class II streams in DMVA are better than those in the proposed solution. In comparison, the MPEG-4 Class III and Class II stream in the proposed solution are sacrificed in order to guarantee the QoS of the Class I stream.

The performance tendency is similar in the scenario of BER = 10^{-5}. Note that only MPEG-4 Class III stream in the proposed solution suffers from high delay and delay jitter in order to preserve the Class I and Class II streams. However, in DVMA, all three classes are not acceptable until about time 1m40s. The results in the scenario of BER = 10^{-5} is better than their counterparts in the scenario of BER = 10^{-4}. This is because the numbers of ARQ in BER = 10^{-5} is less than that in BER = 10^{-4}.

In both BER scenarios, the performance of delay and delay jitter in BE seems better than the other two solutions. However, this is at the cost of unacceptable FER in the voice stream and unpredictable quality (e.g., FER) of streaming video service.

5.3.4 Effect of Streaming Video Handoff

As discussed previously, the maximum end-to-end delay and delay jitter of video streaming allowed in both UMTS and MDMN are 280 ms and 50 ms, respectively, under the test conditions. The handoff latency also should keep below 50 ms as a delay jitter.

Figure 5-50 ~ Figure 5-57 show the performance of end-to-end delay and delay jitter in the scenario of BER = 10^{-5}. Figure 5-58 ~ Figure 5-65 illustrate the performance of end-to-end delay and delay jitter in the scenario of BER = 10^{-4}.

In UMTS, only the performance of maximum end-to-end delay in test case iii (Inter-RAN Handoff, BER = 10^{-4}) exceeds 280 ms. Furthermore, in all the UMTS test cases, the

handoff latency can not satisfy the 50 ms delay bound. Note that, because the distance between the Video Provider and GGSN is unknown, we choose the best case in the UMTS simulation model. That is, the distance between them is only one hop. If the Video Provider is far from the UMTS core network, the performance of maximum end-to-end delay in UMTS may not meet the QoS requirement of UMTS any more.

In comparison, with the proposed stream re-establishing handoff solution, all the test cases in MDMN meet the performance requirement of the maximum end-to-end delay and handoff latency. Because the Video Provider is distributed as the media databases inside the core network, the handoff performance keeps stable in MDMN and does not have the distance problem as mentioned above in UMTS.

A brief comparison of handoff performance in UMTS and MDMN is summarized in Table 5-12. This comparison shows that the improvement of handoff performance ascends with the increase of the scale of the mobile core network. For example, the handoff latency improvement in the intra-RAN scale is 26 ms or 45 ms under different conditions of wireless BER, but in the inter-RAN scale it is 38 ms or 57 ms. Furthermore, the proposed stream re-establishing handoff performance in MDMN is relatively consistent in all scenarios. This validates the enhancement of handoff scalability in MDMN.

Table 5-12 Comparison of handoff performance in UMTS and MDMN

Test Case	End-to-end Delay (mean)	Δ Delay	Handoff Latency (mean)	Δ Handoff Latency
I (UMTS, Intra-RAN)	74 ms	32 ms	68 ms	26 ms
II (MDMN, Intra-RAN)	42 ms		42 ms	
III (UMTS, Inter-RAN)	75 ms	30 ms	91 ms	45 ms
IV (MDMN, Inter-RAN)	45 ms		46 ms	
i (UMTS, Intra-RAN)	79 ms	34 ms	82 ms	38 ms
ii (MDMN, Intra-RAN)	45 ms		44 ms	
iii (UMTS, Inter-RAN)	131 ms	75 ms	104 ms	57 ms
iv (MDMN, Inter-RAN)	56 ms		47 ms	

In this chapter, the simulation models, system setup, test conditions, and simulation results are presented and analyzed. Section 5.1 describes the UMTS and the proposed D-MDMN simulation model, and the simulation pipeline stage of the radio transceiver. In Section 5.2, we discuss the simulation and design issues, such as rate shaping and packetization, BER over Rayleigh fading channels, delay-constrained ARQ, followed by the description of traffic profile and test cases. Section 5.2.5 first evaluates the effect of BER on FER over wireless channels and the effect of AF queue size for optimization of system setup. Then performance evaluation of the proposed IP DiffServ video marking algorithm is undertaken to show that it is more suitable for video streaming in IP mobile networks compared with DVMA solution. Finally, the simulation analysis is concluded by the handoff performance comparison of UMTS versus D-MDMN, indicating that the proposed handoff procedures in D-MDMN have better performance in terms of handoff latency, end-to-end delay and handoff scalability than that in UMTS.

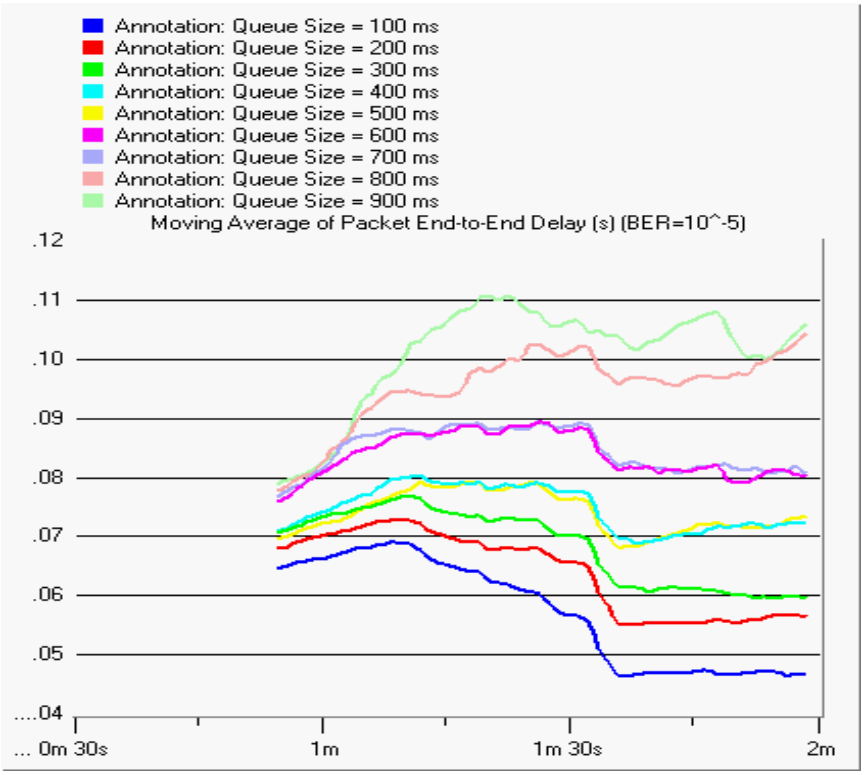


Figure 5-28 Average of End-to-End Delay (Test Case 1~9: BER = 10^{-5})

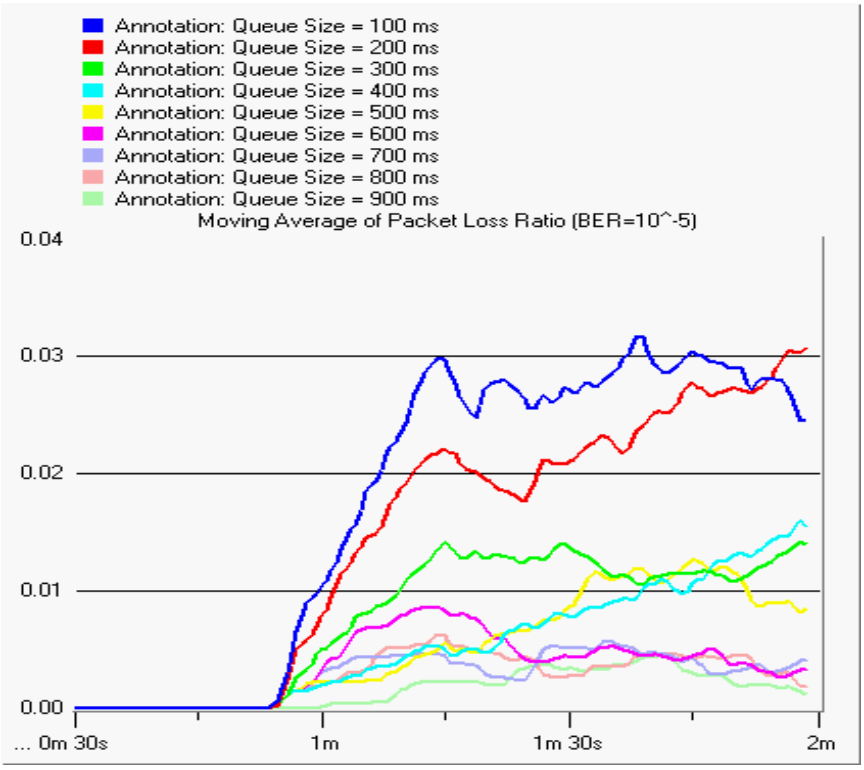


Figure 5-29 Average of Packet Loss Ratio (Test Case 1~9: BER = 10^{-5})

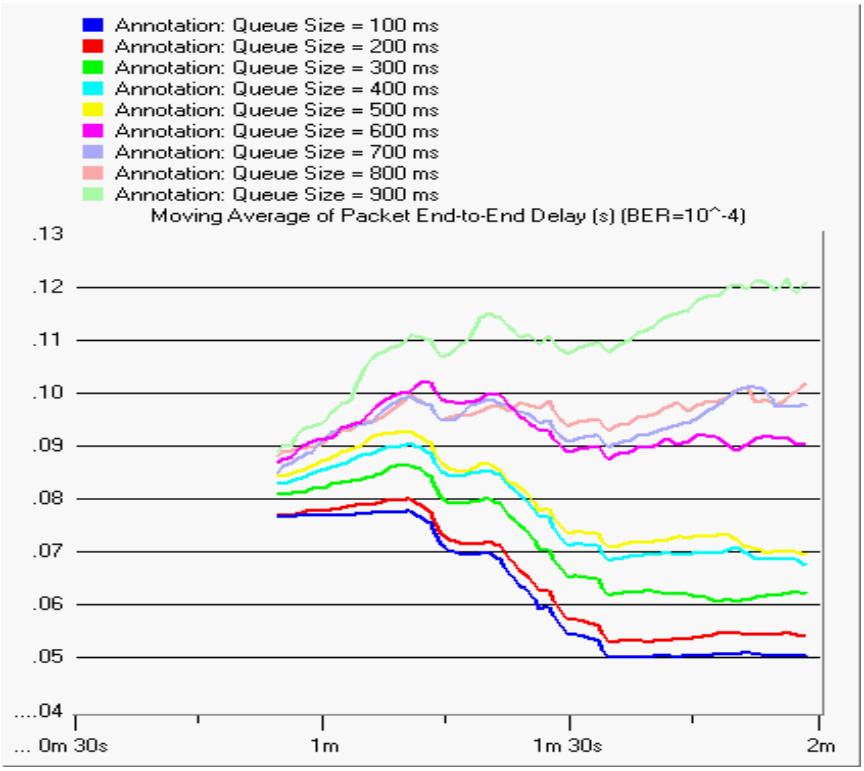


Figure 5-30 Average of End-to-End Delay (Test Case (1)~(9): BER = 10^{-4})

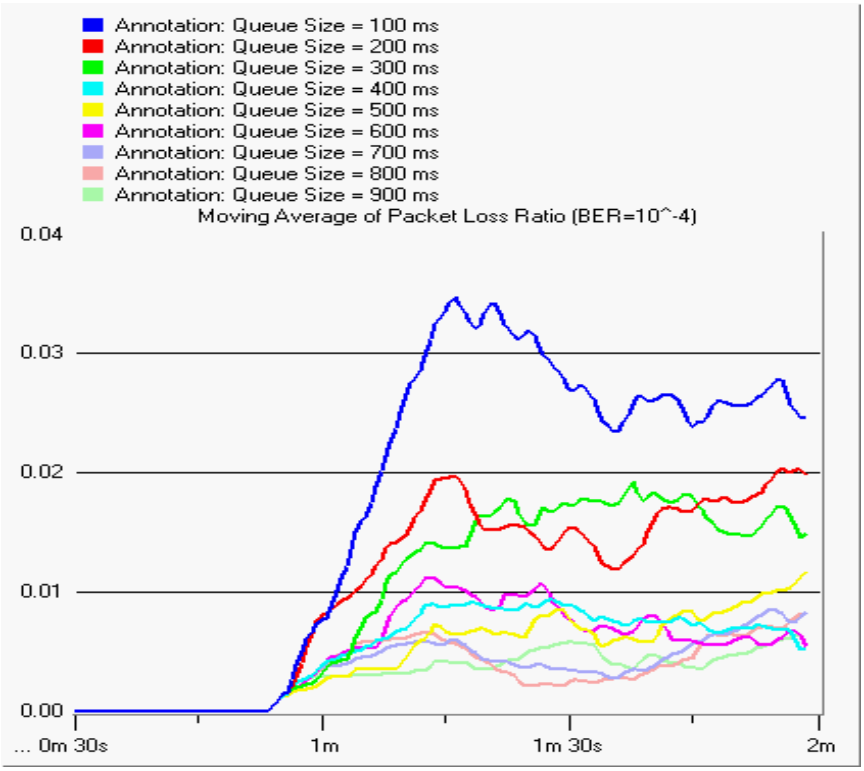


Figure 5-31 Average of Packet Loss Ratio (Test Case (1)~(9): BER = 10^{-4})

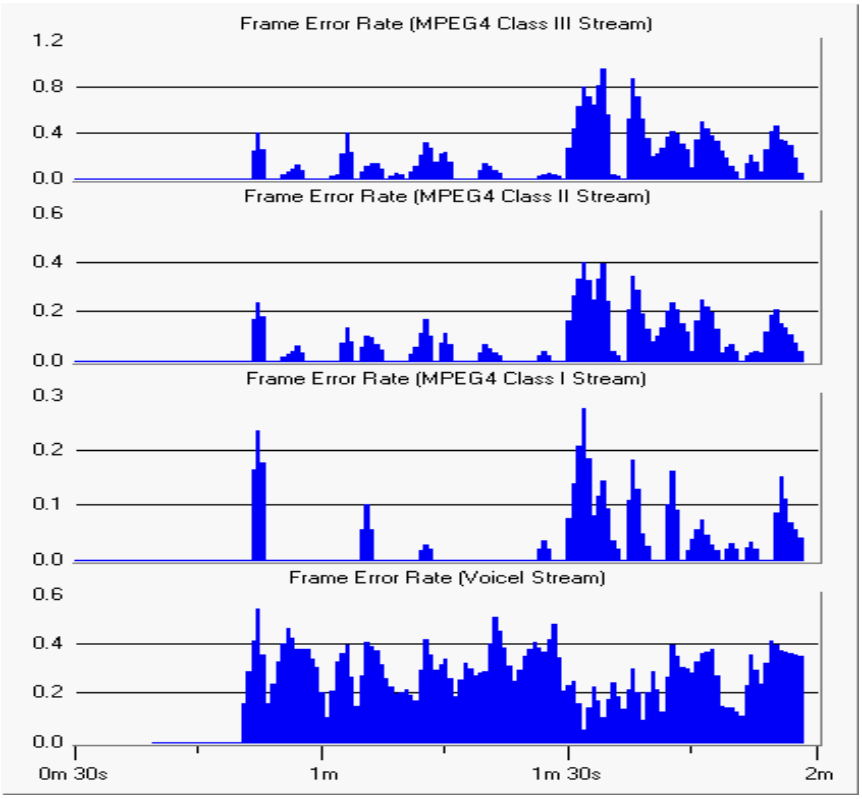


Figure 5-32 FER (Test Case A: Best Effort, FIFO, BER = 10^{-5})

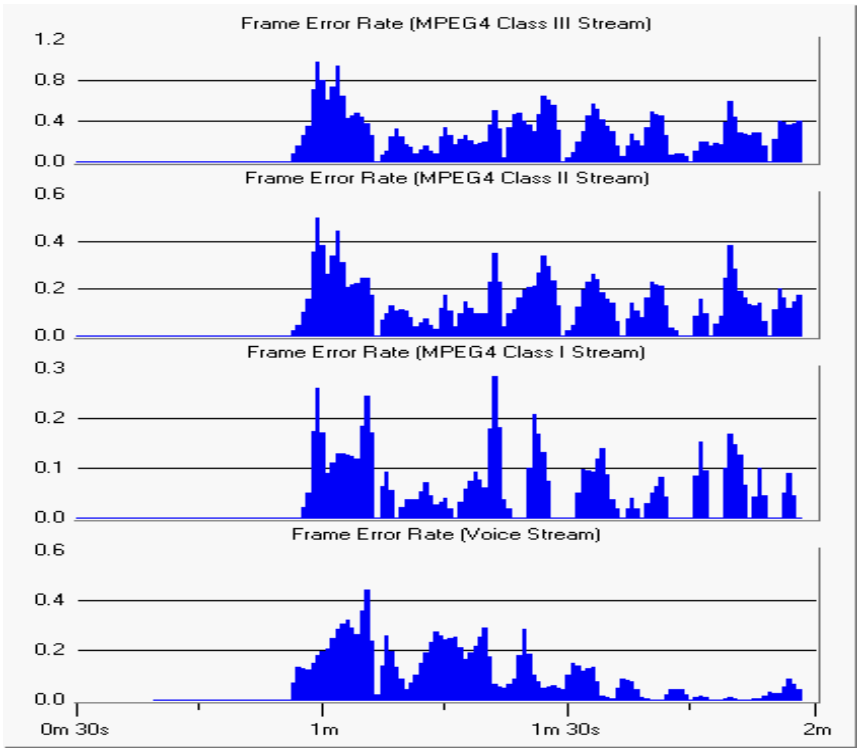


Figure 5-33 FER (Test Case B: DiffServ, WRED, BER = 10^{-5})

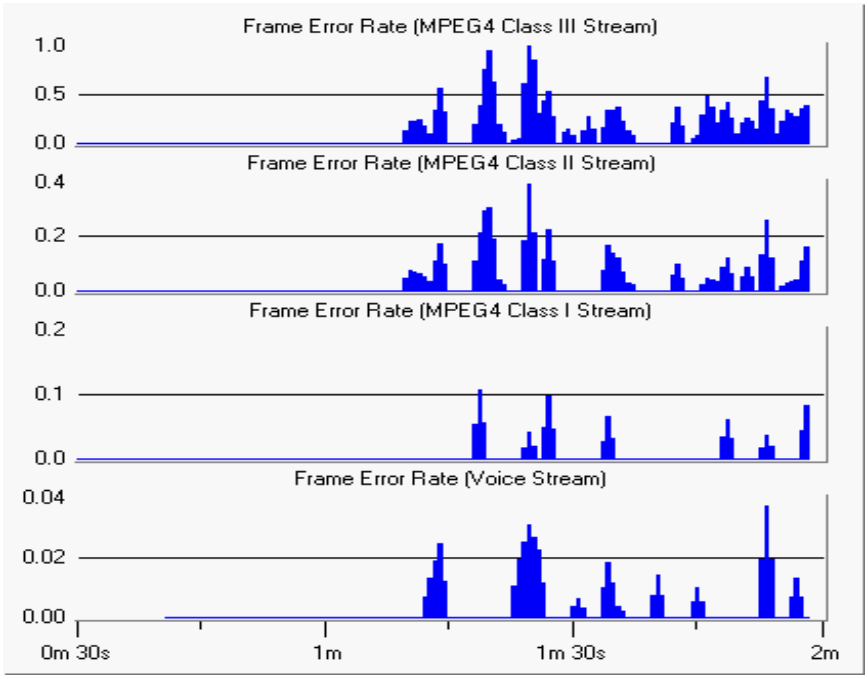


Figure 5-34 FER (Test Case C: DiffServ, WFQ, BER = 10^{-5})

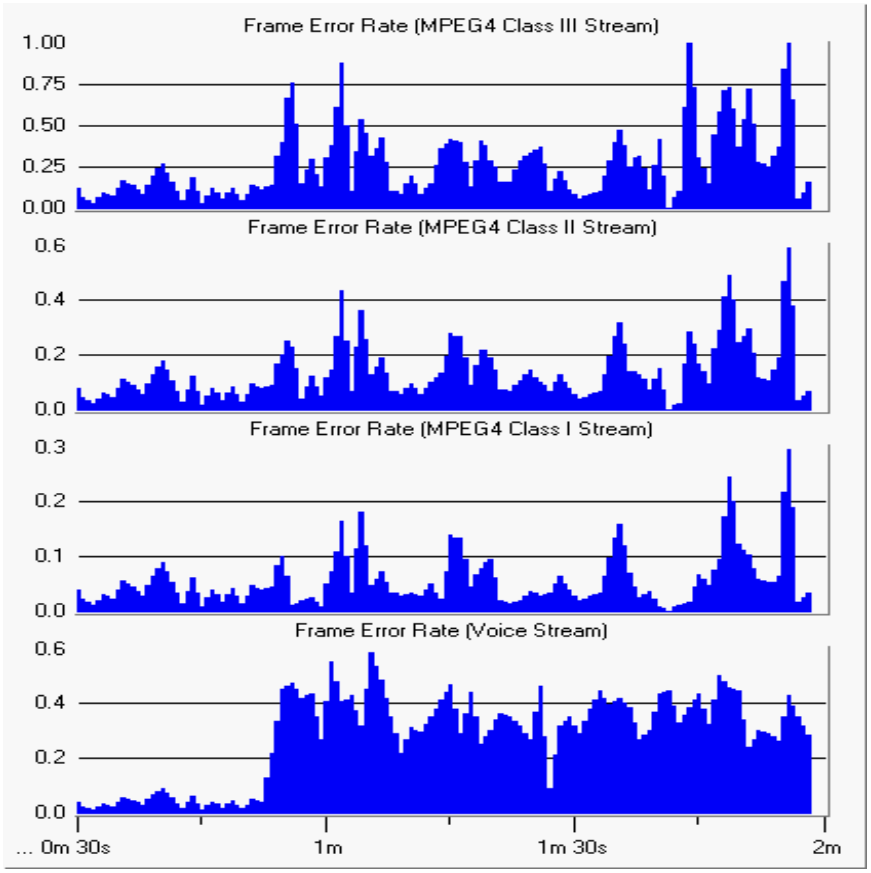


Figure 5-35 FER (Test Case a: Best Effort, FIFO, BER = 10^{-4})

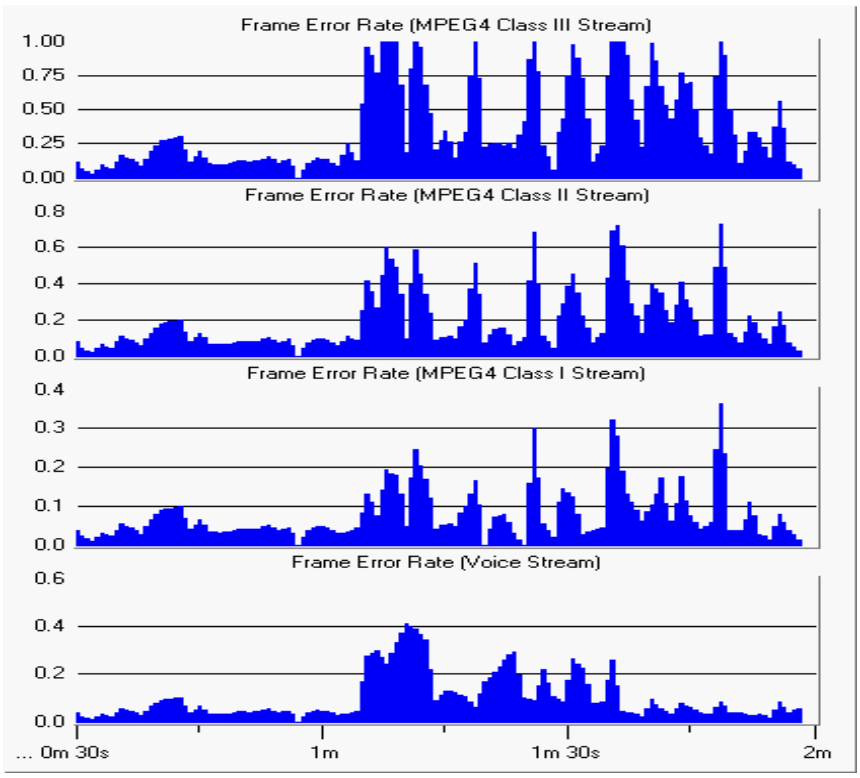


Figure 5-36 FER (Test Case b: DiffServ, WRED, BER = 10^{-4})

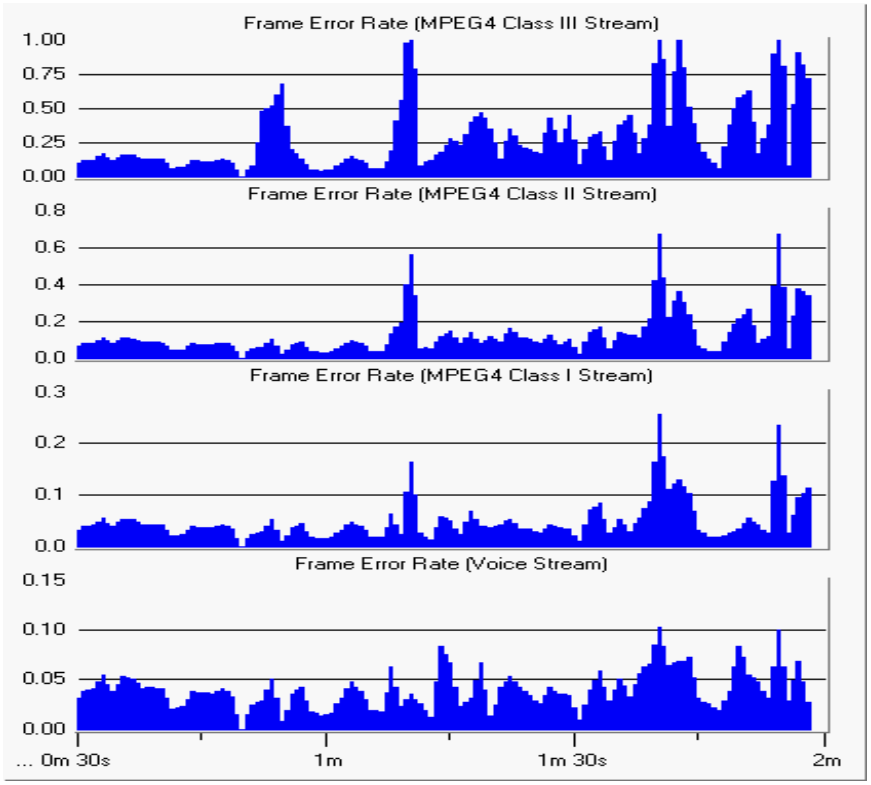


Figure 5-37 FER (Test Case c: DiffServ, WFQ, BER = 10^{-4})

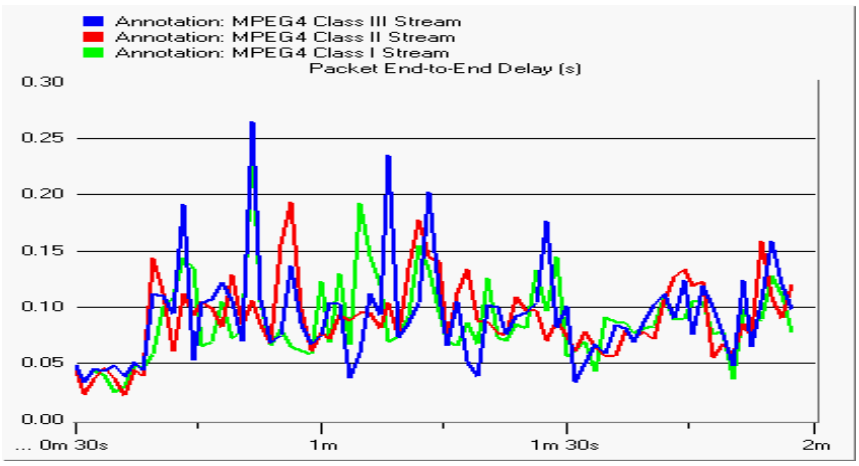


Figure 5-38 End-to-end Delay (Test Case A: Best Effort, FIFO, BER = 10^{-5})

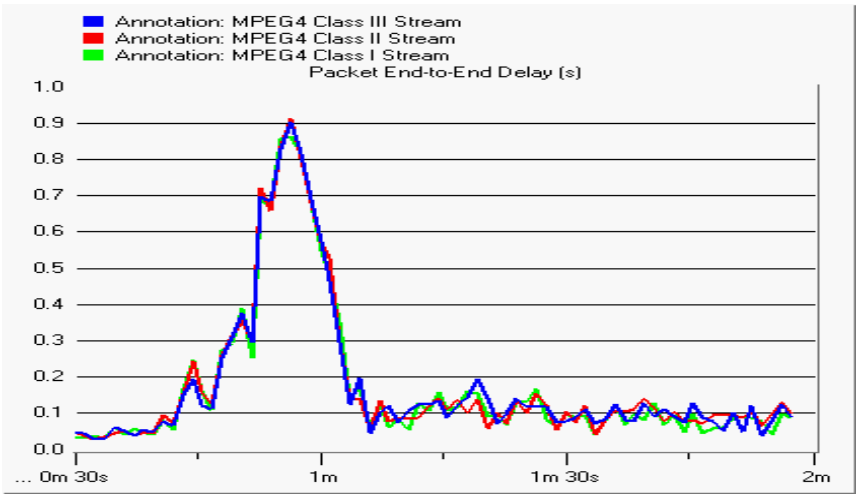


Figure 5-39 End-to-end Delay (Test Case B: DiffServ, WRED, BER = 10^{-5})

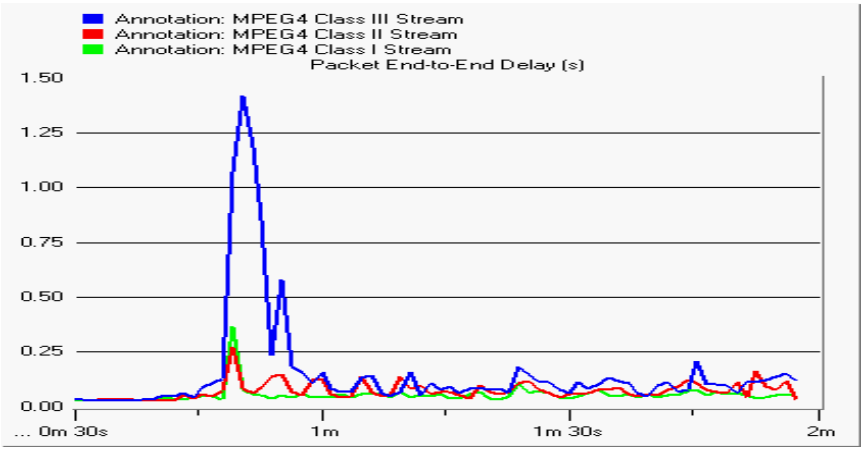


Figure 5-40 End-to-end Delay (Test Case C: DiffServ, WFQ, BER = 10^{-5})

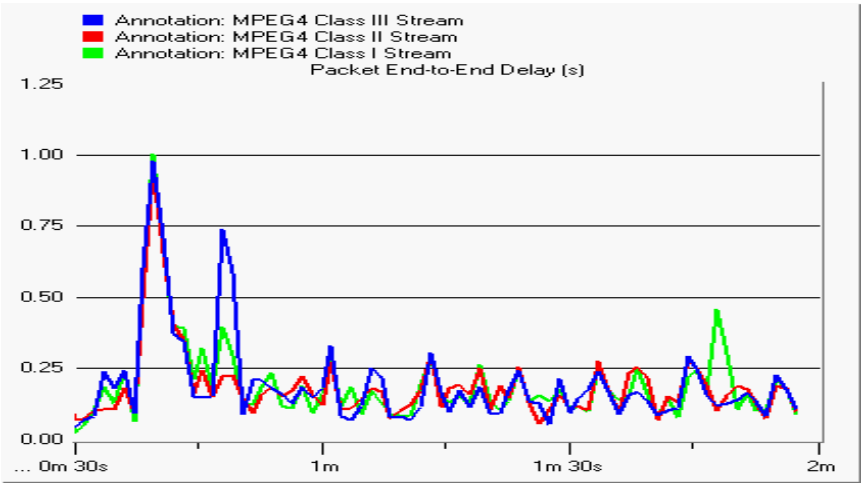


Figure 5-41 End-to-end Delay (Test Case a: Best Effort, FIFO, BER = 10^{-4})

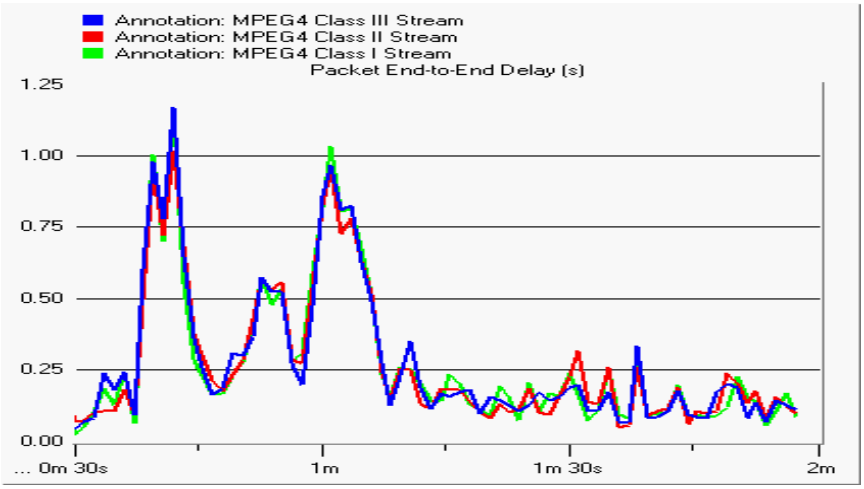


Figure 5-42 End-to-end Delay (Test Case b: DiffServ, WRED, BER = 10^{-4})

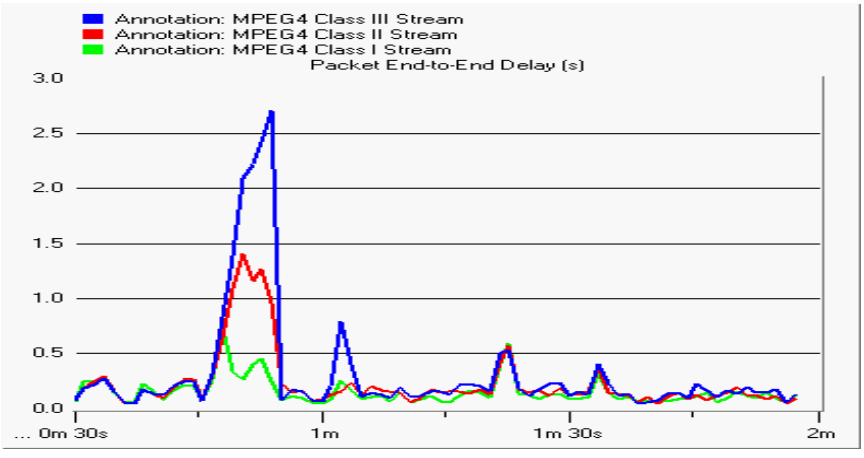


Figure 5-43 End-to-end Delay (Test Case c: DiffServ, WFQ, BER = 10^{-4})

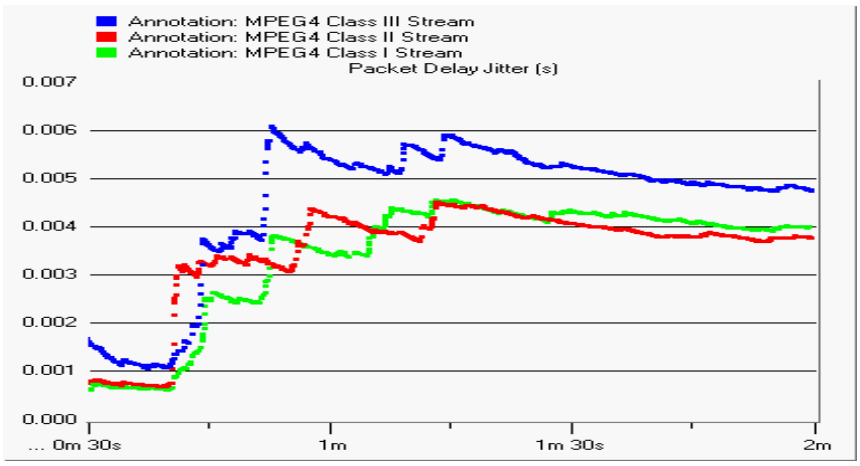


Figure 5-44 Delay Jitter (Test Case A: Best Effort, FIFO, BER = 10^{-5})

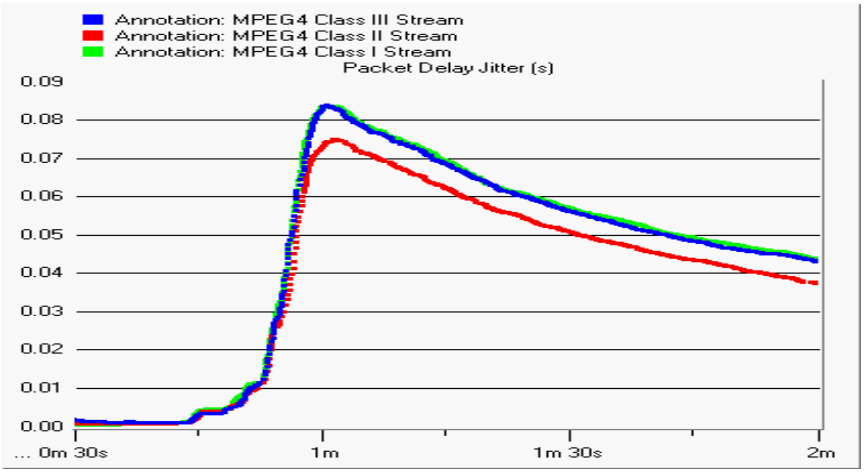


Figure 5-45 Delay Jitter (Test Case B: DiffServ, WRED, BER = 10^{-5})

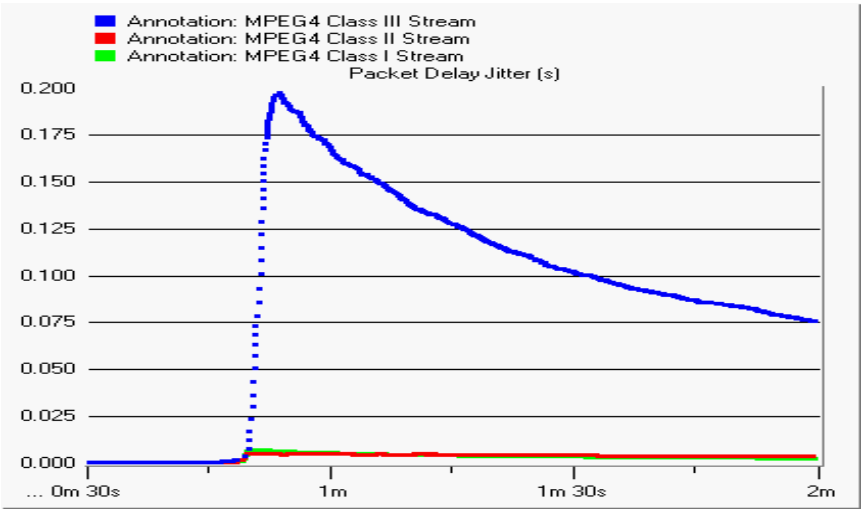


Figure 5-46 Delay Jitter (Test Case C: DiffServ, WFQ, BER = 10^{-5})

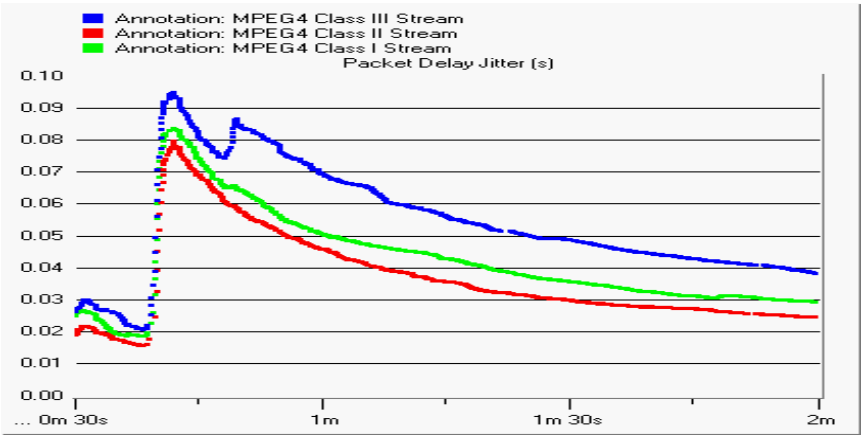


Figure 5-47 Delay Jitter (Test Case a: Best Effort, FIFO, BER = 10^{-4})

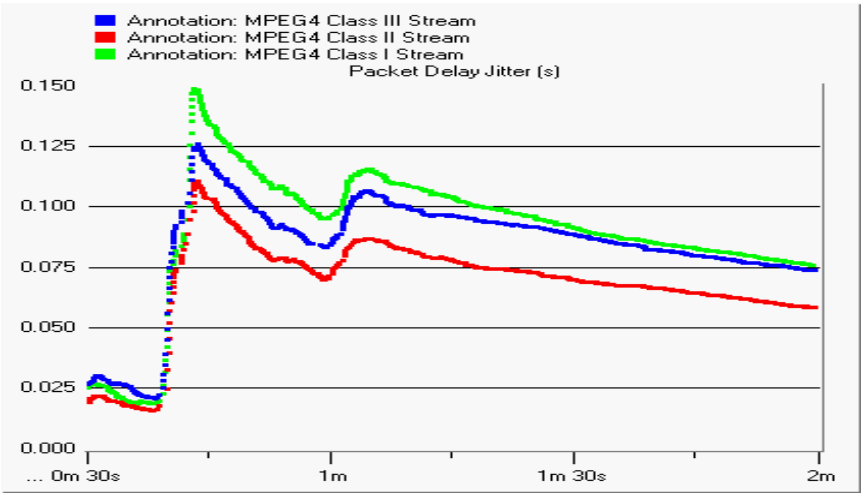


Figure 5-48 Delay Jitter (Test Case b: DiffServ, WRED, BER = 10^{-4})

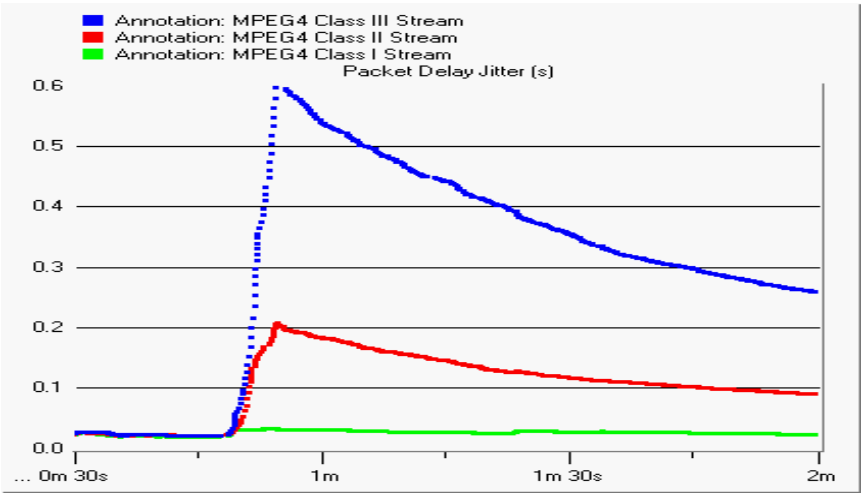


Figure 5-49 Delay Jitter (Test Case c: DiffServ, WFQ, BER = 10^{-4})

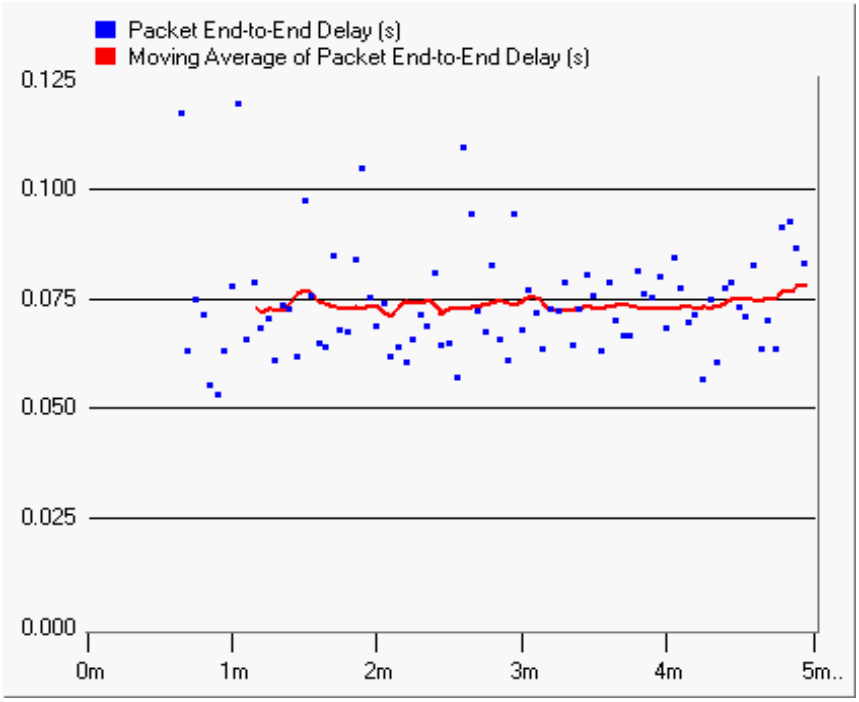


Figure 5-50 End-to-End Delay (Test Case I: UMTS, Intra-RAN, BER = 10^{-5})

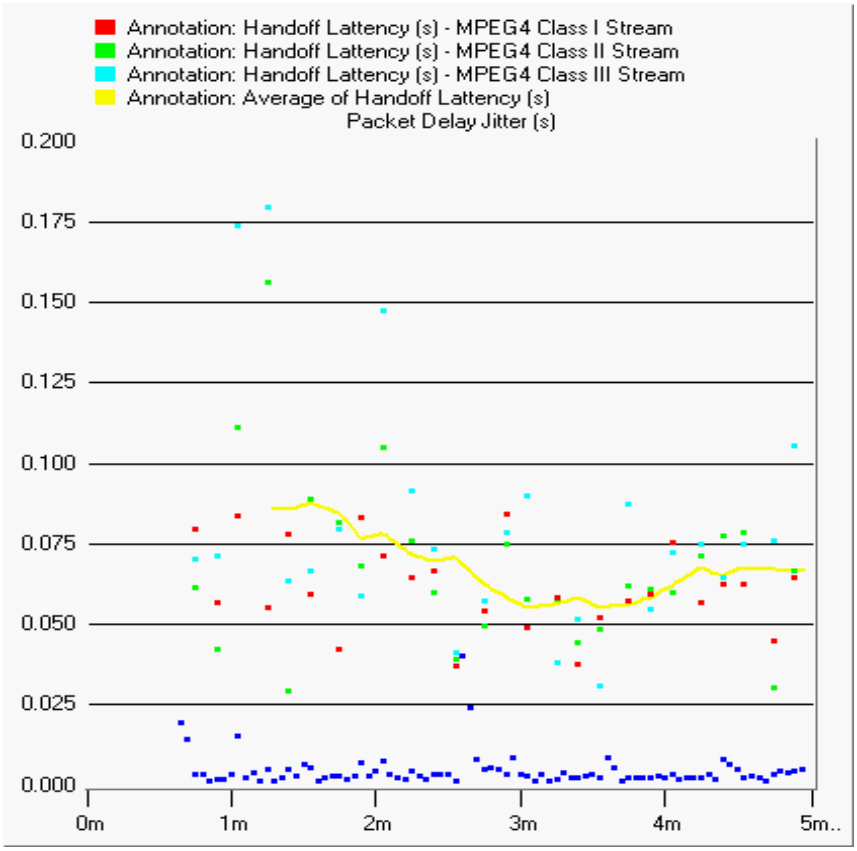


Figure 5-51 Delay Jitter (Test Case I: UMTS, Intra-RAN, BER = 10^{-5})

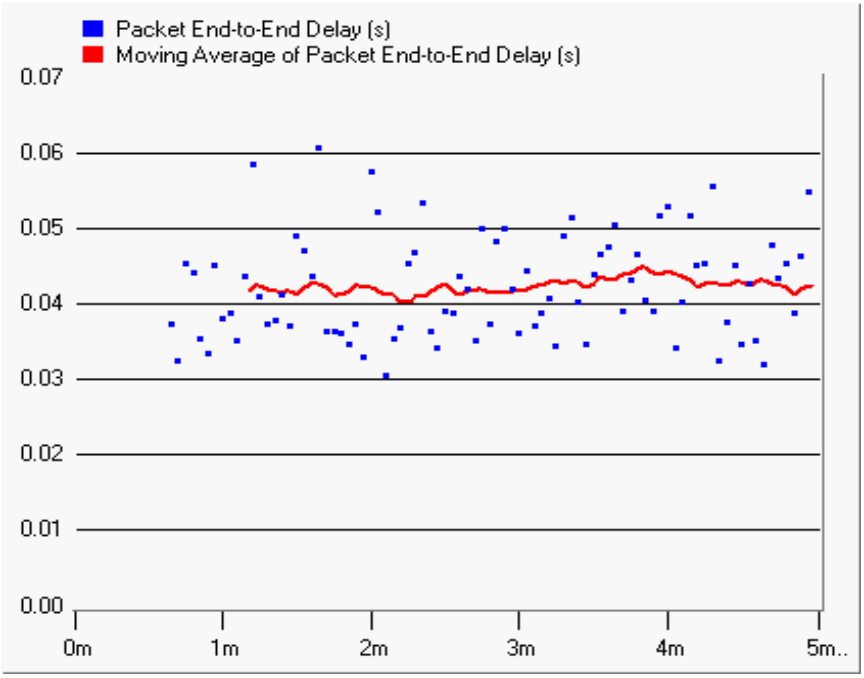


Figure 5-52 End-to-End Delay (Test Case II: MDMN, Intra-RAN, BER = 10^{-5})

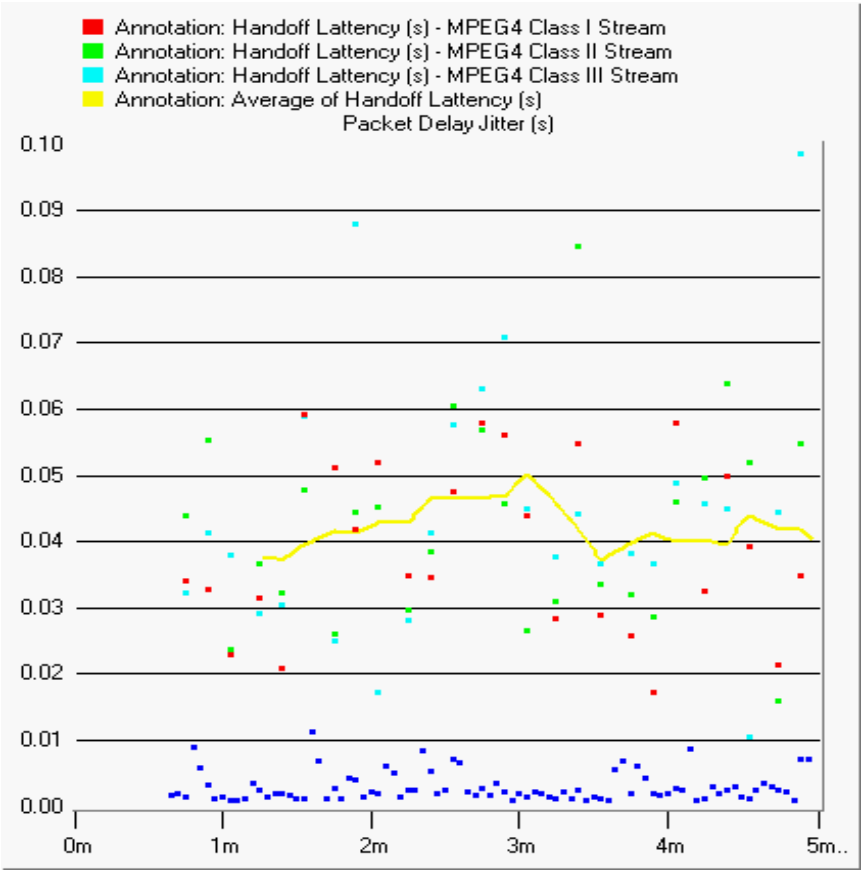


Figure 5-53 Delay Jitter (Test Case II: MDMN, Intra-RAN, BER = 10^{-5})

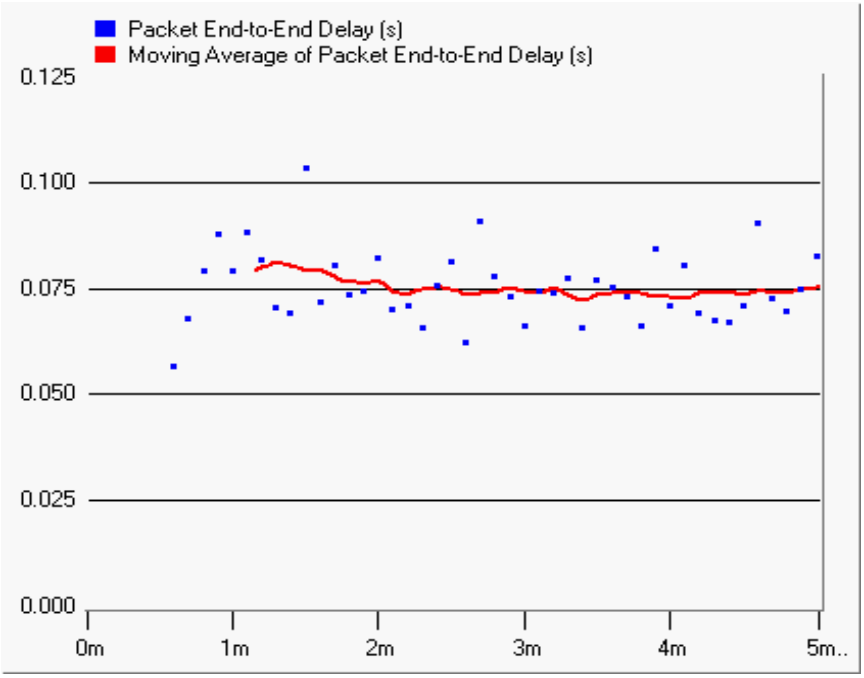


Figure 5-54 End-to-End Delay (Test Case III: UMTS, Inter-RAN, BER = 10^{-5})

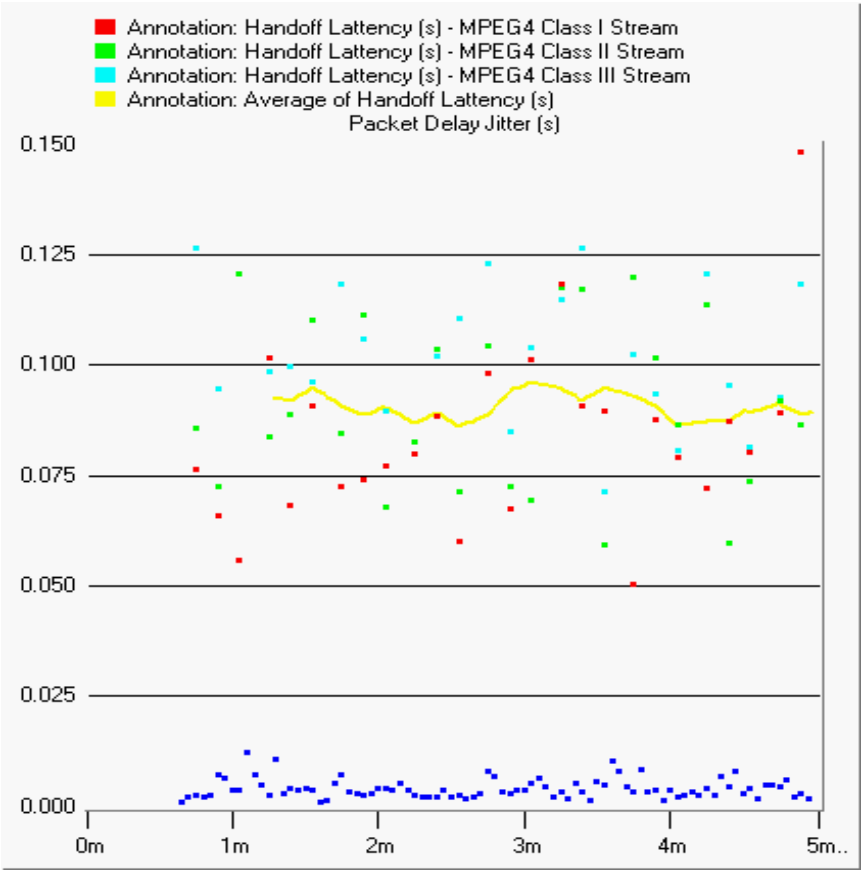


Figure 5-55 Delay Jitter (Test Case III: UMTS, Inter-RAN, BER = 10^{-5})

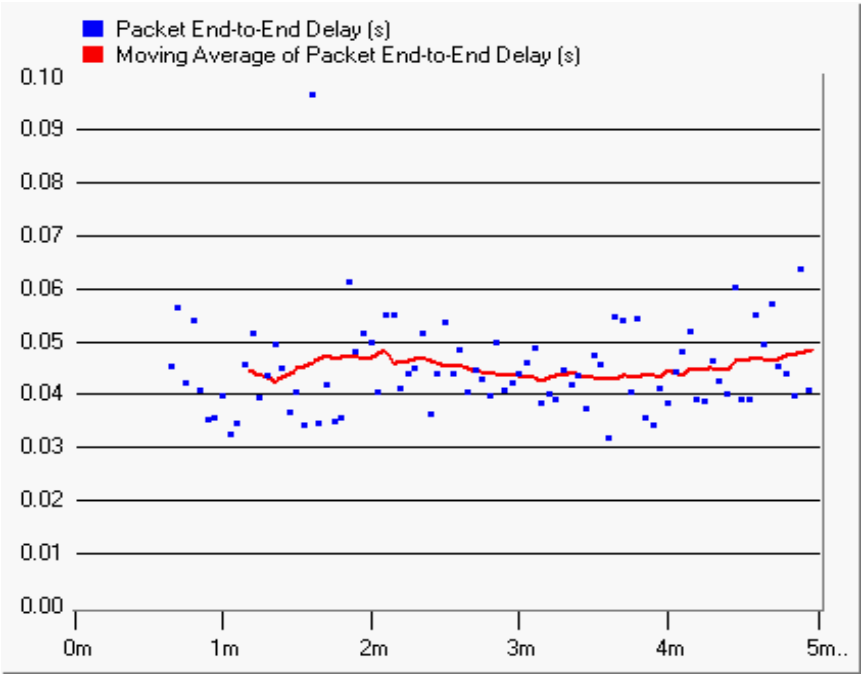


Figure 5-56 End-to-End Delay (Test Case IV: MDMN, Inter-RAN, BER = 10^{-5})

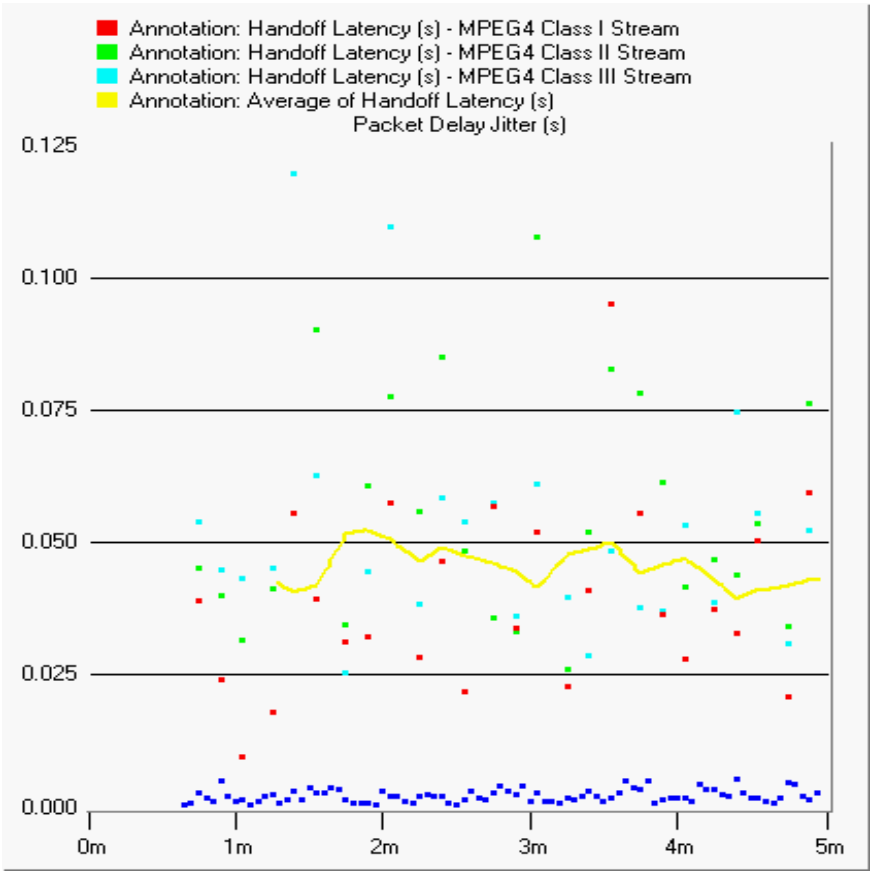


Figure 5-57 Delay Jitter (Test Case IV: MDMN, Inter-RAN, BER = 10^{-5})

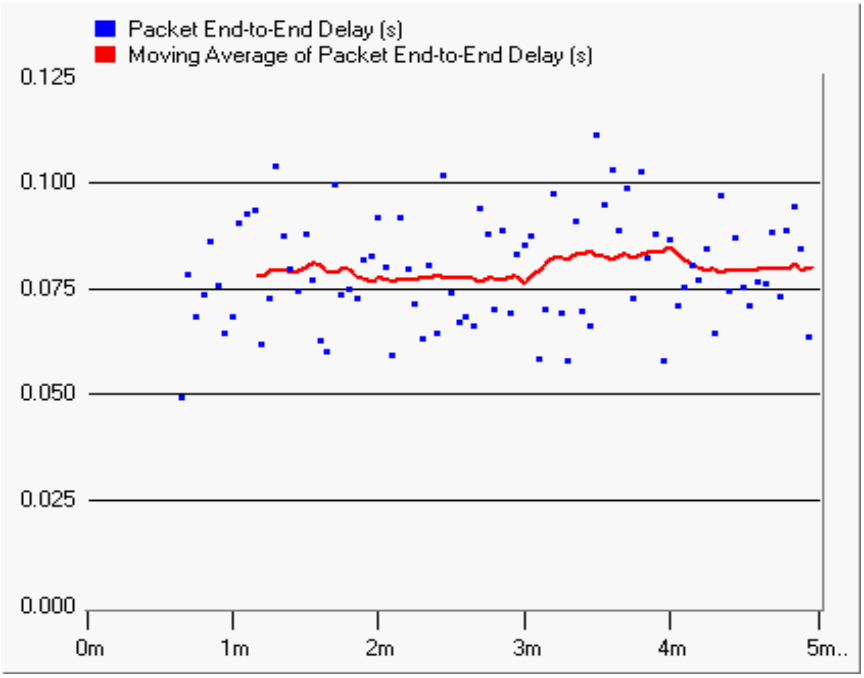


Figure 5-58 End-to-End Delay (Test Case i: UMTS, Intra-RAN, BER = 10^{-4})

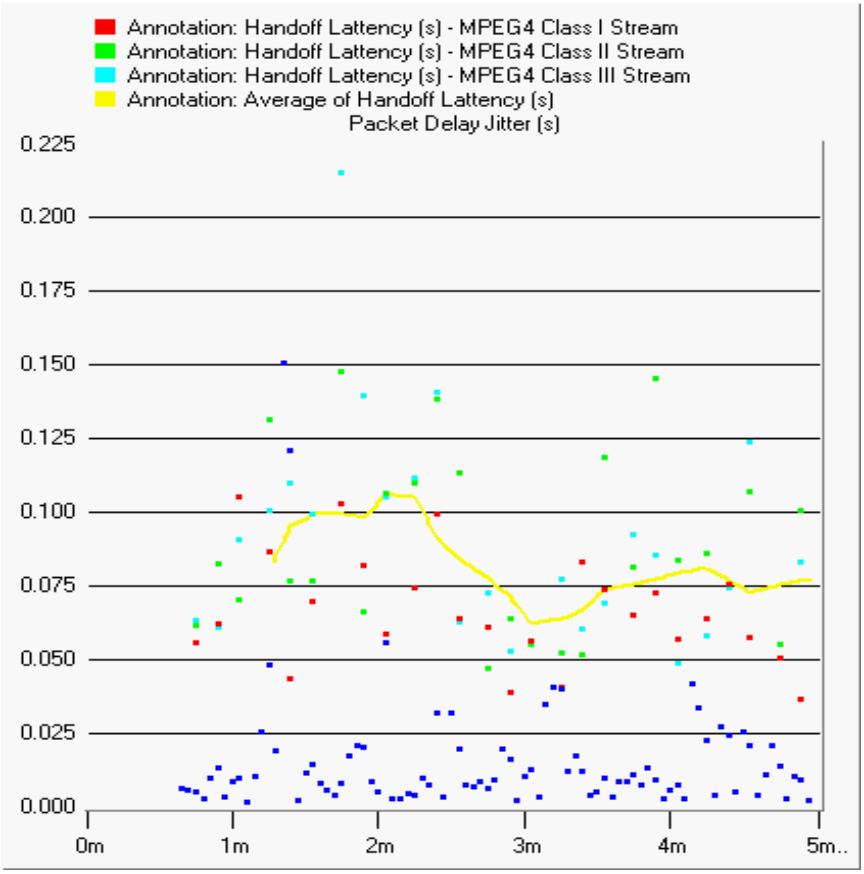


Figure 5-59 Delay Jitter (Test Case i: UMTS, Intra-RAN, BER = 10^{-4})

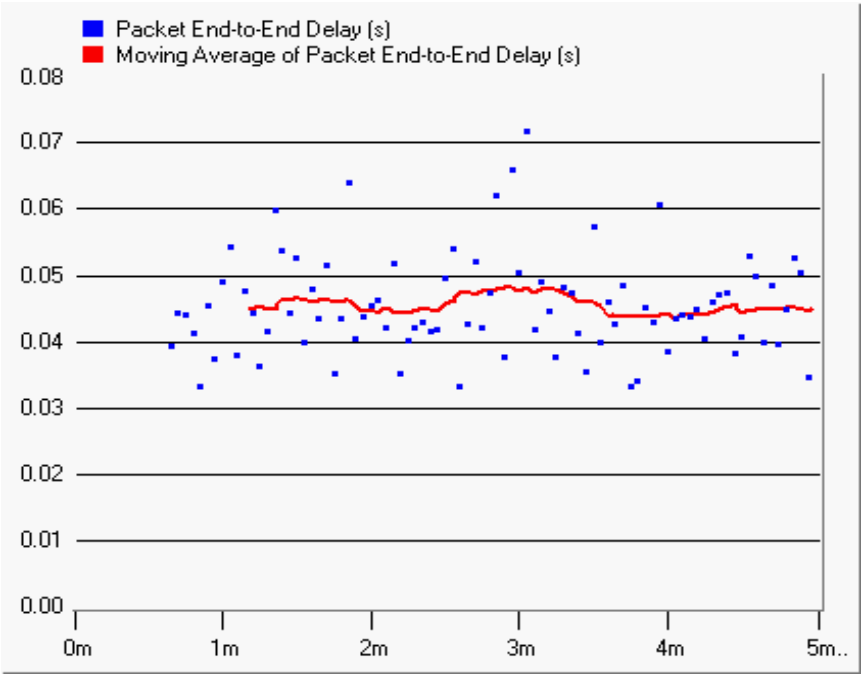


Figure 5-60 End-to-End Delay (Test Case ii: MDMN, Intra-RAN, BER = 10^{-4})

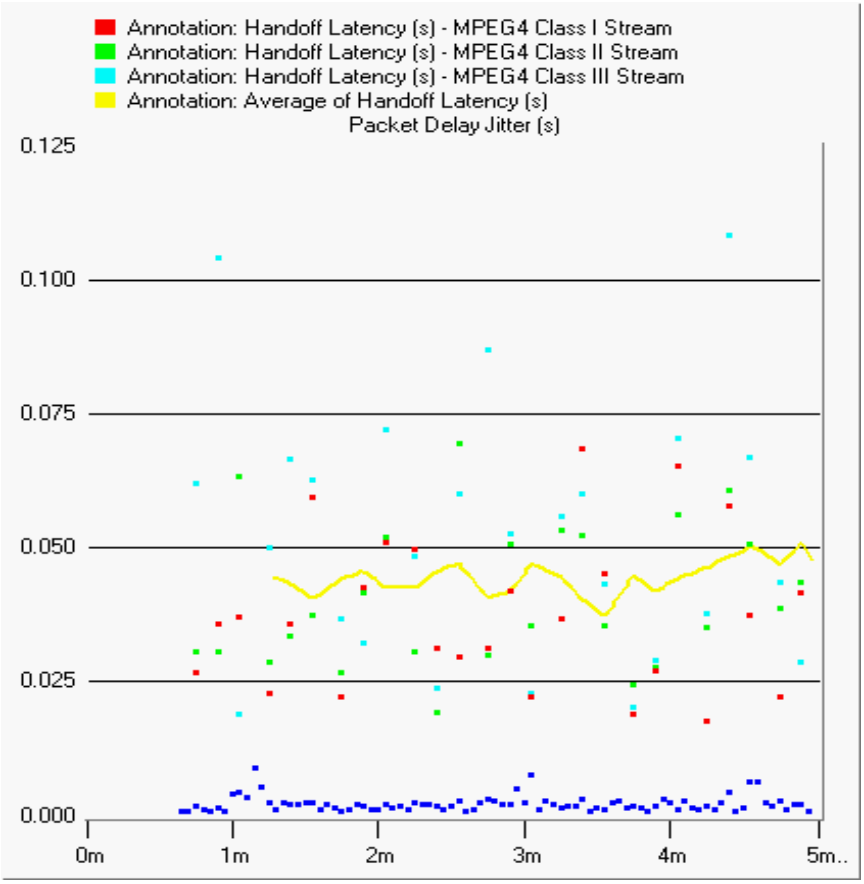


Figure 5-61 Delay Jitter (Test Case ii: MDMN, Intra-RAN, BER = 10^{-4})

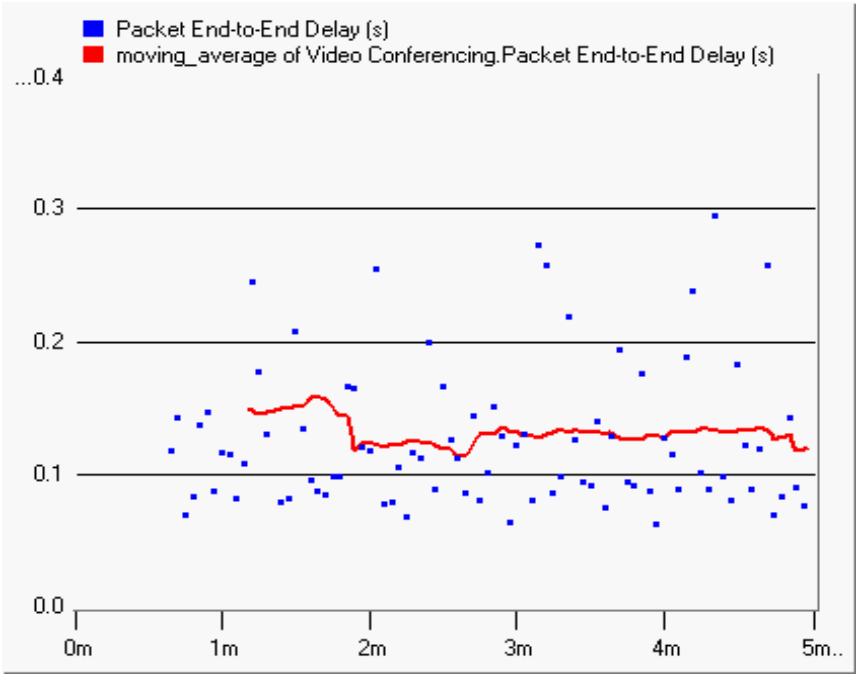


Figure 5-62 End-to-End Delay (Test Case iii: UMTS, Inter-RAN, BER = 10^{-4})

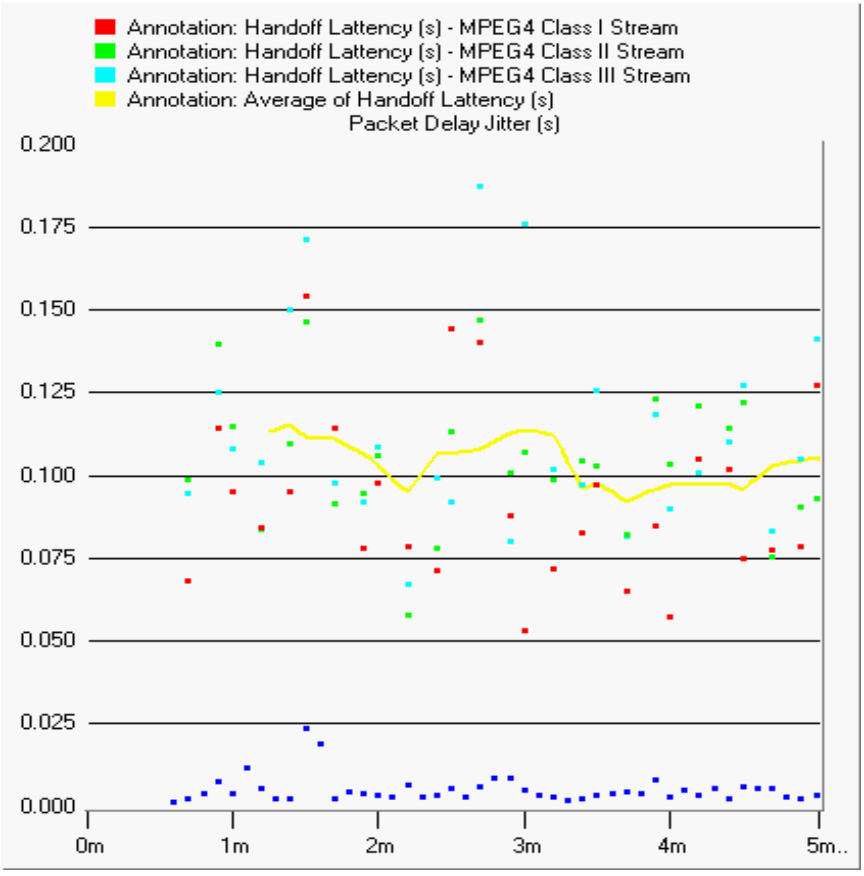


Figure 5-63 Delay Jitter (Test Case iii: UMTS, Inter-RAN, BER = 10^{-4})

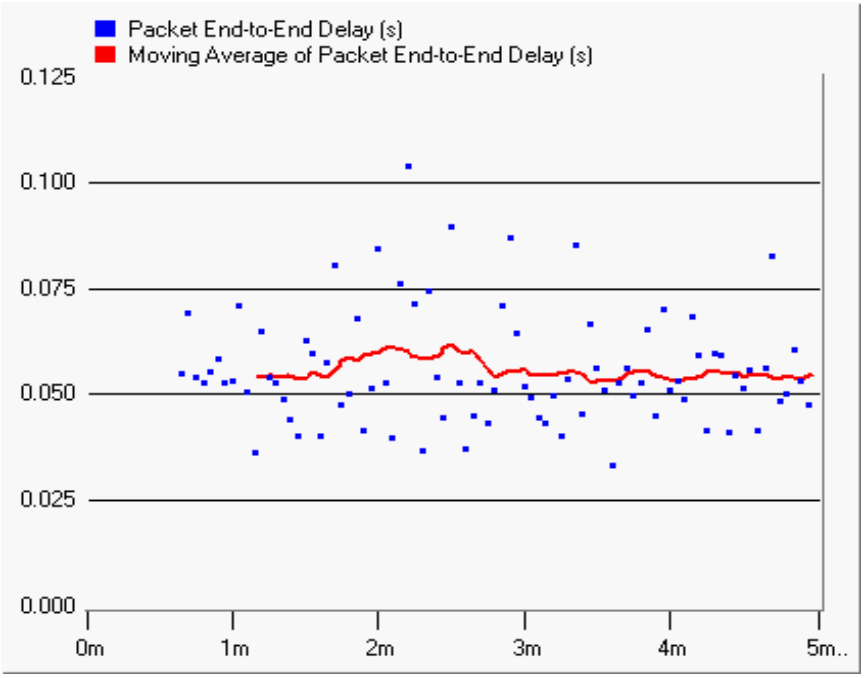


Figure 5-64 End-to-End Delay (Test Case iv: MDMN, Inter-RAN, BER = 10^{-4})

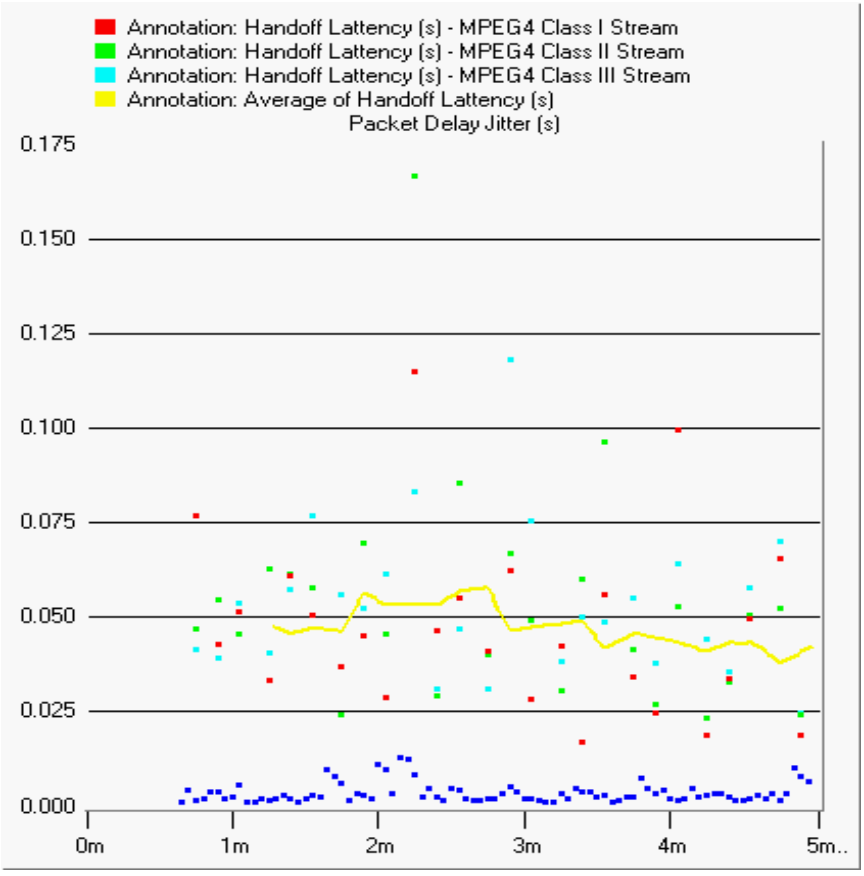


Figure 5-65 Delay Jitter (Test Case iv: MDMN, Inter-RAN, BER = 10^{-4})

6 Conclusions and Future Work

To address the handoff problems in video streaming, as well as the bandwidth fluctuation, packet loss and heterogeneity problems in the wireless networks, and to further enhance the error resilience tools in MPEG-4, the 3G mobile network architecture and the handoff procedures for video delivery in UMTS are studied.

The contributions of this thesis are:

1) A Scalable Multiple Description Coding framework is proposed to explore the joint design of layered coding and multiple description coding.

Under the SMDC framework, MDC components enhance the robustness to losses and bit errors of LC components through path diversity and error recovery. MDC components also reduce the storage, reliability and load balancing requirement among distributed media edge servers. At the same time, LC components not only deal with the unbalanced MD operation at the server end, but also combat the bandwidth frustrations of the time-varying wireless channel. Furthermore, SMDC leverages the distributed multimedia delivery mobile network to provide path diversity to combat video streaming outage due to handoff.

2) A Distributed Multimedia Delivery Mobile Network is proposed for the UMTS core network.

D-MDMN introduces and combines the concepts of CDN and SMDC into the UMTS network in order to solve the video handoff problem and meet the stringent QoS requirements of video streaming in 3GPP. Since the media streaming services are pushed to the edge of core network, it also reduces the media service delivery time, the probability of packet loss, and the total network resource occupation with relatively consistent QoS in all scenarios.

3) Handoff procedures of video streaming in D-MDMN are proposed.

The proposed handoff procedures employ the principle of video stream re-establishing to replace the principle of data forwarding in UMTS. The intra-RAN handoff and inter-RAN handoff procedures are studied in details.

4) A novel IP DiffServ video marking algorithm is proposed to support the unequal error protection of LC components of SMDC.

The proposed algorithm re-organizes the shape, motion, and texture information of video stream into different layers in the proposed SMDC scheme to implement the DiffServ mobile network in UMTS. Furthermore, it spurs the evolution of UMTS toward its final all-IP phase for the purpose of addressing the DiffServ tunneling issue in UMTS.

The above proposed schemes have been validated through the simulation presented in Chapter 5, except that the verification of MDC components of SMDC can not be undertaken because of technical complexity and time limitation.

The limitations and cost of the proposed schemes also can be summarized as follows.

1) The significant performance improvement of SMDC is achieved at the cost of a coding overhead.

2) The D-MDMN network solution is feasible as a client-server solution for video streaming delivery service, but not for end-to-end real-time video conversation.

3) For video streaming delivery, the video descriptions should be distributed in advance into all complementary MDSs at the edge of the RANs according to the service subscription of video mobile users, which adds the complexity to UMTS. A potential video adaptive deployment solution can be given as follow.

Firstly, during the CAC, the video service subscribed by an MS are distributed to the MDSs in the current local RAN and all its neighbor RANs. Secondly, if the MS moves to another RAN, this video service should be simultaneously distributed to the MDSs in all its new neighbor RANs after the handoff successfully takes place.

Further work will include the verification of the proposed SMDC under the object-based MPEG-4 video stream, especially the MDC components over the Rayleigh fading channel, the study of the effect of SMDC coding overhead and the soft handoff procedures in D-MDMN. In addition, the proposed solutions of video streaming may be applicable to audio streaming. The joint design of audio LC and MDC, the distributed audio delivery network, the IP DiffServ audio marking algorithm, and the media synchronization between video stream and audio stream should be studied further.

Appendix

Proposed inter-cell, intra-RNS handoff procedure

In the scenario of inter-cell, intra-RNS handoff, the proposed handoff procedures, shown in Figure 0-1, consists of three phase:

- **Phase I: Preparation of BTS handoff and resource allocation**
- **Phase II: Moving the Serving BTS role to target BTS**
- **Phase III: Releasing resource reservation in the old path**

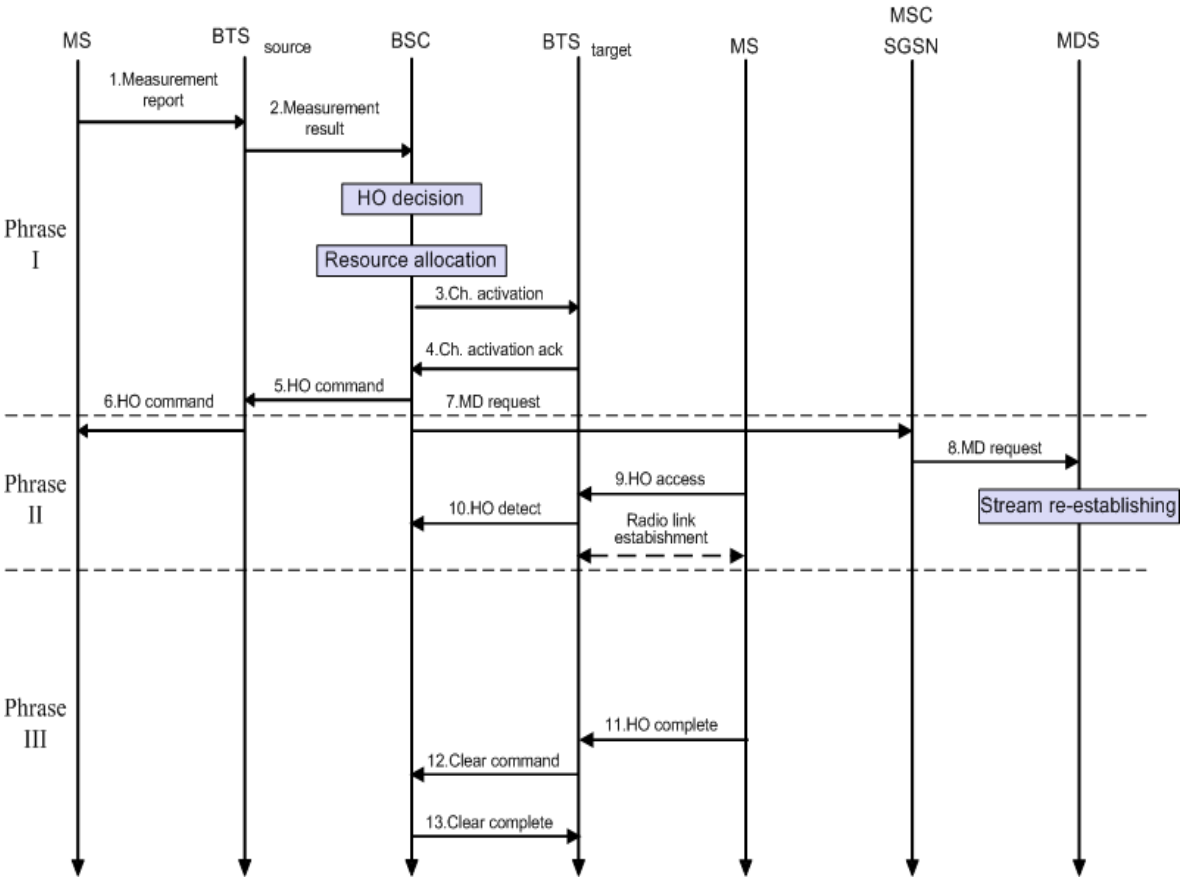


Figure 0-1 Proposed inter-cell, intra-RNS handoff procedure (Control plane)

Acronyms

3GPP	Third Generation Partnership Project
ABR	Available Bit Rate
AF	Assured Forwarding
ALP	Application Level Packets
AMR	Adaptive Multi Rate
AR	Access Router
ARQ	Automatic Repeat reQuest
ATM	Asynchronous Transfer Mode
BE	Best Effort
BER	Bit Error Rate
B-frame	Bi-directionally predicted frame
BIFS	Binary Format for Scene
BPSK	Binary Phase Shift Keying
BR	Border Router
BS	Base Station
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CAC	Call Admission Control
CBQ	Class Based Queuing
CBR	Constant Bit Rate
CCIR	Consultative Committee for International Radiocommunication
CDMA	Code Division Multiple Access
CDN	Content Delivery Network
CIF	Common Interleaved Frame
CLR	Cell Loss Rate
CN	Core Network

CoS	Class of Service
CS	Circuit Switched
DC	Direct Current
DCT	Discrete Cosine Transform
DiffServ	Differentiated Services
DL	Down Link
D-MDMN	Distributed Multimedia Delivery Mobile Network
DS	DiffServ
DSCP	DiffServ CodePoint
DVMA	DiffServ video marking algorithm
EDGE	Enhanced Data rates for GSM Evolution
EF	Expedited Forwarding
ER	Edge Router
FDCT	Forward-DCT
FDD	Frequency Division Duplex
FEC	Forward Error Correction
FER	Frame Erasure Rates
FGS	Fine granularity scalability
FP	Frame relay transport Protocol
GERAN	GSM/EDGE radio access network
GFR	Guaranteed Frame Rate
GGSN	Gateway GPRS Support Node
GOB	Group Of Blocks
GOP	Group Of Pictures
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
GTP	GPRS Tunneling Protocol
HLR	Home Location Register
IDCT	Inverse-DCT
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force

I-frame	Intra-coded frame
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
I-VOP	Intra-coded Video Object Plane
LAN	Local Area Network
LC	Layered Coding
MC-CMDA	Multiple Carrier CDMA
MD	Multiple Description
MDC	Multiple Description Coding
MDS	Media Description Server
MPEG	Moving Pictures Experts Group
MS	Mobile Station
MSC	Mobile Switching Center
nrt-VBR	non real-time Variable Bit Rate
NSS	Network Subsystem
OD	Object Descriptor
OFDM	Orthogonal Frequency Division Multiplexing
PDU	Protocol Data Unit
PFGS	Progressive Fine Granularity Scalability
PFGST	PFGS Temporal
P-frame	Predictively coded frame
PHB	Per-Hop Behavior
PS	Packet Switched
PSS	Packet-switched Streaming Service
PSTN	Public Switched Telephone Network
P-VOP	Predictively coded Video Object Plane
QCIF	Quarter Common Interleaved Frame
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAN	Radio Access Network

RED	Random Early Detection
RFC	Request For Comments
RFL	Radio Frequency Layer
RNC	Radio Network Controller
RNL	Radio Network Layer
RNS	Radio Network Subsystems
RPS	Reference Picture Selection
RS	Redirection Server
RTCP	Real Time Control Protocol
RTP	Real Time Protocol
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
SD	Single Description
SDP	Session Description Protocol
SDU	Service Data Unit
SGSN	Serving GPRS Support Node
SIP	Initiation Protocol
SLA	Service Level Agreement
SMDC	Scalable Multiple Description Coding
SNR	Signal-to-Noise Ratio
sRNS	source RNS
sSGSN	source SGSN
TCP	Transmission Control Protocol
TD-CDMA	Time Division-CDMA
TDD	Time Division Duplex
tRNS	target RNS
tSGSN	target SGSN
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UEP	Unequal Error Protection

UL	Up Link
UMTS	Universal Mobile Telecommunications System
URL	Uniform Resource Locator
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
VO	Video Object
VoIP	Voice over IP
VOP	Video Object Planes
VP	Video Packet
WAN	Wide Area Network
WCDMA	Wideband-CDMA
WFQ	Weighted Fair Queuing
WRED	Weighted Random Early Detection

References

[1] Dapeng Wu, Y. T. Hou, Wenwu Zhu, Ya-Qin Zhang, Jon M. Peha, Streaming "Video over the Internet: Approaches and Directions", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11 No. 3 , pp. 282 -300, Mar. 2001

[2] Dapeng Wu, Y. T. Hou, Ya-Qin Zhang, "Transporting Real-Time Video over the Internet: Challenges and Approaches", *Proc. of the IEEE* , Vol. 88, No. 12 , pp. 1855 -1877, Dec. 2000

[3] Frank G. Lin, "Transparent MPEG Video Filtering for Wireless Networks", Master's thesis, University of Waterloo, Canada, 2000.

[4] S. Floyd, and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", *IEEE/ACM Trans. on Networking*, V.1 N.4, pp. 397-413,Aug. 1993.

[5] G. Blakowski, R. Steinmetz, "A Media Synchronization Survey: Reference Model, Specification, and Case Studies", *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 1, pp. 5 -35, Jan. 1996

[6] H.M. Radha, M. van der Schaar, Yingwei Chen, "The MPEG-4 Fine-grained Scalable Video Coding Method for Multimedia Streaming over IP", *IEEE Trans. on Multimedia*, Vol. 3, No. 1 , pp. 53 -68, Mar. 2001

[7] Weiping Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11, No. 3 , pp. 301 -317 , Mar. 2001

[8] Weiping Li, Fan Ling, Xuemin Chen, "Fine Granularity Scalability in MPEG-4 for Streaming Video", *Proc. IEEE Int. Symp. on Circuits and Systems*, Vol. 1, pp. 299 -302 , 2000

[9] "Transparent End-to-end Packet Switched Streaming Service (PSS); Protocols and codecs", 3GPP TS 26.234 V5.2.0, Sept. 2002

[10] Sumit Roy, Bo Shen, Vijay Sundaram, "Application Level Hand-off Support for Mobile Media Transcoding Sessions", *Pro. of the 12th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video*, May 2002

[11] Roger Karrer, Thomas Gross, "Dynamic Handoff of Multimedia Streams", *Proc. of the 11th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video*, Jan. 2001

[12] J. Kangasharju, F. Hartanto, M. Reisslein, K.W. Ross, "Distributing Layered Encoded Video through Caches", *IEEE Trans. on Computers*, Vol. 51, No. 6, pp. 622 -636, June 2002

[13] A.K. Katsaggelos, L.P. Kondi, F.W. Meier, J. Ostermann, G.M. Schuster, "MPEG-4 and Rate-distortion-based Shape-coding Techniques", *Proc. of the IEEE* , Vol. 86, No. 6, pp. 1126 -1154, June 1998

[14] H. Schulzrinne, A. Rao, R. Lanphier, "Real Time Streaming Protocol (RTSP) ", IETF RFC 2326, Apr. 1998.

[15] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", IETF RFC 1889, Jan. 1996.

[16] Jiangchuan Liu, Huai-Rong Shao, Bo Li, Wenwu Zhu, Ya-Qin Zhang, "A Scalable Bit-stream Packetization and Adaptive Rate Control Framework for Object-based Video Multicast", *Proc. of IEEE Global Telecommunications Conf.*, Vol. 3, pp. 2020 -2025, 2001

[17] Feng Wu, Shipeng Li, Ya-Qin Zhang, "A Framework for Efficient Progressive Fine Granularity Scalable Video Coding", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11, No. 3, pp. 332 -344, Mar. 2001

[18] S. Wenger, G.D. Knorr, J. Ott, F. Kossentini, "Error Resilience Support in H.263+", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8, No. 7, pp. 867 -877, Nov. 1998

[19] Dapeng Wu, Y.T. Hou, Ya-Qin Zhang, "Scalable Video Transport over Wireless IP Networks", *Proc. of IEEE Int. Symp. on Indoor and Mobile Radio Communications*, Vol. 2, pp. 1185 -1191, 2000

[20] R. Aravind, M.R. Civanlar, A.R. Reibman, "Packet Loss Resilience of MPEG-2 Scalable Video Coding Algorithms", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 6, No. 5, pp. 426 -435, Oct. 1996

[21] A. Ortego, K. Ramchandran, "Rate-distortion Methods for Image and Video Compression", *IEEE Signal Processing Magazine*, Vol. 15, No. 6, pp. 23 -50, Nov. 1998

[22] J. Padhye, V. Firoiu, D.F. Towsley, J.F. Kurose, "Modeling TCP Reno Performance: A Simple Model and its Empirical Validation", *IEEE Trans. on Networking*, Vol. 8, No. 2, pp. 133 -145, Apr. 2000

[23] Qian Zhang, Wenwu Zhu, Ya-Qin Zhang, "Network-adaptive Scalable Video Streaming over 3G Wireless Network", *Proc. of IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 579 -582, 2001

[24] S. Floyd, K.Fall, "Promoting the Use of End-to-end Congestion Control in the Internet", *IEEE Trans. on Networking*, Vol. 7, No. 4, pp. 458 -472, Aug. 1999

[25] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, "On the Self-similar Nature of Ethernet Traffic", *IEEE Trans. on Networking*, Vol. 2, No. 1, pp. 1 -15, Feb. 1994

[26] John G. Apostolopoulos, Susie J. Wee, "Unbalanced Multiple Description Video Communication Using Path Diversity", *Proc. of Int. Conf. on Image Processing*, Oct. 2001

[27] M. Gallant, F. Kossentini, "Rate-distortion Optimized Layered Coding With Unequal Error Protection for Robust Internet Video", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11, No. 3, pp. 357 -372, Mar. 2001

[28] Yao Wang, L. Karam, G.P. Abousleman, T. Key, B. Razzouk, "Error Resilient Video Coding Techniques", *IEEE Signal Processing Magazine*, Vol. 17, No. 4, pp. 61 -82, July 2000

[29] I. Moccagatta, S. Soudagar, J. Liang, H. Chen, "Error-resilient Coding in JPEG-2000 and MPEG-4", *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 6, pp. 899 -914, June 2000

[30] R. Talluri, "Error-resilient Video Coding in the ISO MPEG-4 Standard", *IEEE Communications Magazine*, Vol. 36, No. 6, pp. 112 -119, June 1998

[31] G. Cote, B. Erol, M. Gallant, F. Kossentini, "H.263+: Video Coding at Low Bit Rates", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8, No. 7, pp. 849 -866, Nov. 1998

[32] A.S. Tosun, W.C. Feng, "Efficient Multi-layer Coding and Encryption MPEG Video Stream", *Proc. of IEEE Int. Conf. on Multimedia and Expo*, Vol. 1, pp. 119 -122, 2000

[33] S. Dogan, S. Eminsoy, A.H. Sadka, A.M. Kondoz, "Personalised Multimedia Services for Real-time Video over 3G Mobile Networks", *Proc. of Third Int. Conf. on 3G Mobile Communication Technologies*, pp. 366 -370, 2002

[34] Yao Wang, Qin-Fan Zhu, "Error Control and Concealment for Video Communication: A Review", *Proc. of the IEEE* , Vol. 86, No. 5, pp. 974 -997, May 1998

[35] M. B. Jill, Robert D. Gaglianello, "Packet Loss Effects on MPEG Video Sent Over the Public Internet", *ACM Multimedia 98 - Electronic Proc.*, Sept. 1998

[36] Dapeng Wu, Y.T. Hou, Wenwu Zhu, Hung-Ju Lee, Tihao Chiang, Ya-Qin Zhang, H.J. Chao, "On End-to-end Architecture for Transporting MPEG-4 Video over the Internet", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 10, No. 6, pp. 923 -941, Sept. 2000

[37] Chung-Ming Huang, Ming-Yuhe Jang, "Handoff Architectures and Protocols for Transmitting Compressed Multimedia Information in Mobile PCSs", *IEEE Trans. on PCSs, Consumer Electronics*, Vol. 43, No. 3, pp. 784 -794, Aug. 1997

[38] Y.Wang, Jörn Ostermann, Ya-Qin Zhang, *Video Processing and Communications*, Prentice Hall, 2002

[39] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group", IETF RFC 2597, June 1999.

[40] D. Black, "Differentiated Services and Tunnels", IETF RFC 2983, Oct. 2000.

[41] B. Davie , A. Charny, J.C.R. Bennett, K. Benson, J.Y. Le Boudec, W. Courtney, et al, "An Expedited Forwarding PHB (Per-Hop Behavior) ", IETF RFC 3246, Mar. 2002.

[42] N. Brady, "MPEG-4 Standardized Methods for the Compression of Arbitrarily Shaped Video Objects", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp. 1170 -1189, Dec. 1999

[43] J. Apostolopoulos, T. Wong, Wai-tian Tan, S. Wee, "On Multiple Description Streaming with Content Delivery Networks", *Proc. of Twenty-First Annual Joint*

Conf. of the IEEE Computer and Communications Societies, Vol. 3, pp. 1736 - 1745, 2002

[44] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", IETF RFC 2474, Dec. 1998.

[45] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", IETF RFC 2475, Dec.1998.

[46] J. Apostolopoulos, "Error-Resilient Video Compression Through the Use of Multiple States", *Proc. of Conf. on Image Processing*, Sept. 2000

[47] J. Apostolopoulos, "Reliable Video Communication over Lossy Packet Networks Using Multiple State Encoding and Path Diversity", *Proc. of Visual Communications and Image Processing*, pp. 392-409, Jan. 2001

[48] S. Nanda, K. Balachandran, S. Kumar, "Adaptation Techniques in Wireless Packet Data Services", *IEEE Commun. Mag.*, vol. 38, pp.54–64, Jan. 2000

[49] V.K. Goyal, "Multiple Description Coding: Compression Meets the Network", *IEEE Signal Processing Magazine*, Vol. 18, No. 5, pp. 74 -93, Sept. 2001

[50] "IP in the RAN as a transport option in 3rd generation mobile systems", Mobile Wireless Internet Forum MTR-006 v2.0.0, June 2001

[51] "General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp Interface", 3GPP TS 29.060 V5.4.0, Dec. 2002

[52] John W. Mark and Weihua Zhuang, *Wireless Communications and Networking*, Prentice Hall, New Jersey, 2002

[53] Qian Zhang, Wenwu Zhu, Ya-Qin Zhang, "Resource Allocation for Multimedia Streaming over the Internet", *IEEE Trans. on Multimedia*, Vol. 3, No. 3 , pp. 339 -355, Sept. 2001

[54] Y. Bernet, S. Blake, D. Grossman, A. Smith, "An Informal Management Model for Diffserv Routers", IETF RFC 3290, May 2002

[55] J. Schiller, *Mobile Communications*, Addison-Wesley, 2000

[56] "Combined GSM and Mobile IP Mobility Handling in UMTS IP CN", 3G TR 23.923 V.3.0.0, May 2000

[57] "Handovers for real-time services from PS domain", 3GPP TR 25.936 V4.0.1, Dec. 2001

[58] "Handover procedures", ETSI TS 100 527 V7.0.0, Aug. 1999

[59] "Handover procedures", 3GPP TS 23.009 V5.3.0, Dec. 2002

[60] "Network architecture", 3GPP TS 23.002 V5.9.0, Dec. 2002

[61] "Quality of Service (QoS) concept and architecture", 3GPP TS 23.107 V5.7.0, Dec. 2002

[62] J. Wiljakka, "Transition to IPv6 in GPRS and WCDMA Mobile Networks", *IEEE Communications Magazine*, Vol. 40, No. 4, pp. 134 -140, Apr. 2002

[63] Wei-Ge Chen, Ming-Chieh Lee, "α-channel Compression in Video Coding", *Proc. of IEEE Int. Conf. on Image Processing*, Vol. 1, pp. 500 -503, 1997

[64] J. De Vriendt, P. Laine, C. Lerouge, Xiaofeng Xu, "Mobile Network Evolution: A Revolution on the Move", *IEEE Communications Magazine*, Vol. 40, No. 4, pp. 104 -111, Apr. 2002

[65] David Leon, "RTP Retransmission Framework", IETF Internet draft, Mar. 2002

[66] J. Widmer, R. Denda, M. Mauve, "A Survey on TCP-friendly Congestion Control", *IEEE Network*, Vol. 15, No. 3, pp. 28 -37, May-June 2001

[67] "General Packet Radio Service (GPRS); Service description; Stage 2", 3GPP TS 23.060 V5.4.0, Dec. 2002

[68] S. Das, A. Misra, P. Agrawal, "TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility", *IEEE Personal Communications*, Vol. 7, No. 4, pp. 50 -58, Aug. 2000

[69] F.M. Chiussi, D.A. Khotimsky, S. Krishnan, "Mobility Management in Third-Generation All-IP Networks", *IEEE Communications Magazine*, Vol. 40, No. 9, pp. 124 -135, Sept. 2002

[70] F.A. Chiussi, D.A. Khotimsky, S. Krishnan, "A Network Architecture for MPLS-based Micro-mobility", *Proc. of Wireless Communications and Networking Conf.*, Vol. 2, pp. 549 -555, 2002

[71] T. Ahmed, A. Mehaoua, G. Buridant, "Implementing MPEG-4 Video on Demand over IP Differentiated Services", *Proc. of IEEE Global Telecommunications Conf.*, Vol. 4, pp. 2489 -2493, 2001

[72] T. Ahmed, G. Buridant, A. Mehaoua, "Encapsulation and Marking of MPEG-4 Video over IP Differentiated Services", *Proc. of IEEE Symp. on Computers and Communications*, pp. 346 -352, 2001

[73] Jitae Shin, Jong Won Kim, C.C.J. Kuo, "Quality-of-service Mapping Mechanism for Packet Video in Differentiated Services Network", *IEEE Trans. on Multimedia*, Vol. 3, No. 2, pp. 219 -231, June 2001

[74] "UTRAN Overall Description", 3GPP TS 25.401 V5.5.0, Dec. 2002

[75] T. Yoshimura, T. Ohya, T. Kawahara, M. Etoh, "Rate and Robustness Control with RTP Monitoring Agent for Mobile Multimedia Streaming", *Proc. of IEEE Int. Conf. on Communications*, Vol. 4, pp. 2513 -2517, 2002

[76] R.Bennett and H.Zhang, "WF^2Q: Worst-case Fair Weighted Fair Queuing", *Proc. of IEEE INFOCOMM '96*, pp. 120-128, March 1996.

[77] *OPNET Modeler 9.0 Documentation*, OPNET Technologies, Inc., 2002.