High Performance QoS Traffic Transmission Scheme for Real-Time Multimedia Services in Wireless Networks

Moonsik Kang

Department of Electronic Engineering, GangneungWonju National University / Gangneung, South Korea
mskang@gwnu.ac.kr

* Corresponding Author: Moonsik Kang

Received November 21, 2012; Revised December 5, 2012; Accepted December 12, 2012; Published December 31, 2012

Abstract: This paper proposes a high performance QoS (Quality of Service) traffic transmission scheme to provide real-time multimedia services in wireless networks. This scheme is based on both a traffic estimation of the mean rate and a header compression method by dividing this network model into two parts, core RTP/UDP/IP network and wireless access parts, using the IEEE 802.11 WLAN. The improvement achieved by the scheme means that it can be designed to include a means of provisioning the high performance QoS strategy according to the requirements of each particular traffic flow by adapting the header compression for real-time multimedia data. A performance evaluation was carried out to show the effectiveness of the proposed traffic transmission scheme.

Keywords: Traffic transmission scheme, High performance QoS, Traffic estimation, Header compression, Core RTP/UDP/IP network, IEEE 802.11 WLAN

1. Introduction

In the history of telecommunication, the technology development has always been driven by the human need to communicate with the ever increasing amount of information across increasing distances [1, 2]. Moreover, the recent rapid growth of the wireless mobile network has created new demands for multimedia applications over the Internet including both wired and wireless parts [8, 9]. To cope with this increasing traffic requirements, several studies have focused on providing users with the required Quality of service (QoS) at different network layers [2, 15, 18]. Here, QoS is considered the ability to provide different priorities to different applications, users, or data flows or to guarantee a certain level of performance to data flow. RTP (Real Time Protocol) is widely used as the protocol for various real-time communications and multimedia services using IP Networks, such as multi-user online games and video/audio streaming or remote video conferencing and remote patient care. When audio/video data is transferred through cable or wireless media, the packets are capsulated into an IP header or RTP, with a total header size of 32 (52 bytes in case of IPv6) bytes [11]. The header compression technique can be regarded as a process for reducing the header size by eliminating the increasing packet header overheads, which includes two representative methods of the IPHC and ROHC [10, 11]. These utilize the repetitiveness of header fields, and enable more efficient multimedia communications.

Until now, the authors have searched for the most effective method to meet the diverse QoS multimedia traffic requirements, such as efficiency and scalability. Here, this paper proposes a high performance QoS traffic transmission scheme (HPQTS), which adopts an adaptive resource control strategy including both the traffic estimation scheme and header compression method. This is a more effective solution for the required QoS from one end of the network to the other, and might include the connection between the access point and router, which is located in the border of the core-IP (/RTP) network. Many proposals suggest both DS and wireless LAN QoS methodologies to provide a better service for specific classes of traffic [3, 5, 16], but not for particular end-to-end flows [4]. The solution proposed in this paper can be used to optimize the performance of the network for different classes of traffic and apply them to the dynamic provisioning using traffic estimation of mean rate. The

Preliminary results of this paper were presented at the ICITCS2012. This present paper is an improved version which has been extended to include a modified scheme and comprehensive performance analyses.

proposed scheme was designed to allow real-time traffic transmission at the RCA (Resource Control Agent) part.

The remainder of this paper is organized as follows. Section 2 describes the related studies. The QoS service model and core network design method will be followed. Section 3 presents a new HPQTS considering the network traffic conditions and HC scheme over the core network was proposed. Section 4 evaluates the performance and discusses the effect of the proposed scheme. The final section concludes this paper.

2. Related Work

The problem of providing QoS for real-time multimedia applications in IP networks has received considerable attention [2, 6, 15, 18, 19]. On the other hand, supporting QoS over wireless links and integrating QoS mechanisms with mobility is still an open problem. Recent surveys have analyzed a range of issues and identified many research directions [8, 9, 19]. The analysis follows from some of their conclusions and applies them to the problem of providing QoS for the real-time wireless multimedia data based on traffic estimation architecture. Several authors examined a completely different approach to QoS differentiation in wireless networks by extending or modifying the MAC layer [7, 16]. Some studies related to an understanding of the QoS possibility of the IEEE 802.11e draft standard have been reported [13, 14, 17]. These papers discussed the implications of the proposed protocol through simulations.

An idea related to the present work including header compression (HC) over IP/RTP [7, 10, 11], discussed the complementary nature of 802.11e and HC, and how internetworking between them can support an end-to-end architecture [7]. Some papers discussed a different approach to the problems of providing throughput guarantees over the wireless access part. They focused on the access point (AP), which changes the contention window of all wireless stations on the flow. Work related to an end-to-end QoS architecture including the wireless network are reported in references [16, 18, 19]. The common feature appears to be the use of a DS core. As progress moves towards future networks, it appears that DS may become the QoS mechanism of choice in the core IP network, which also approaches the problem of maintaining state in the presence of mobility. The approach taken in their work in the core network, however, is similar to the present approach, differing mainly in the fact that a change in the DS core with HC scheme over IP/RTP is proposed to make it dynamic.

3. QoS Service Model and Core Network Design

The purpose of the QoS scheme is to provide prioritized services by controlling the bandwidth, jitter and latency, as well as improving the loss characteristics. In addition, it is important to ensure that providing a priority for one or more flows does not make the other flows fail. The QoS scheme can provide a better service to certain flows, by either raising the priority of a flow or limiting the priority of another flow. Referring to congestion management, this study attempted to increase the priority of flow by queuing and servicing queues in different ways. The queue management method for congestion avoidance adapts the priority by dropping the lower-priority flows before servicing the higher-priority flows.

3.1 QoS Strategy and Differentiated Service Model

Policing and shaping provide a priority to a flow by limiting the throughput of the other flows. In the BE service model, the actual forwarding treatments do not characterize any service levels but have evolved by experience to a single BE service class. In such a single service class network, there is no classification of the traffic based on the characteristics of the applications. To account for multiple traffic flows through the network and avoid the congestion caused by an uneven network operation utilization, the BE model uses the *over-provisioning* technique, in which the network simply provides sufficient bandwidth to always match the committed network service guarantees. This method involves dropping the additional incoming bandwidth on any network link at the time of congestion. The idea behind it is that the packets will not be dropped if the network load is kept low, nor will they experience high delays caused by waiting in long queues. The limitations of this method are that not every network problem can be fixed with the same remedy. Indeed, no matter how much capacity is provided by the network, new applications always appear to consume it. Therefore, congestion will inevitably appear and the QoS requirements of some applications will not be met. Thus far, the Internet has experienced tremendous growth in traffic volume as well as the diversity in the type of carried traffic, such as voice and video data. For interactive applications, the end-to-end delay also becomes a significant factor. Instead of maintaining state information, DS (differentiated service) applies different PHBs (per hop behaviors) to packets, which is specified by a DSCP (DS code point) in the DS (or type of service) field of IP header. This achieves scalability by aggregating the traffic classification state for the IP-layer packets in DS network. The PHB refers to the externally observable forwarding behavior applied at a DS-compliant node to a DS behavior aggregate. According to the different applications, two types of PHB behaviors are defined in the DS model [3]: AF (assured forward) PHB behavior and EF (expedited forward) PHB. The routers in a DS domain perform different functions. The routers can be classified into two types, boundary router and interior (or core) router. The boundary routers are responsible for connecting a DS domain to a node either in another DS domain or in a non-DS domain. SLA (service level agreement) is a type of contract between the customer and service provider, which typically covers the issue of QoS service specifications that are to be met by the service

provider [13].

A TCA (traffic conditioning agreement) is a part of the DS SLA. The TCA represents a filter to which the specific SLA is bound. This filter is a classifier for separating the traffic stream for processing. In the DS region, SLA is mapped to a conceptual network model, which is then applicable to the configuration of the individual elements within the network. This requires translation from the SLA to a more detailed SLS. The SLS is defined as a set of parameters and their values, which together define the service offered to a traffic stream by the DS domain. In addition, as a part of SLA, the TCA is translated to a DS specific traffic conditioning specification (TCS). The TCS is defined as a set of parameters specifying the traffic *profile*. The Traffic Conditioning Framework (TCF) consists of two parts, such as the traffic classifier and traffic conditioner. The former is used to select packets from an incoming packet stream according to predefined rules. In addition, two types of classifiers are defined in the DS model, which might be located at the ingress nodes or at interior nodes in the DS domain. The classifier located at the ingress node is generally a MF (multi-field) classifier. The other is a BA (behavior aggregate) classifier located at the interior routers, and is based on the DSCP value. The DSCP is a reformatted DS field of the IP header, which is used to define the class of packet. This class specifies both the forwarding scheduling and path selection.

The main components of the traffic conditioner are the meter, marker and shaper/dropper. The meter is used to measure the temporal properties of traffic flowing and its comparison against those specified in a TCA. A packet in agreement with the TCS (*in profile*) is treated differently from those that are *out of profile*. The marker is used to perform the DSCP marking of the packets. The DSCP value is defined as the SLA between the host and service provider, and PHB behaviors are determined according to the DSCP value. The shaper is used to delay the packets at the nodes where they are active and the dropper is used to discard them. The MF classifier will classify traffic from multiple sources. After this classification process, an appropriate TCS will be assigned to the traffic of each source. The BA classification is then performed based on DSCP [13]. A meter then compares the marked traffic with the traffic profile in the TCS. The conformance status of the packet is decided, and the non-conforming packets are re-marked for a lower service-level, or shaped to conform to the TCS, otherwise they are dropped.

The policing process is used to delay some or all of the packets in a traffic stream to bring the stream into compliance with a traffic profile. This normally has a finite-size buffer, and packets may be discarded if there is insufficient buffer space to hold the delayed packets. The EF PHB is intended to provide a building block for low delay, low jitter and low loss, assured bandwidth, end-to-end service through the DS domain. The AF PHB is applied to those applications such that the traffic *out of profile* will be delivered with a lower probability compared with the traffic *in profile*. Therefore a provider in the DS domain offers different levels of forwarding assurances for IP packets. In addition, four AF classes have been defined in each node by allocating a certain portion of the forwarding resources. The packets for the assured forwarding are marked with a code point by mapping them to one of these classes. The packets within the classes can be assigned to one of three-drop precedence.

3.2 Core Network and Access Network

Because of the limitations of the wireless medium, providing QoS is a far more challenging issue in 802.11 networks. In the advent of QoS in the IP Core Network, it has become imperative that the Wireless Access network also provide the required QoS. The end-to-end QoS requires not only a QoS support mechanism in the core-IP network, but also in the access networks. The 802.11e proposes two new schemes for wireless access. Enhanced Distributed Co-ordination Function (EDCF) and Hybrid Co-ordination Function (HCF). The two co-ordination functions are based on the existing co-ordination functions, and are designed to be backward compatible with them. EDCF is an extension of the existing DCF scheme with some of the elements of the MAC parameterized per Traffic Category (TC), and works to prioritize traffic based on the access categories (AC) [13, 14]. Each MAC Frame is tagged with traffic category identification (TCID) by mapping the TCIDs to the ACs. The TCIDs are not strictly in numerical order because the EDCF mechanism has been designed to be compatible with IEEE 802.1D/Q. The TCIDs are the same as the user priority tag of the 802.1D/Q Header. The EDCF works to provide a type of statistical priority for Traffic based on the TCID values.

The basic item that needs to be achieved by internetworking is to actually translate the 802.11e parameters to DS parameters. This can be achieved as shown in the table below. The four classes of traffic, as specified in the table, map the different TCIDs within the 802.11e framework. Therefore, the table is a direct map of the DSCP field to the TCID field and vice versa. This provides a simple mechanism for translation. The effort is simply replacing the appropriate fields using the mapping of Table 1. The relationship between the four traffic classes and four priority queues will be explained in detail in section 4.3

Table 1. Mapping table.

Traffic Class	Service Data Types	DSCP	TCID
TC1	Voice (VoIP)	(101)xxx for EF	7
TC2	Video Streaming	(100)xxx for AF4x	5
TC3	Signaling Data	(010) xxx for AF2x	3
TC4	Normal Data (e.g. E-mail, Web)	(000)000 for default BE	1

The contention window (CW) and Inter-frame Spacing times are adjusted so that a higher priority class has more likelihood of gaining access to the medium. The DIFS (DCF inter-frame spacing) of the DCF mechanism is now replaced with AIFS (arbitration IFS). The minimum size of the AIFS is equal to the DIFS. The others are longer than this value. Therefore, the highest priority traffic has the shortest AIFS. The lower priorities have correspondingly

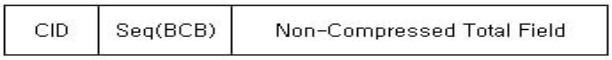
longer AIFS values. The core IP network model is an appropriate architecture for implementing a scalable service differentiation on the Internet by aggregating the traffic classification state [18]. Because of the aggregation function, the core routers in the DS network only maintain the minimum state information and provide the required QoS. A forwarding treatment is a set of rules defining the importance of a class compared to other classes. These rules characterize the relative amount of resources, which should be dedicated for a particular class in the scheduler, and the packet dropping order during congestion. The Traffic conditioner is used to determine if the offered traffic is in compliance with the agreed profile. Two types of routers have been identified in the core domain, border routers and core routers. Border routers exchange packets with other domains and perform traffic conditioning, which are allowed to maintain per-flow information. In the event of a QoS in the IP core network, it is imperative that the wireless access network also provide the required QoS. The end-to-end QoS requires not only a QoS support mechanism in the core network, but also in the access networks. The 802.11e WLAN standard should propose an EDCF for wireless access.

The scheme uses RTP as the upper layer protocol to meet a range of diverse QoS multimedia traffic requirements, On the other hand, in that case, the audio/video multimedia data is transferred by being capsulated into an IP header (20 octets) or RTP (12 octets), with a total header size of 32 bytes. Therefore it might be necessary to reduce the header size by eliminating the increasing packet header overheads. Therefore, this paper proposes a high performance QoS traffic transmission scheme with both a traffic estimation of the traffic mean rate and a header compression method for the effective solution for the required QoS from one end to the other in wireless networks.

3.3 Compressed Header Format

As mentioned in a previous section, the header compression technique is the process for reducing the header size, which enables more efficient multimedia communications. The existing header compression methods (VJHC, IPHC, and ROHC) begin to exchange the delta value of SN (sequence number) or TS (time stamp). After that, its size will be decreased. The method proposed in this paper compresses both SN and TS simultaneously. Indeed, the 48-bits of SN and TS can be reduced by 3-bits. The following shows the process of transmitted and received as well as the type of the compressed header. The following figure presents the packet format used in the compression negotiation stage.

First, the value of CID (context identification) is determined and the context DB in compressor and decompressor (C/D) is created, of which the content is precisely the same as its value. The CID becomes the key of the context DB. The context is created in the N-state (negotiation) and used in the C-state (compression), which consists of static fields (including the source and destination address fields). The values of the fields will not change between streaming units, and when the N-state begins, the context is constructed using the uncompressed full header. Here, the decompressor can process the compressed packet using the received CID. The following figure shows the first full header packet format transmitted by the compressor after a negotiation process.



The following figure shows the typical packet form in the compressed stage, in which the compressor is in a normal operation with the best compression rate.

Therefore, a method using both the context and NCB bits can be used to compress the header more efficiently in cooperation with the upper layer protocols, such as Application, RTP, UDP and IP.

4. The Proposed Scheme (HPQTS)

The access part of the proposed network model will also support legacy 802.11e users, which should be capable of interfacing with the remainder of the core network. In this regard, the AP becomes the end point for PHB operation as a service level specification (SLS).

4.1 Network Model for QoS Traffic

Fig. 1 shows the proposed model for high performance QoS traffic transmission scheme (HPQTS). The boundary entity is co-located at the ingress router to the core part as a border router (BR), which has a number of functions and is under the control of the resource control agent (RCA). The BR is in charge of receiving packets from the access router (AR) and marking them with an appropriate DS code point (DSCP), which is not necessarily a direct translation of the UP tag. If the incoming traffic is in excess of what is expected it will be marked simply as BE traffic. In this way, the BR performs admission control for incoming traffic. Another important function of the BR is the marking of packets so that the core network can easily recognize it, which is important for the translation of information between the two networks. The RCA can also instruct the BR to drop packets from certain users, or of a certain flow. The BR forwards all incoming traffic information to the RCA and should provide policing to account for falsification. This can be achieved using standard means, such as token bucket/leaky bucket policing. Here, the internal routers within the core network are called core routers (CR). The BR, can forward messages to the RCA, which ultimately decides whether or not to grant a certain request.

The CR performs the PHB function of the DS network,

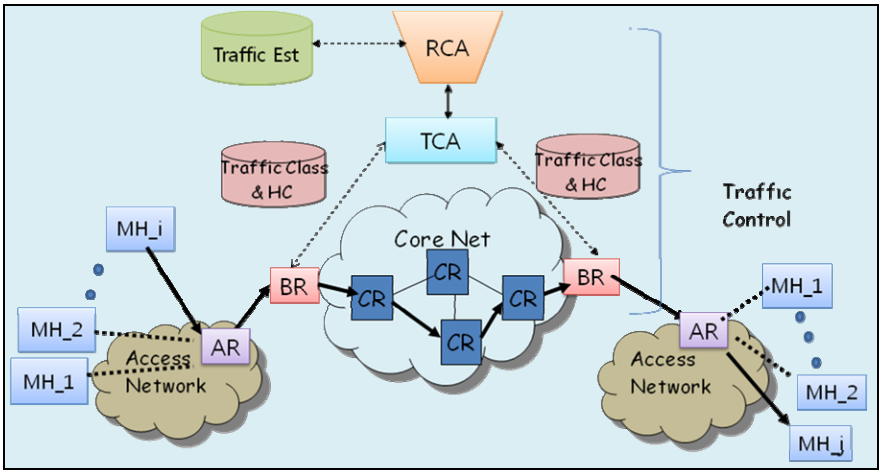


Fig. 1. HPQTS Network model for real-time multimedia services.

which is divided into two parts. The first is the priority queuing part, i.e. EF, AF and BE. The other is the drop precedence part, in which the packets in the DS network can also be marked with a particular drop precedence. When the CR needs to drop packets, this part will be used to determine which packets to drop. This concept is different from the weighted queuing because the drop precedence affects the service in a different manner. Both the weighting of individual queues and the drop precedence are under the control of the RCA. The RCA can dynamically change the configuration of the CR. RCA can be used as the central management entity, which is in charge of traffic monitoring and dynamically provisioning the core network using traffic measurements. This may be supported by a database to store SLA information and traffic measurement information.

A range of experiments on the MPEG streaming data with RTP/UDP/IP packets in networks were performed to find an efficient header compression method of the RTP protocol in multi-user, online game's peer-to-peer multimedia services or video streaming services using IP wireless networks. For the highest header compression rate for the SN (sequence number) and TS (time stamp), which increases dynamically and constantly among the many fields in the RTP header, the proposed scheme has the following features [11]:

(1) The increase in both SN value and TS value at the packets between compressor and decompressor (C/D) should be maintained uniformly in the following manner. The SN increases in increments of one if the BCB (basic compression bit) is determined to be the demanded bit value resulting from the compression. The TS increases according to the PCTSI (picture clock interval) times.

(2) The total amount of (SN plus TS) has a value of 48 bits. The C/D first decides how many of these 48 bits should be compressed. This can be decided as the 3-bits as BCB. If this value is inadequate, the value of BCB is decided by an adjustment between them, which is called the NCB (negotiation compression bit). The NCB value is the same as BCB if 3-bit is suitable.

(3) Finally the compressor compresses the total 48-bits (16-bit SN plus 32-bit TS) into n NCB bits, and n bits are then transferred to the decompressor. The SN value is than calculated using (SN_{n-1} + 1) with the previous values and then restores the TS as (PCTSI + TS_{n-1}). As a result, the 48-bits might be compressed into n bits.

Furthermore, consider that the BCB of SN and the TS between C/D nodes can be set to 3-bits. The C/D determines the final NCB according to the characteristics of the payload traffic transferred through the RTP, and both SN and TS are compressed based on this result. When the NCB value is determined to be n, a decision is made to determine to whether to compress the 48-bits of SN and the TS in 3-bits (BCB) or to compress them with an additional n bits in 3+n (NCB). As a result, the 48-bits of (SN and TS) value are sent from the compressor after being compressed in NCB (BCB + n bits, $0 \leq n < 48$). This means that long streamed or large video traffic can be compressed with a value larger than 3-bits, whereas small files or small traffic is compressed by the BCB, 3-bits. Here, the n NCB is determined by calculating the basic and extended bits to increase the compression rate and prevent cases where the decompressor itself has problems of restoring the lost consecutive packets of more than 2^n - 1. Considering the ability to restore consecutive packet loss, the n-bit size was set according to dependence on the negotiated results between C/D. Consequently, the number of compressed bits can be extended to NCB (BCB+n). In other words, if a large number of video packets are being transmitted through the wireless network and the link status is improper, the robustness against multiple packet losses can be provided using the diverse expansion of the compression range. In other case, the compression rate can increase by configuring the bit value at the BCB. If the 48-bits of SN and TS are compressed in BCB 3-bits, the compression rate increases, whereas the robustness of the protocol is lowered. On the other hand, the compression rate is lowered if they are compressed in (3 + n) by adding the extended_n bits to the BCB 3-bits, but the protocol robustness increases. Fig. 2 summarizes the determination process of NCB.

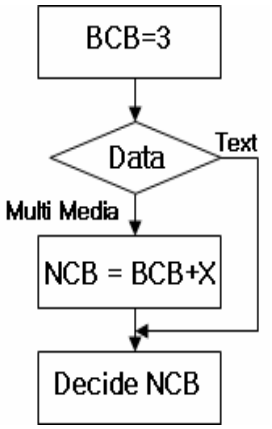


Fig. 2. Determination of NCB.

4.2 SN and TS Compression method in Core Network

The finite state machine of a decompressor consists of three states: Build-state, Normal-state, and Repair-state, as shown in Fig. 3. At the build-state, the context will be constructed and then the first packet transmitted to, while negotiating the value of the NCB. The normal-state sends the packets with the highest compression rate, corresponding to the C-state. When an error occurs, the state of the decompressor will be changed to the repair-state. After repairing the errors, it returns to the normal state.

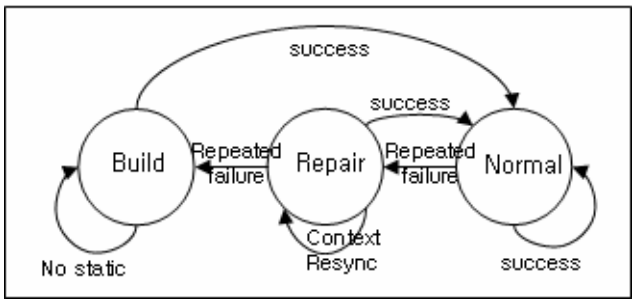


Fig. 3. Decompressor finite state machine.

An existing method for the sequence number (SN) in the RTP is to increase the sequence numbers of the consecutive packets by increments of one. That is, the RTP packet with a 16-bit SN increases by 1 from 0 to 65535. This is because it is easy to detect the loss of a packet in the part of the reception. On the other hand, consider the case of achieving a higher compression rate. When the n^{th} packet's SN is 3000, the SN of the $(N+1)^{th}$ SN is $3000 + 8(2^3) = 3008$. Here, SN_n is the original sequence number in the RTP packet, and the NCB is the basic 16-bit SN compression bits. Therefore, if n=3, the 16-bit SN is compressed in 3-bits. When the range is expanded in the proposed method, the SN can be represented in equation (1), where SN_n is the RTP's nth packet's sequence and SN_{n-1} is the most recently decoded sequence number.

$$SN_n = SN_{n-1} + 1 \tag{1}$$

If the SN is compressed in such a way, the SN compression rate of the RTP header can be calculated using the following relationship.

$$C_{SN} = \frac{SN - SN_{proposed}}{SN} = 1 - \frac{SN_{proposed}}{SN} \tag{2}$$

Here, the SN is the uncompressed packet header size (byte), and the $SN_{proposed}$ is the proposed compressed size of the SN. For example, if the NCB is 3 bits, the value of $SN_{proposed}$ becomes 3 bits, i.e. the same size of NCB. The C_{SN} is the compression rate of the SN, i.e. the relative ratio of the compressed SN size. Regarding the constitution of the SN, to achieve higher compression efficiency, the changes in the current RTP TS (Time Stamp) value should be set at constant multiple values. This paper proposes that the TS value of the RTP header increases by the PCF (picture clock frequency) value. In a video stream, within the same picture, the difference in TS becomes 0. The TS value for the Intra-coded frames (I-frames), Inter-predicted frames (p-frames) or the following Bi-directionally predicted frames (B-frames) can be greater than PCF or larger than 0. The PCF of H.261, H.263 ver. 1 is 29.97Hz, which means that the increase between the two coded pictures is 3003. The following represents why such a PCTSI (picture clock interval) value is calculated. When the video is compressed using RTP, a packet may not include more than 1 picture. A single picture may be divided into two or more packets. The header compression profile regarding video information uses 90 kHz as the reference clock [8].

$$PCTSI = \frac{Video\ packet\ reference\ clock\ (90KHz)}{PCF} \tag{3}$$

In addition, the PCF values of the H.263 ver. 2 and MPEG-4 become 25Hz and 30Hz, respectively. These values increase in multiples of 3600 and 3000, respectively. Although there is a loss in the transfer process between C/D because the TS increases by the constant multiple of PCTSI, it is restored to the normal state immediately with the TS value of the previous packet. Similarly, if the TS value of the RTP header is configured to increase by the constant multiple of PCTSI, the TS value can simply be compressed and restored using the following equation.

$$TS_n = TS_{n-1} + PCTSI \tag{4}$$

If TS_n is the time stamp of the n^{th} packet of the RTP and TS_{n-1} is the most recently decoded time stamp, when the range of the time stamp is expanded by the suggested method, it can be represented by Eq. (4). If the change in TS between two packets is configured to be a multiple of PCTSI and the PCTSI value is shared between C/D, the decompressor can use the previously mentioned BCB n-bit because it restores the TS value easily. Using this method the compressor might use either 3-bits or (3+extended_n) bits to compress and restore the 16-bit SN and 32-bit TS. The compression rate of the TS of RTP header can be calculated using Eq. (5).

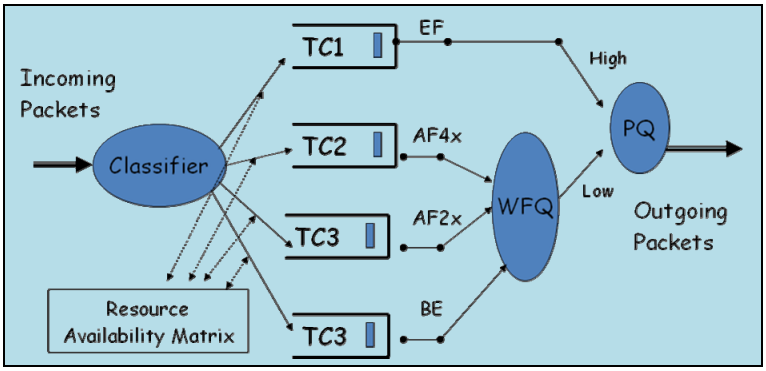


Fig. 4. Bandwidth allocation scheme and traffic classifier.

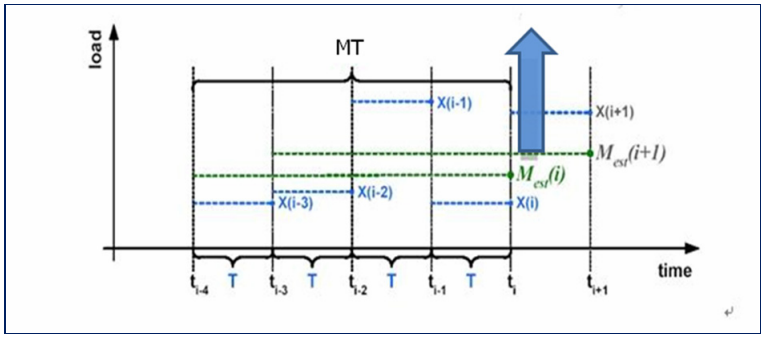


Fig. 5. Traffic estimation using the moving average method.

$$C_{TS}=1-\frac{TS_{proposed}}{TS} \quad (5)$$

where TS is the uncompressed packet header size and the $TS_{proposed}$ is the proposed compressed TS size. For example, $TS_{proposed}$ might be regarded as 0 when the TS is decompressed using NCB bits at the decompressor. The C_{TS} is the compression rate of TS, i.e. the relative ratio of the TS size.

4.3 Bandwidth Allocation

In this model, the ingress traffic can be monitored and measured continuously by obtaining the moving average of the rate. In addition, the network continuously collects data traffic from each link. Over time, this collection of data is used to characterize the behavior of the traffic, e.g. what kind of traffic dominates at a particular time of the day. This enables the data for the traffic patterns to be compiled over time, which helps define the required parameters in the core of the network to support this variation in traffic with time.

As shown in Fig. 4, the basic concept within a CR is that of four priority queues: One for EF (for high priority traffic i.e., Expedited Forwarding traffic), two queues for AF4x and AF2x (i.e., Assured Forwarding traffic), and one for BE (i.e., Best effort traffic). The different priorities for weighting algorithms, such as Priority Queuing (PQ), govern each of these queues. Fig. 4 shows that the TC1 queue, which is the highest priority one, passes through a single weighing stage of PQ, whereas TC2 and TC3 traffic pass through two levels of the weighing stage of WFQ. Therefore, the TC1 has highest precedence and the smallest number of weighing stages. Similarly, the network service maps to different Traffic Classes. For example, the real time voice traffic is for TC1, which is implemented using EF.

The traffic matrix is specified based on a traffic estimation according to the traffic measurement method, as shown in Fig. 5. This figure shows the method for estimating the mean rate using the moving average obtained by calculating with the following Eq. (6). Here, T represents the sampling interval, X(i) means the measured mean rate in i, and M means the window size (number of sampling intervals). The values within such a pattern matrix will be typical or normal values, of which the normal values are within a predefined threshold. In the normal state, the network can be provisioned safely according to the proposed pre-specified matrix by allocating the optimal bandwidth.

$$M_{est}(i)=\frac{1}{M}\sum_{j=0}^{M-1}X(j-j) \quad (6)$$

Continuous monitoring of the incoming traffic can be used to determine if at any given time, the incoming traffic is within the bounds of the expected traffic. Fig. 6 shows the resource control strategy according to the traffic condition. In the presence of sudden changes, the network enters an abnormal state. In this case, the network reacts by further changing the weight in discordance with the matrix above. As soon as the network returns to the normal state, the parameters are brought back to the recommended

values. When the network condition is monitored continuously and the moving average rate is estimated, it provides an opportunity to record the variations in traffic pattern over longer periods of time. As a result, the weights within the matrix can be defined and re-defined over time as a continuously varying function.

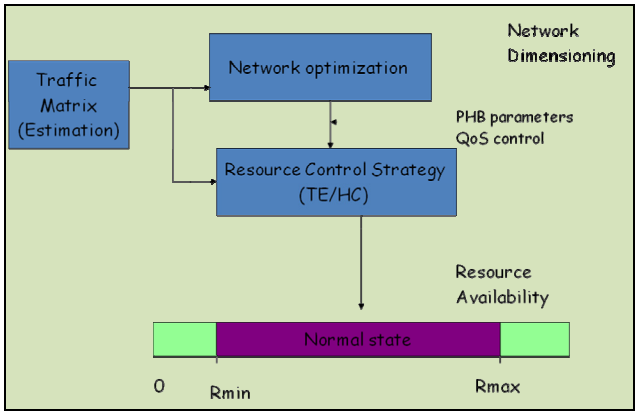


Fig. 6. Resource control strategy according to the traffic condition.

5. Performance Evaluation

Some QoS parameters, such as the delay, packet loss and throughput, were considered to evaluate the performance of the proposed model. The delay parameter might refer to either the propagation delay or round-trip delay. The packet loss parameter refers to the ratio among the numbers of lost packets from the source to destination. In addition, the throughput can be defined as the bit rate coming out of the last hop of the service scope in the destination domain. These parameters may or may not be considered an independent parameter, because they can be calculated as a function of both the transmission rate and packet loss depending on the definition of the traffic profile. The state of the network is determined by the rate at which the packets arrive and depart from various queues as well as by the set of entities waiting for service.

The proposed network model may be described as a collection of individual CR nodes and link states of which the traffic consists of the entities in the core network as well as the access part. The traffic is represented explicitly because they determine the behavior of the nodes and links at a particular point in time. In the present simulation, the performance was evaluated from the view of the network efficiency by the mean delay. Here, the moving average method of the mean rate model was used for traffic measurement. The traffic arrival process is selected as the Poisson model with the mean throughput of 150Mb/s for the simulation network model. In addition, to test the operation of the header compression, the Maximum Transmit Unit (MTU) in the RTP layer was assumed to be 500 bytes. This means that video frames larger than 500 bytes are segmented and transported after being divided into 2 or more RTP packets. On the other hand, in this simulation, the RTP packets were assumed to divide into the same size, and the RTP packets were assumed to have a constant size less than 500 bytes. Two channels were used between the two nodes, and the data channel was used to transmit data between C/D, while the feedback channel was used to transmit the ACK or NAK signal from the decompressor. The average data transmission speed was approximately 8Kbps, the frame rate was 10fps, and the video frame size was 100 bytes, which means the packet segmentation procedure is ignored in this simulation.

Fig. 7 shows the simulation results for the error restoration rate for MPEG4 video streaming data. Here, the x- and y-axis represents the length of the consecutive packet loss and the decompression success rate, respectively. The results show that the success rate for the 3-bit compression method increases for smaller lengths of consecutive packet loss, whereas the error rate for the method of large compression bit number decreases with increasing length of consecutive packet loss. In addition, for 3-bit compressions, the difference between the consecutive error rates of 10 and 50 was more than 8 times, whereas the 9-bit compression had a similar effect from the consecutive error lengths, as the error rate was maintained from 0.4 to 0.6. This means that it would be more efficient to perform compression and restoration by increasing the compression bit numbers of SN and TS to (BCB + extended_n) in the environment of a high network error rate, or the large consecutive packet error lengths.

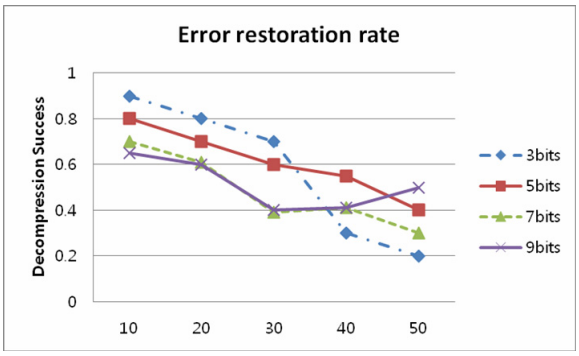


Fig. 7. Error restoration rate for MPEG4 video streaming.

In the simulation, the proposed scheme based on traffic estimation (HPQTS_TE) and weighted fair queuing (WFQ) functions were implemented between the BR and the core routers in the core network, respectively. Fig. 8 shows a comparison of the mean access delay according to each traffic class for two cases. Here, the x- and y-axis represents the traffic class and mean access delay time in msec, respectively. Case 1 means that the traffic amount of each class occurs at the same rate, whereas case 2 means that the traffic occurrence rates for TC1, TC2, TC3 and TC4 are 38%, 22%, 15% and 25%, respectively. These results show that the delay performance of TC1 (the traffic with relative high priority) of HPQTS_TE was improved considerably compared to other lower priority traffic, and TC1 showed much better performance in case 2.

Fig. 9 shows the mean delay performance for the over traffic case in the presence of TC1 traffic over 50% compared with the others in the network (55%, 15%, 18%, and 12%, respectively). Here, the x- and y-axis represents

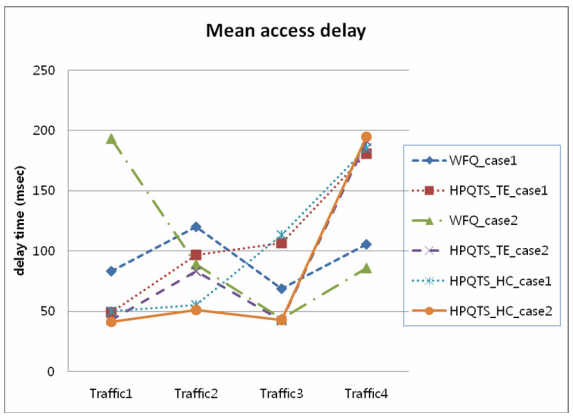


Fig. 8. Mean access delay according to the traffic types for two cases.

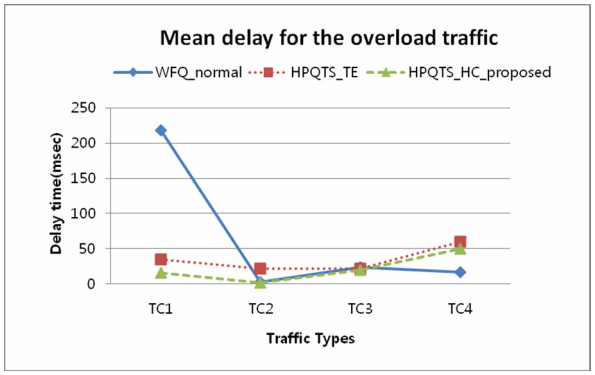


Fig. 9. Average traffic delay for the overload traffic condition.

the traffic types (class) and mean delay time in msec, respectively. This result shows that the delay performance for TC1 traffic of HPQTS_TE is improved considerably, compared to TC2, TC3 and TC4, which require attention because of the proper operation of the proposed scheme for the high priority class. In particular, when the HC scheme is applied to TC2 (for video traffic) and the traffic density is heightened, the proposed HPQTS_HC performance is much greater for that traffic than for the other schemes.

6. Conclusion

A high performance QoS traffic transmission scheme (HPQTS) based on a traffic estimation of the mean rate and header compression technique was introduced for the multimedia applications in wireless networks. This scheme can provide efficient end-to-end QoS performance between the mobile hosts in wireless access parts through an IP (/RTP) core network. The multimedia data was classified into four different types and forwarded to their destination by passing through BR and CR via both the wireless access part and core part. The performance of the proposed scheme was evaluated using the simulation network model at some aspect of the mean access delay. The simulation showed that the proposed HPQTS has much better performance with a low delay due to header compression over the core network. This solution will be also a scalable one in the core network because of the aggregation behavior. A further study will attempt to find the optimal solution for the adaptive QoS resource control strategy to cooperate with cognitive networks.

References

[1] Carolina Fortuna and M. Mohorcic, “Tends in the development of communication networks: Cognitive networks”, *Journal of computer Networks, ELSEVIER*, Vol.53, 2009. Article (CrossRef Link)
[2] A.F. Bayan and Tat-Chee Wan, “A Scalable QoS Scheduling Architecture for WiMAX Multi-Hop Relay Networks”, *2010 ICETC*, 2010.
[3] Article (CrossRef Link)
[4] Fei Zhang and James Macnicol, “Efficient Streaming Packet Video Over Differentiated Service Networks”, *IEEE Trans. on Multimedia*, Vol. 8, No. 5, Oct. 2006.
[5] Article (CrossRef Link)
[6] D. Qin and N. Shroff, “A Predictive Flow Control Scheme for Efficient Network Utilization and QoS “, IEEE/*ACM Trans. On Networking*, Vol. 12, No. 1, Feb. 2004. Article (CrossRef Link)
[7] S. Wang, Dong Xuan, Wei Zhao, "Providing Absolute Differentiated Services for Real Time Applications in Static Priority Scheduling Networks", *IEEE/ACM Transaction on Networking*, Vol. 12, No. 2, Apr. 2004.
[8] Article (CrossRef Link)
[9] Xavi Masip-Bruin, Marcelo Yannuzzi, et al., “The EuQoS System: A Solution for QoS Routing in Heterogeneous Networks”, IEEE Communications Magazine, Feb. 2007.
[10] Article (CrossRef Link)
[11] M. Kang, “An Efficient High-Speed Traffic Control Scheme for Real-Time Multimedia Applications in Wireless Networks”, *LNEE Vol.215 (ICITCS2012)*, Dec. 2012. Article (CrossRef Link)
[12] Xavi Masip-Bruin, Marcelo Yannuzzi, et al., “The EuQoS System: A Solution for QoS Routing in Heterogeneous Networks”, IEEE Communications Magazine, February 2007.
[13] Article (CrossRef Link)
[14] Lajos HanzoII and R. Tafazolli, “A Survey of QoS Routing Solutions for Mobile AD-HOC Networks”, IEEE Communications, Vol. 9, No. 2, 2007. Article (CrossRef Link)
[15] Haipeng Jin, "Performance Comparison of Header Compression Schemes for RTP/UDP/IP Packets”, *IEEE Communications Society,* 2004.
[16] Article (CrossRef Link)
[17] M. Kang et al., “Header Compression of RTP/UDP/IP Packets for Real Time High-Speed IP Networks”, *LNCS Vol.4413*, Dec. 2007. Article (CrossRef Link)
[18] V.kikuchi, T. Nomura, "RTP Payload Format for MPEG-4 Audio/Visual Streams", *RFC3016*, Nov. 2000. Article (CrossRef Link)
[19] IEEE 802.11 WG Draft Supplement to International Standard, IEEE 802.11e/D2.0, Nov. 2001. Article

(CrossRef Link)
[20] Mary Baker et al, "Using IEEE 802.11e MAC for QoS over Wireless", *Proceedings of IPCCC 2003*, April 2003. Article (CrossRef Link)
[21] Garcia-Macias et al, "Quality of Service and Mobility for the Wireless Internet", *ACM WMI*, 2001. Article (CrossRef Link)
[22] I. Mahadevan, K. M. Sivalingam, "Architecture and Experimental Framework for Supporting QoS in Wireless Networks using Differentiated Services", ACM 2001. Article (CrossRef Link)
[23] Albert Banchs et al, "Providing Throughput Guarantees in IEEE 802.11e Wireless LANs", *Proceedings of WCNC 2002*, Mar.2002. Article (CrossRef Link)
[24] Panos Trimintzios et al, "Quality of Service Provisioning for Supporting Premium Services in IP Networks", *Proceedings of the IEEE Global Telecommunications Conference*, Nov. 2002. Article (CrossRef Link)
[25] Sorniotti et al, "Design Guidelines for an Internet Scaled QoS Framework", *Proceedings of ECUMN'07*, 2007. Article (CrossRef Link)

Moonsik Kang is Professor of department of Electronic Engineering at College of Engineering, Gangneung Wonju National University (GWNU), South Korea. He is Director of the Institute of Engineering Research at GWNU. He received his B.S. and M.E. degrees in Electronic Engineering from Yonsei University, South Korea, in 1985 and 1988, respectively and his Ph.D. in Electronic Engineering from the same University in 1993. Dr. Kang was a post doctorate research associate at department of Electrical and Electronic Engineering, University of Pennsylvania, PA, USA. Also he worked as a research associate at department of Electronic and Computer Engineering, Illinois Institute of Technology, IL, USA. In addition, he had worked as a Researcher with Samsung Electronics, South Korea. He also served or is currently serving as a reviewer and Technical Program Committee for many important Journals, Conferences, Symposiums, Workshops in Computer Networks area. His research interests include High-Performance Wired/Wireless Network Protocols, Convergence Technology for Next Generation Networks, QoS traffic control schemes, and Mobile multimedia traffic modeling and Applications.