A Practical RTP Packetization Scheme for SVC Video Transport over IP Networks

Kwang-deok Seo, Jin-soo Kim, Soon-heung Jung, and Jeong-ju Yoo

Scalable video coding (SVC) has been standardized as an extension of the H.264/AVC standard. This paper proposes a practical real-time transport protocol (RTP) packetization scheme to transport SVC video over IP networks. In combined scalability of SVC, a coded picture of a base or scalable enhancement layer is produced as one or more video layers consisting of network abstraction layer (NAL) units. The SVC NAL unit header contains a (DID, TID, QID) field to identify the association of each SVC NAL unit with its scalable enhancement layer without parsing the payload part of the SVC NAL unit. In this paper, we utilize the (DID, TID, QID) information to derive hierarchical spatio-temporal relationship of the SVC NAL units. Based on the derivation using the (DID, TID, QID) field, we propose a practical RTP packetization scheme for generating single RTP sessions in unicast and multicast transport of SVC video. The experimental results indicate that the proposed packetization scheme can be efficiently applied to transport SVC video over IP networks with little induced delay, jitter, and computational load.

Keywords: Scalable video coding, unicast/multicast, RTP packetization, RTP payload format.

I. Introduction

The Joint Video Team of the ITU-T VCEG and the ISO/IEC MPEG has recently standardized a scalable video coding (SVC) as a scalable extension of H.264/AVC with the aim of enabling the creation of a compressed bitstream that can be rapidly and easily adapted to fit with the bit-rates of various transmission channels and with the display capabilities and computational resource constraints of various receivers [1], [2]. The scalability in SVC is achieved by taking advantage of the layered approach already known from previous video coding, and the scalability is represented by three fundamental types, namely, spatial, temporal, and quality (SNR) scalabilities [1]. The layers of SVC include one base layer and one or more scalable enhancement layers that can be stacked on top of each other. The more scalable enhancement layers the SVC stacks, the more diverse bit-rates, frame-rates, and resolutions it is possible to support. Thus, the SVC is the most promising coding technique suitable for multimedia contents service in a universal media access (UMA) environment that can solve the problem of variability in bandwidth, the problem of variability in receiving terminal performance and resolution, the problem of consumers' various preferences to contents, and so on [3]. SVC inherits the structure of H.264/AVC, which is divided into two parts, namely, the so-called video coding layer (VCL) and the network abstraction layer (NAL).

SVC in general provides a good basis for efficiently adapting the media content to diverse usage contexts that may even change dynamically. However, in a streaming scenario there are several options for actually deploying SVC and SVC-based content adaptation. The options range from simple server-client architectures to more complex delivery architectures involving several adaptation nodes located along the content delivery

Manuscript received Sept. 1, 2009; revised Jan. 14, 2010; accepted Jan. 15, 2010.

This work was supported by the IT R&D Program of MKE/KCC/KEIT (Development of Open-IPTV Technologies for Wired and Wireless Networks, 2008-S-006-02).

Kwang-deok Seo (phone: +82 33 760 2788, email: kdseo@yonsei.ac.kr) is with the Computer and Telecommunications Engineering Division, Yonsei University, Wonju, Gangwon, Rep. of Korea.

Jin-soo Kim (email: jskim67@hanbat.ac.kr) is with the Department of Multimedia Engineering, Hanbat National University, Daejeon, Rep. of Korea.

Soon-heung Jung (email: zeroone@etri.re.kr) and Jeong-ju Yoo (email: jjyoo@etri.re.kr) are with the Radio & Telecommunications Convergence Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.10.1409.0031

path. Such architectures aim at minimizing adaptation delay by placing adaptation nodes close to a location where dynamically changing usage environments are expected, such as in the wireless access networks of the end consumers. Another aim of such in-network adaptation architectures is to save bandwidth in service scenarios where multiple consumers wish to consume the same content. In such scenarios, a single SVC stream is delivered to the access network of the content consumers and only there it is replicated for, and adapted to, the usage environment of each consumer, thus saving bandwidth in the core network.

To transport SVC video encapsulated in NAL units over Internet protocol (IP) in real-time, real-time transport protocol (RTP) has been generally employed, [4], [5]. Although an RTP payload format for loading the SVC NAL units onto an RTP payload part is currently disclosed in an Internet-draft document *draft-ietf-avt-rtp-svc-18.txt* [6], no research has provided a result yet on a practical RTP packetization scheme that can support the standardized RTP payload format for SVC in transporting since the SVC bitstream based on combined scalability is of a complicated structure that stores scalable enhancement layer NAL units as well as base layer NAL units in a single bitstream. Therefore, we propose a practical RTP packetization scheme for combined scalability of SVC in transporting SVC video over IP networks in unicast and multicast services.

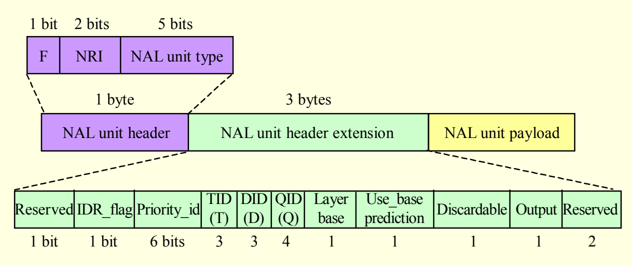


Fig. 1. SVC NAL unit structure [7].

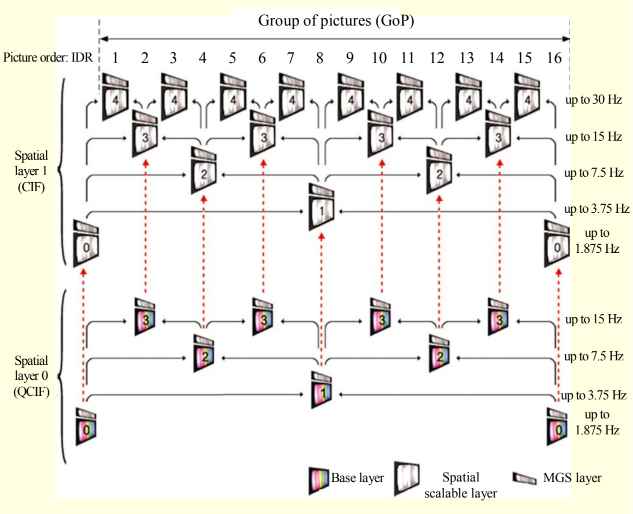


Fig. 2. Exemplary SVC bitstream structure with combined scalability.

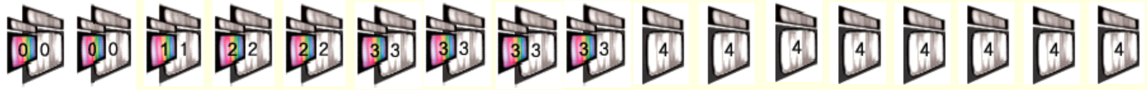
II. Structure of SVC NAL Unit

An SVC bitstream consists of one or more NAL units. Each NAL unit consists of a header of four octets and the payload byte string as shown in Fig. 1 [1], [6]. The first octet of the header shares the syntax with the one presented in H.264/AVC and indicates the type of the NAL unit (*NAL_unit_type*), the potential presence of bit errors or syntax violations in the NAL unit payload (*forbidden_zero_bit*), and information on the relative importance of the NAL unit for the decoding process (*nal_ref_idc*). In the H.264/AVC standard, NAL unit types 14, 15, and 20 have been reserved for future extensions. SVC now uses these three NAL unit types. NAL unit type 14 is used for the prefix NAL unit, NAL unit type 15 is used for SVC sequence parameter sets, and NAL unit type 20 is used for coded slice in scalable extension. The prefix NAL unit includes descriptive information of the following H.264 compatible base layer NAL unit which does not have a (*DID, TID, QID*) field in its header. The scalable NAL unit of type 20 consists of a header of four octets and the payload byte string to encapsulate VCL data. Therefore, NAL unit types 14 and 20 indicate the presence of three additional octets in the NAL unit header extension as shown in Fig. 1. The NAL unit header extension part extends the NAL unit header conforming to H.264/AVC by three additional octets and mainly provides the layer decoding dependency information. In the NAL unit header extension part, *TID* (*temporal_id*) indicates the hierarchy between temporal layers for temporal scalability, *DID* (*dependency_id*) denotes the inter-layer coding dependency hierarchy between higher/lower scalable enhancement layers for spatial scalability, and *QID* (*quality_id*) designates the quality level hierarchy of medium grain scalability (MGS) layers for quality scalability.

For more details on the syntax and semantics of the SVC NAL unit header, please refer to [1] and [6].

III. Structure of SVC Bitstream

This section describes the structure of the SVC bitstream. Figure 2 shows how to construct a combined scalability of SVC with two spatial scalability layers and five temporal scalability levels in a single bitstream. Each spatial layer consists of a quality base layer and a quality enhancement layer (MGS layer). The input pictures in spatial layer 0 are created by down-sampling the input pictures in spatial layer 1 by a



Picture order:	IDR	16	8	4	12	2	6	10	14	1	3	5	7	9	11	13	15
Quality base layer of spatial layer 0 (DID=0, QID=0)	5/(000)	1/(000)	1/(010)	1/(020)	1/(020)	1/(030)	1/(030)	1/(030)	1/(030)	→ (*DID*, *TID*, *QID*) field is referred from prefix NAL unit.							
MGS layer of spatial layer 0 (DID=0, QID=1)	20/(001)	20/(001)	20/(011)	20/(021)	20/(021)	20/(031)	20/(031)	20/(031)	20/(031)								
Quality base layer of spatial layer 1 (DID=1, QID=0)	20/(100)	20/(100)	20/(110)	20/(120)	20/(120)	20/(130)	20/(130)	20/(130)	20/(130)	20/(140)	20/(140)	20/(140)	20/(140)	20/(140)	20/(140)	20/(140)	20/(140)
MGS layer of spatial layer 1 (DID=1, QID=1)	20/(101)	20/(101)	20/(111)	20/(121)	20/(121)	20/(131)	20/(131)	20/(131)	20/(131)	20/(141)	20/(141)	20/(141)	20/(141)	20/(141)	20/(141)	20/(141)	20/(141)

Fig. 3. NAL units order in an SVC bitstream and the corresponding set of *NAL_unit_type*/(*DID*, *TID*, *QID*) information.

factor of two. A group of pictures (GoP) with size 16 is encoded by the hierarchical B-picture technique to obtain four temporal scalability levels in spatial layer 0 and five temporal scalability levels in spatial layer 1. The pictures of the coarsest temporal resolution are encoded first, and then B-pictures are inserted at the next finer temporal resolution level in a hierarchical manner. The lowest spatial layer (layer 0) has quarter common intermediate format (QCIF) resolution and four temporal scalability levels with frame rates of 1.875 Hz, 3.75 Hz, 7.5 Hz, and 15 Hz, respectively. The higher spatial layer (layer 1) has CIF resolution and five temporal scalability levels that give the additional maximum frame rate of 30 Hz.

To represent different resolutions in different spatial layers, the *DID* value in the (*DID, TID, QID*) field is used [1]. That is, in Fig. 2, the NAL unit with *DID*=0 corresponds to a picture with a resolution of QCIF, and the NAL unit with *DID*=1 corresponds to a picture with a resolution of CIF. A hierarchical B-picture technique is applied for provision of temporal scalability, and the *TID* value in the (*DID, TID, QID*) field is used to display a supportable frame rate.

In Fig. 2, the *TID* value is displayed in the middle part of each picture indicated in a rectangle. In the case of transmitting only a key-picture with *TID*=0, the frame rate can be supported at frame rates up to 1.875 Hz. In the case of additionally transmitting B-pictures with *TID*=3 and *TID*=4, the frame rate can be extended up to 15 Hz and 30 Hz, respectively.

If the NAL units at the same point of time which belong to spatial layers 0 and 1 have the same *TID* value, inter-layer prediction can be executed in the direction of the arrows indicated by dotted lines in Fig. 2. Since each spatial layer generates one MGS layer for support of quality scalability, the NAL units pertaining to the quality enhancement layer are all set to *QID*=1.

Figure 3 shows the NAL unit order in an SVC bitstream for a GoP with two spatial layers consisting of a quality base layer and a quality enhancement layer as in Fig. 2. The figure also shows the corresponding set of *NAL_unit_type*/(*DID, TID, QID*) information of each NAL unit. The NAL units are serialized in decoding order, but not in picture display order. It begins with the lowest temporal level. The temporal level is increased after the NAL units of all spatial layers for a temporal level are arranged. For each NAL unit, the corresponding *NAL_unit_type* and (*DID, TID, QID*) information contained in its NAL unit header is denoted below the NAL unit in Fig. 3.

As seen in Fig. 3, the encoding of an IDR picture occurs first. For the first IDR picture, one base layer NAL unit having the one octet basic header of Fig. 1 is generated in the base layer (quality base layer of spatial layer 0), and three scalable enhancement layer NAL units including the three octets extension header of Fig. 1 are generated in the three scalable enhancement layers (MGS layer of spatial layer 0, quality base layer of spatial layer 1, and MGS layer of spatial layer 1). As shown in Fig. 3, for the combined scalability of the SVC, analyzing the *NAL_unit_type* and (*DID, TID, QID*) field of the NAL units can detect the payload type of the data contained in the NAL units and derive the temporal and spatial hierarchy between the NAL units. In this paper, such information can be utilized in designing an RTP packetization scheme for SVC video transport over IP networks.

IV. Suggested RTP Packetization Mode for SVC Video Transport

In a previous study [8], we demonstrated an appropriate RTP packetization mode for SVC unicast transport over IP networks. In this paper, we extend the scope to multicast transport over IP networks.

To transport an SVC bitstream over IP networks, a new payload format for RTP is currently being specified in IETF [6]. An RTP stream carrying only one layer would carry NAL units belonging to that layer only. An RTP stream carrying a complete scalable video bitstream would carry NAL units of a base layer and one or more enhancement layers. In the former case, however, the system administrator of the server should open a separate user diagram protocol (UDP) port for each RTP session to carry a single layer. Thus, the server should open as many ports as required to transport all the layers. System administrators would like to avoid opening too many UDP ports in their firewalls because of the security risk and the administrative effort. Moreover, for mass deployment to a number of end terminals, it is desirable to reduce the number of UDP ports in a firewall to the absolute minimum, ideally to a single one [7]. In this respect, the latter approach is much preferred to the former one.

This line of thought leads to the unicast transport scenario as depicted in Fig. 4, where the server opens only a single RTP session to carry one or more layers for unicast, utilizing only a single transport address for video [8]. For each terminal, the server composes a bitstream tailored to the terminal's needs by aggregating NAL units of appropriate layers. A single RTP session generator is used to aggregate the extracted contents from potentially more than one scalable enhancement layer into a single RTP stream carrying one or more layers [8].

Another network distribution model that can take full advantage of the multilayered characteristics of SVC video includes multicast transport with two different topologies: first, layered multicast of video data to receivers with heterogeneous connectivity, where layers are transported in separate RTP sessions on separate transport addresses [9], and second, multicast on the server side, with a media-aware network element (MANE) to aggregate and/or trim sessions. The NAL units of the aggregated and/or trimmed sessions are conveyed jointly on a single transport address and in a single RTP session as in the case of unicast [6]. The use scenario based on layered multicast is depicted in Fig. 5. A server carries one base layer and multiple enhancement layers, forming a hierarchy. Each SVC video layer is transported in its own IP multicast group identified by its own IP multicast address, and terminals tune into the desired layers by subscribing to the IP multicast group, normally by using Internet group management protocol (IGMP) [10]. This implies that one terminal may have to subscribe to many IP multicast groups for the best possible video quality. However, practical constraints, namely, the existence of network address translations (NATs) and firewalls, make such an approach only feasible in certain academic and research environments. As discussed in [6], practical constraints including NATs and firewalls lead to the concept of MANE, a "middlebox" in the network that aggregates for each client one or more layers into a single RTP stream tailored to the client's requirements as shown in Fig. 6. The MANE is a system that meaningfully manipulates the incoming RTP stream based on information available only in the signaling and in the RTP header, RTP payload, and perhaps the NAL unit header. The basic processing procedure of MANE, shown in Fig. 7, includes RTP depacketization, adaptation on the bitstream level (NAL units) for the target client's requirements by using an adaptation decision taking engine (ADTE) [10], and RTP packetization for the generation of a single RTP session.

To implement the unicast and multicast scenarios transporting

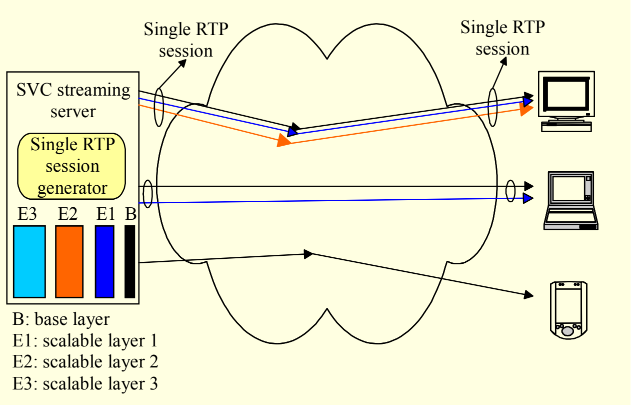


Fig. 4. Unicast transport of SVC video.

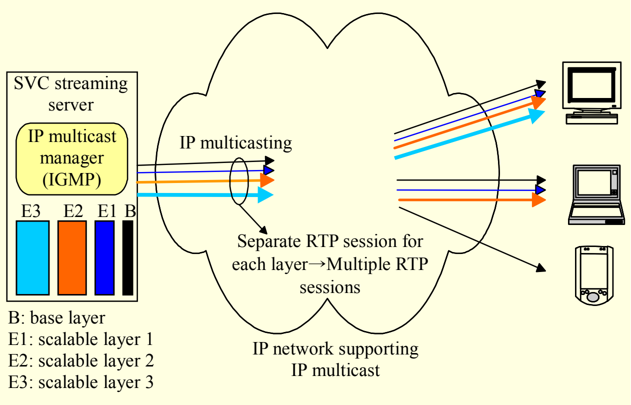

SVC streaming server
IP multicast manager (IGMP)
E3 E2 E1 B
IP multicasting
Separate RTP session for each layer→Multiple RTP sessions
IP network supporting IP multicast
B: base layer
E1: scalable layer 1
E2: scalable layer 2
E3: scalable layer 3

Fig. 5. Layered multicast transport of SVC video.

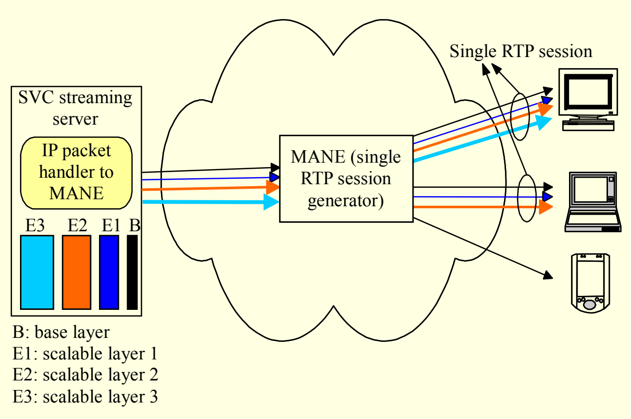


Fig. 6. MANE-based multicast transport of SVC video.

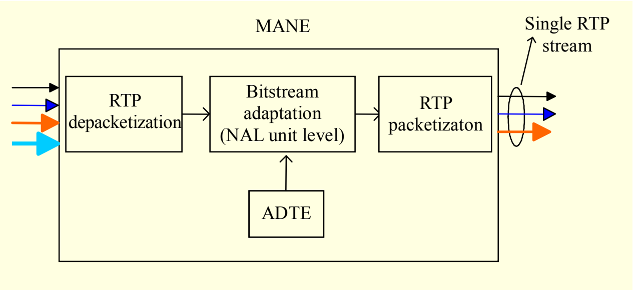


Fig. 7. Basic functional structure of MANE.

SVC video in a single RTP session as required by the server of Fig. 4 and by the MANE of Fig. 6, we need to devise an appropriate RTP packetization scheme. It is particularly necessary to support encapsulating NAL units from multiple SVC layers into a single RTP packet following the standard payload format as specified in [6].

The IETF specification on the RTP payload format for SVC contains four basic mechanisms such as a single NAL unit (SNU), a single-time aggregation packet (STAP), and a multi-time aggregation packet (MTAP) to aggregate more than one NAL unit into a single RTP packet, as well as another mechanism called a fragmentation unit (FU) to split excessively large NAL units into multiple RTP packets [6], [7].

Figure 8 shows the basic principle of forming the four RTP packet types. In Fig. 8, the SNU type can load only one NAL unit (NAL1 or NAL2) in one RTP packet, and the STAP allows encapsulation of more than one NAL unit (NAL1 and NAL2) into one RTP packet that stem from the same picture. This STAP is divided into an STAP-A type that loads NAL units in an RTP packet in the same order as encoding and an STAP-B type that loads NAL units in an RTP packet without considering the encoding order for interleaving purposes. The MTAP can be used to aggregate NAL units (NAL3 and NAL4) from different pictures into one RTP packet, and it basically supports interleaving. This MTAP is divided into an MTAP-16 type supporting a 16-bit time offset and an MTAP-24 type supporting a 24-bit time offset, depending on the size of a time offset field for displaying the difference in the presentation time instants of the NAL units. STAP or MTAP is needed if an NAL unit is much smaller than the maximum transmission unit (MTU). This would result in small RTP packets and cause significant overhead, because the packet header is large in comparison to the transferred payload data. By aggregating several NAL units into a single RTP packet, we can mitigate this problem. When an NAL unit does not fit into a single RTP packet, an FU is used to split the NAL unit (NAL5) into several parts, each fitting into a single RTP packet that does not exceed an MTU size in order to prevent the occurrence of packet fragmentation during transmission.

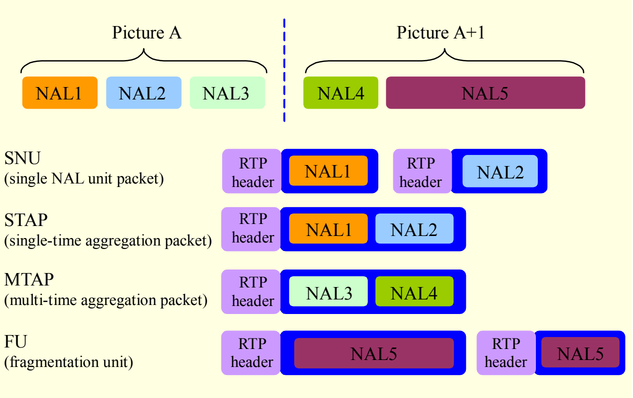


Fig. 8. Basic RTP packet types for SVC.

Table 1. Allowed packet types for RTP packetization modes of SVC.

NAL unit type	Packet type	SNU mode	Non-interleaved mode	Interleaved mode
0	Undefined	Ignore	Ignore	Ignore
1 - 23	SNU	Yes	Yes	No
24	STAP-A	No	Yes	No
25	STAP-B	No	No	Yes
26	MTAP16	No	No	Yes
27	MTAP24	No	No	Yes
28	FU-A	No	Yes	Yes
29	FU-B	No	No	Yes
30 - 31	Undefined	Ignore	Ignore	Ignore

Three fundamentally different packetization modes of operation are supported in [6]: SNU mode, non-interleaved mode, and interleaved mode. Table 1 summarizes the allowed RTP packet types for each packetization mode [6]. The SNU mode is able to support only the SNU type that can load only one NAL unit having 1 to 23 *NAL_unit_types* in an RTP packet, and its application field is restrictive. Thus, the latest Internet-draft document designates that the SNU mode shall not be used for RTP packetization for SVC video [6]. In non-interleaved mode, the NAL units should be aggregated in decoding order by adopting STAP-A, whereas in interleaved mode, NAL units belonging to a picture or multiple pictures can be aggregated out of decoding order by adopting STAP-B and MTAP. Non-interleaved mode is intended to avoid excessive RTP/UDP/IP header overhead that would result when small NAL units are encapsulated in each single RTP packet. On the other hand, interleaved mode provides an error resilience tool against burst errors by shuffling the order of the transmitted data.

STAP-A supported in non-interleaved mode aggregates NAL units with identical NALU times, whereas MTAP supported in interleaved mode aggregates NAL units with different NALU times. Here, NALU time is defined as the value that the RTP timestamp would have if that NAL unit was transported in its own RTP packet. In Fig. 2, pictures belonging to different spatial layers but having the same picture number (or display time) must have the same NALU time. Thus, by adopting STAP-A, it is far more feasible to provide synchronization between pictures belonging to different spatial enhancement layers but with identical NALU times. Therefore, non-interleaved mode is more suitable for systems that require very low end-to-end latency and timely synchronization among NAL units from multiple SVC layers aggregated in an RTP packet. Furthermore, as shown in Table 1, except for the SNU mode, only non-interleaved mode supports the single-NAL-unit type that can contain only a single NAL unit in the RTP payload. As a result, non-interleaved mode can be suggested as a mandatory packetization mode for fast and real-time video transport requiring timely synchronization among SVC layers in a single RTP session, and interleaved mode can be considered as an optional mode for error resilience by providing interleaving function against burst packet loss. Accordingly, we employ non-interleaved mode as a basic RTP packetization mode for the single RTP session generator shown in Figs. 4 and 6.

V. Proposed RTP Packetization Scheme

In this section, we propose a practical non-interleaved mode RTP packetization scheme for generating a single RTP session that is used in the unicast and MANE-based multicast service scenarios shown in Figs. 4 and 6. In the non-interleaved mode for SVC video transport, three RTP packet types such as SNU, FU-A, and STAP-A are supported, as shown in Table 1. Generally, the NAL units belonging to the base layer have a higher importance and priority in transmission than the NAL units belonging to the scalable enhancement layers, and they are processed to be robust against errors through channel coding, separately from the scalable enhancement layer information. Therefore, the NAL units of the base layer are not loaded in an RTP packet by mixing with the NAL units of the scalable enhancement layers. Rather, they are loaded independently in an RTP packet. Accordingly, the STAP-A type is not applied to the NAL units of the base layer; either the SNU or FU-A type is selected and used to load the NAL units in an RTP packet by considering the length of the NAL units. The three packet types (SNU, FU-A, and STAP-A) are all applied to the NAL units belonging to the scalable enhancement layers.

Here, an algorithm based on a look-ahead scheme to identify the scalable enhancement layer NAL units to which the STAP-A type is to be applied will be described. The (*DID, TID, QID*) information of NU_i which is the *i*-th NAL unit being input to the loop of the present algorithm, is indicated by (D_i, T_i, Q_i). The next NAL unit to be analyzed one step in advance by the look-ahead scheme is designated by NU_{i+1}, and (*DID, TID, QID*) information of NU_{i+1} is indicated by (D_{i+1}, T_{i+1}, Q_{i+1}). To determine whether to apply the STAP-A type, (D_{i+1}, T_{i+1}, Q_{i+1}) information of NU_{i+1} is extracted in advance and compared. The sequential conditions that should be satisfied in order to aggregate NU_{i+1} and NU_i and add them to one RTP payload are the following:

i) NU_{i+1} should not be the NAL unit belonging to the base layer.

ii) NU_{i+1} should have the same *TID* value as NU_i.

iii) The sum of the size of the NAL units accumulated until NU_i in an RTP payload plus the size of NU_{i+1} should be smaller than the size of an MTU. In the case of transmitting an RTP packet greater than the MTU, the RTP packet is fragmented into several packets by the fragmentation function of a router or gateway during transmission, thereby causing a burden to the network and the client.

iv) The following conditions should be satisfied depending on the magnitude correlation of Q_i and Q_{i+1}.

(a) If $Q_{i+1}>Q_i$, this means that the quality level of an MGS layer increases. This phenomenon occurs only to the NAL units belonging to the same picture number; thus, the condition of STAP-A is satisfied. Therefore, NU_{i+1} and NU_i can be loaded together in an RTP payload.

(b) If $Q_{i+1}\leq Q_i$, this means that the quality level of an MGS layer does not increase. The situations where this phenomenon occurs can be divided into the situation of $D_{i+1}>D_i$ and vice versa. The situation of $D_{i+1}>D_i$ occurs to the NAL units that always exist within the same picture number; thus, NU_i and NU_{i+1} can be targets of STAP-A. However, the situation of $D_{i+1}=D_i$ occurs between the NAL units having different picture numbers, that is, different presentation time instants; thus, NU_{i+1} and NU_i cannot be targets of STAP-A. The situation of $D_{i+1}<D_i$ never occurs under the condition of $Q_{i+1}\leq Q_i$.

In summary, to perform RTP packetization in the STAP-A type, NU_i and NU_{i+1} should sequentially satisfy all of i, ii, iii, and iv-(a) steps among the above conditions, or should sequentially satisfy all of i, ii, iii, and iv-(b) steps. Figure 9 shows a detailed flowchart of the proposed scheme in order to perform RTP packetization by determining the SNU, FU-A, and STAP-A packet types based on the conditions for determining the packet type as STAP-A. The RTP packet type is determined only based on the *NAL_unit_type*, and the (*DID, TID, QID*) information. The steps i, ii, iii, and iv-(a) and iv-(b), which are the conditions for determining the packet type as

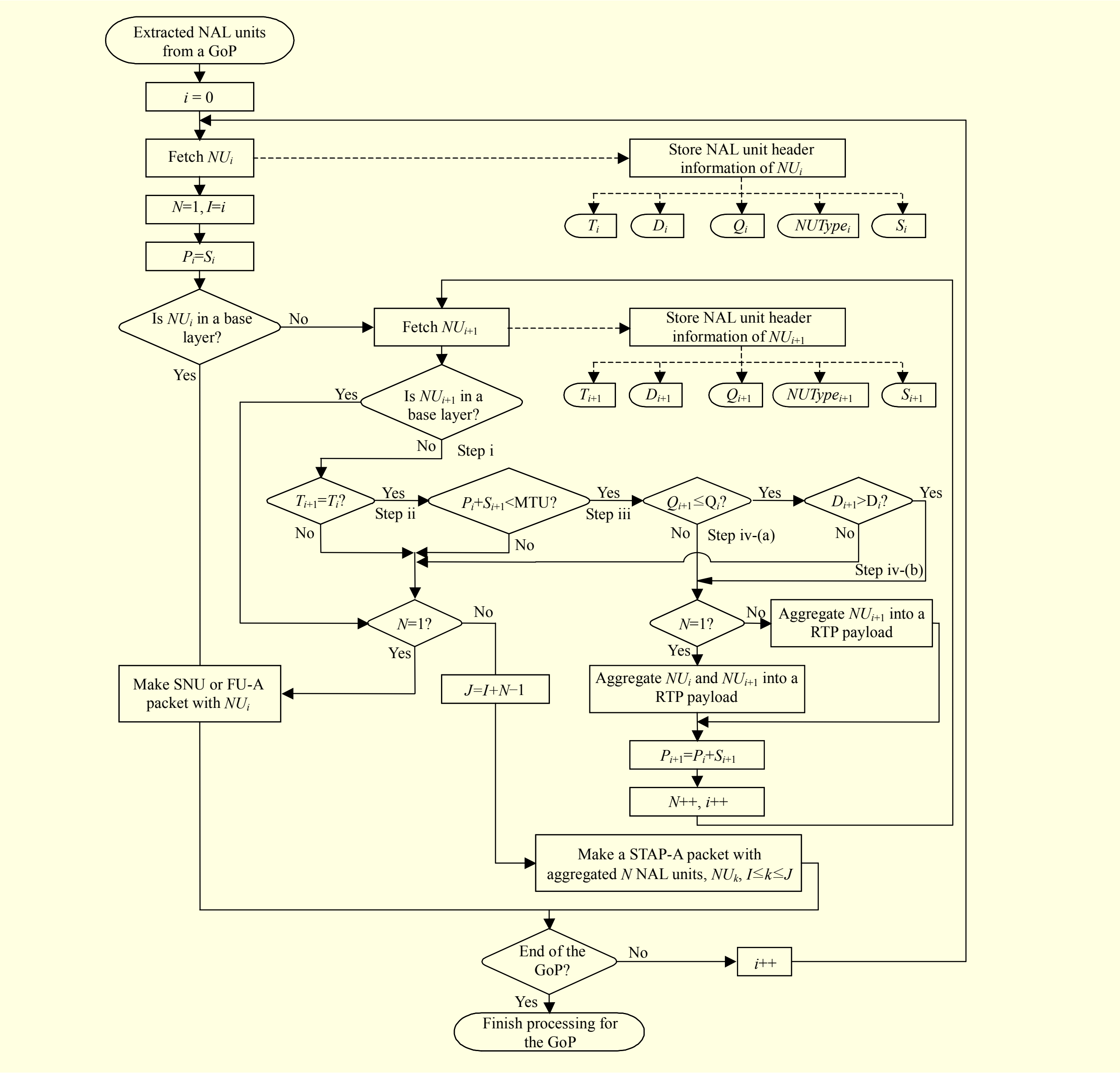

Extracted NAL units from a GoP
$i = 0$
Fetch NU_i
Store NAL unit header information of NU_i
T_i
D_i
Q_i
$NUType_i$
S_i
N=1, I=i
P_i=S_i
Is NU_i in a base layer?
No
Yes
Fetch NU_{i+1}
Store NAL unit header information of NU_{i+1}
T_{i+1}
D_{i+1}
Q_{i+1}
$NUType_{i+1}$
S_{i+1}
Yes
Is NU_{i+1} in a base layer?
No
Step i
T_{i+1}=T_i?
Yes
Step ii
No
P_i+S_{i+1}<MTU?
Yes
Step iii
No
Q_{i+1}≤Q_i?
Yes
No
Step iv-(a)
D_{i+1}>D_i?
Yes
No
Step iv-(b)
N=1?
No
Yes
N=1?
No
Yes
Aggregate NU_{i+1} into a RTP payload
Make SNU or FU-A packet with NU_i
J=I+N−1
Aggregate NU_i and NU_{i+1} into a RTP payload
P_{i+1}=P_i+S_{i+1}
N++, i++
Make a STAP-A packet with aggregated N NAL units, NU_k, I≤k≤J
End of the GoP?
No
i++
Yes
Finish processing for the GoP

Fig. 9. Detailed flowchart of the proposed RTP packetization scheme for non-interleaved mode.

STAP-A, are indicated on the corresponding blocks of Fig. 9, respectively. In Fig. 9, N implies that a packetizing process that is currently in progress is in the process of loading an N-th NAL unit in an RTP payload. Parameters I and J are used to indicate the start position and end position of the N-number of NAL units to be loaded in the RTP payload, S_i means the size of NU_i, and P_i denotes the size of total packets accumulated in the RTP payload including NU_i and is used to check whether or not the size of the total packets accumulated in the RTP payload exceeds that of the MTU.

The algorithm shown in the flowchart is carried out in the look-ahead scheme of investigating NU_{i+1} in advance to compare the STAP-A type condition. Therefore, if the packet type is determined as STAP-A when N=1, NU_i and NU_{i+1} are simultaneously loaded in the RTP payload, whereas if the packet type is determined as STAP-A when N>1, only NU_{i+1} is loaded in the RTP payload. If the packet type is determined as not STAP-A when N=1, the packet type is determined as SNU or FU-A by checking whether the size of NU_i exceeds that of the MTU. On the other hand, if the packet type is determined not to be STAP-A when N>1, N NAL units accumulated in an RTP payload up to the present are loaded and transmitted in one RTP packet. Then parameters N and I are updated to generate a new RTP packet, and this is followed by repeating the entire process.

Based on this process, the computational complexity of the proposed method could be claimed to be considerably low, in that it requires only the knowledge of *NAL_unit_type* and (*DID, TID, QID*) information which is directly available from the header extension part of the SVC NAL unit and some arithmetic comparison as shown in Fig. 9. The effect of the proposed non-interleaved mode packetization could be reflected as a reduction of the generated number of RTP packets, which thereby reduces the bit-rate. This effect is due to the efficient encapsulation of NAL units preserving the standard RTP payload format for SVC. By the proposed packetization method, we can reduce the generated number of packets by employing STAP-A. This reduces the overhead data volume required for building a UDP header, an IP header, and an RTP header.

To extend the proposed method to accommodate interleaved mode packetization supporting MTAP and STAP-B, some additional processing time for interleaving the order of NAL units is required during packetization. This corresponds to additional time overhead incurred during packetization. Moreover, the interleaved mode requires additional fields called decoding order number and time-offset in the RTP payload part for providing sampling time and decoding order information of NAL units.

VI. Experimental Results

To verify the effectiveness of the proposed non-interleaved mode RTP packetization, we implemented an SVC-based streaming system. For unicast service, as shown in Fig. 4, the extraction based on ADTE is performed at the server side, while it is performed at the MANE side for multicast case as shown in Fig. 6. More details on ADTE, which is beyond the scope of this paper, can be found in [11]. In the experiment, the proposed algorithm shown in Fig. 9 was employed for non-interleaved mode RTP packetization. To show the effectiveness of the proposed scheme, the following metrics were used:

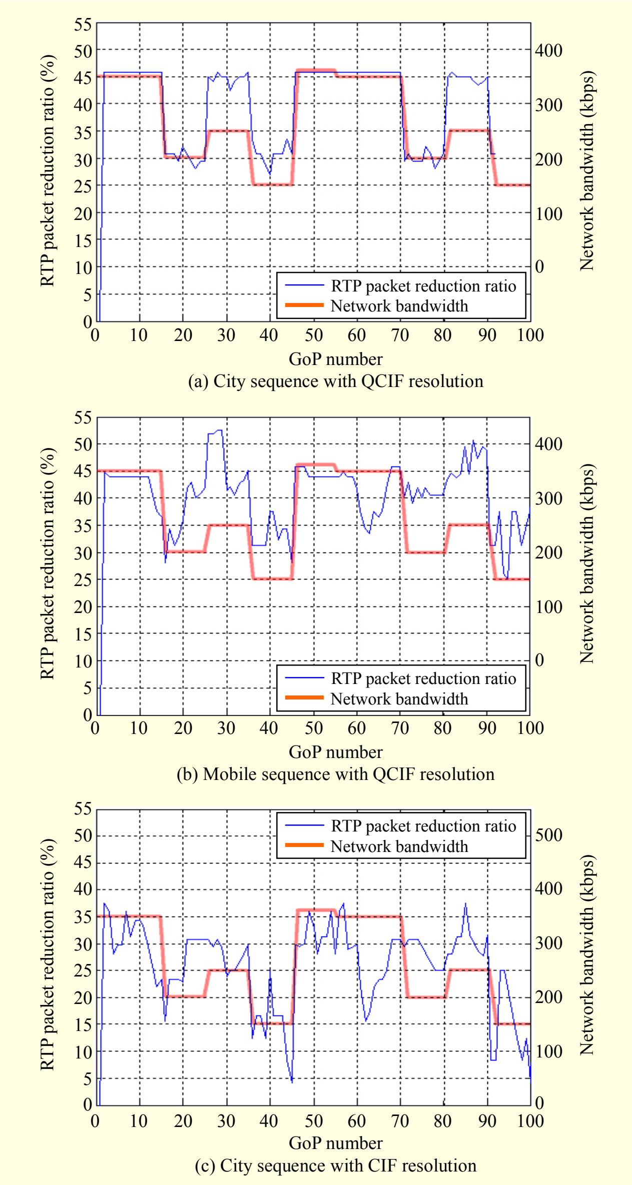
- number of RTP packets reduced by applying the proposed non-interleaved mode RTP packetization,
- bit-rate reduction ratio obtained by applying the proposed non-interleaved mode RTP packetization,
- transmission delays between server and clients, to a large degree determined by the delays incurred by MANE,
- load on and scalability of MANE in terms of CPU usage.

Simulations for evaluating the first two metrics were carried out using three test sequences: *City* with QCIF resolution, *Mobile* with QCIF resolution, and *City* with CIF resolution. The sequences were encoded using the Joint Scalable Video Model (JSVM) [2] software following the same GoP structure shown in Fig. 2. First, we compared the number of RTP packets generated by employing both SNU and FU-A types with the number generated by additionally applying STAP-A type to observe the significant reduction effect in the number of RTP packets by employing STAP-A type for non-interleaved mode RTP packetization. Accordingly, the following equation was used to evaluate the packet reduction ratio, P_r, achieved by STAP-A for each GoP:

$$P_r = \frac{N_{\text{SNU,FU}} - N_{\text{STAP}}}{N_{\text{SNU,FU}}} \times 100\ (\%), \qquad (1)$$

where $N_{\text{SNU,FU}}$ denotes the number of RTP packets generated per GoP when both the SNU and FU-A types are employed, and N_{STAP} designates the number generated by additionally applying the STAP-A type. Figure 10 shows the simulation results on the packet reduction ratio when the available network bandwidth varies over time in an IP network with an MTU size of 1,500 bytes. In Fig. 10, the right-hand vertical axis represents the available network bandwidth for each GoP, and the left-hand vertical axis shows the packet reduction ratio for each GoP evaluated by (1). We use a series of predetermined values to emulate the monitored network bandwidth as marked in red in Fig. 10. From the simulation results, we can observe significant reduction in the number of RTP packets ranging from 25% to 50% for QCIF resolution sequences and from 10% to 35% for CIF resolution sequence. We can also notice that the packet reduction ratio is affected by the given network bandwidth since the number of NAL units that can be aggregated in an RTP packet by the STAP-A type highly depends on the number of scalable enhancement layers pertaining to the extraction point (*DID, TID, QID*) determined by the extractor. In general, we can usually observe higher packet reduction ratios at high network bandwidths. For higher network bandwidths, a larger number of NAL units can be aggregated to a single RTP packet. On the other hand, the result in Fig. 10(c) shows a smaller packet reduction ratio when compared to Figs. 10(a) and (b). The reason for this is that spatial layer 1 in the case of Fig. 10(c) contains NAL units that correspond to a CIF resolution, unlike the cases of Figs. 10(a) and (b) which employ a CGS layer as their spatial layer 1. Thus, the NAL unit size in spatial layer 1 of Fig. 10(c) is usually much larger than those of Figs. 10(a) and (b), thereby resulting in a smaller number of NAL units that can be aggregated to a single RTP packet by STAP-A type considering the MTU size constraint of the network.

Figure 11 shows the simulation results on the bit-rate reduction effect achieved by employing the proposed non-interleaved mode RTP packetization. The simulations were carried out on the three test sequences originally used in Fig. 10

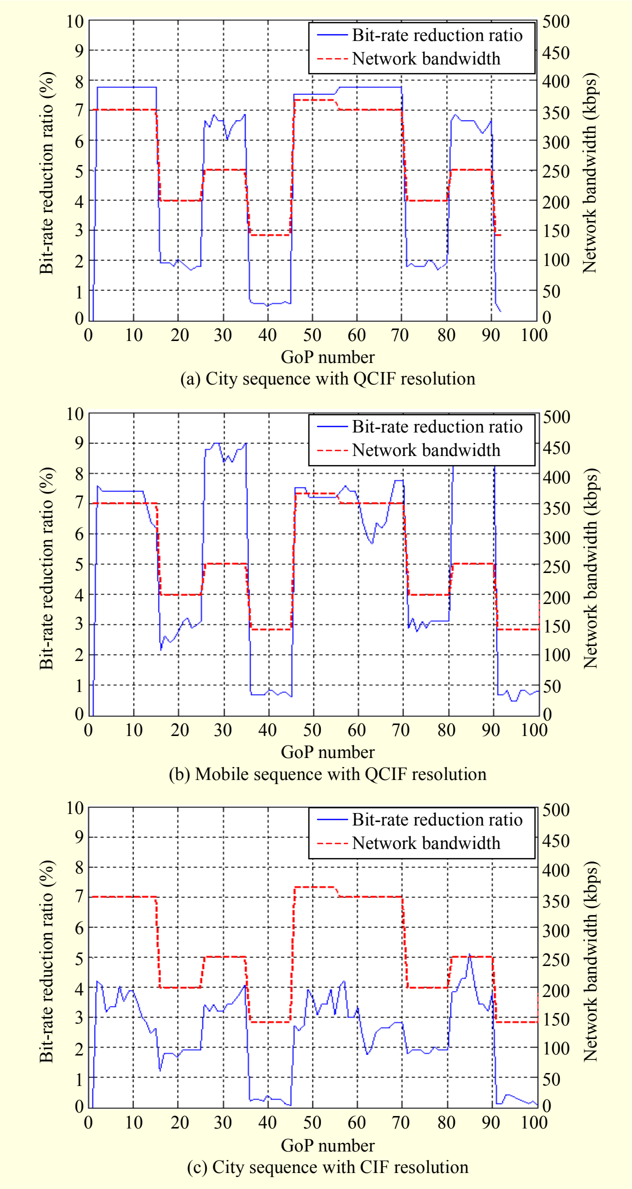


(a) City sequence with QCIF resolution

(b) Mobile sequence with QCIF resolution

(c) City sequence with CIF resolution

Fig. 10. Simulation results on the packet reduction ratio for the three test sequences.



(a) City sequence with QCIF resolution

(b) Mobile sequence with QCIF resolution

(c) City sequence with CIF resolution

Fig. 11. Simulation results on the bit-rate reduction ratio for the three test sequences.

under the same simulation conditions. In Fig. 11, the left-hand vertical axis represents the bit-rate reduction ratios for each GoP. The bit-rate reduction effect results from the reduced number of RTP packets by applying the STAP-A type. For each saved RTP packet resulting from application of the STAP-A type, we do not need to build a UDP header and an IP header in addition to an RTP header. Accordingly, the total number of octets we can reduce for each saved RTP packet reaches up to 40 octets considering the basic RTP header size of 12 octets, UDP header size of 8 octets, and IP header size of 20 octets. The bit-rate reduction ratio in Fig. 11 corresponds to the ratio of the reduced number of octets achieved by additionally employing STAP-A to the total number of octets generated by not employing STAP-A type (that is, only applying SNU and FU-A types) for each GoP. As seen in Fig. 11(a) and (b), the bit-rate reduction ratio reaches up to around 10%. However, the maximum reduction ratio in Fig. 11(c) is reduced by half to reach up to 4% to 5%. This observation can be anticipated from the results of Fig. 10 where we noticed higher packet reduction ratios in Fig. 10(a) and (b) than in Fig. 10(c).

For the case of MANE-based multicast, a number of clients could be simultaneously connected to the MANE to receive proper video service. Therefore, the delay time and CPU usage required for processing incoming RTP packets at the MANE are important aspects to evaluate the proposed RTP packetization

Table 2. Video sequences used for evaluation of delay time and CPU usage.

Video sequence	Base layer	First spatial enh. layer (quality enh.)	Second spatial enh. layer (quality enh.)	Bit-rate (kbps)
City1	QCIF@15Hz (3 MGS)	CIF@30Hz (3 MGS)	None	652
Mobile1	QCIF@15Hz (3 MGS)	CIF@30Hz (3 MGS)	None	735
Foreman1	QCIF@15Hz (3 MGS)	CIF@30Hz (3 MGS)	None	843
City2	CIF@30Hz (3 MGS)	4CIF@30Hz (3 MGS)	None	1,287
Mobile2	CIF@30Hz (3 MGS)	4CIF@30Hz (3 MGS)	None	1,452
Foreman2	CIF@30Hz (3 MGS)	4CIF@30Hz (3 MGS)	None	1,609
City3	QCIF@15Hz (1 MGS)	CIF@30Hz (2 MGS)	4CIF@30Hz (3 MGS)	2,081
Mobile3	QCIF@15Hz (1 MGS)	CIF@30Hz (2 MGS)	4CIF@30Hz (3 MGS)	2,242
Foreman3	QCIF@15Hz (1 MGS)	CIF@30Hz (2 MGS)	4CIF@30Hz (3 MGS)	2,575

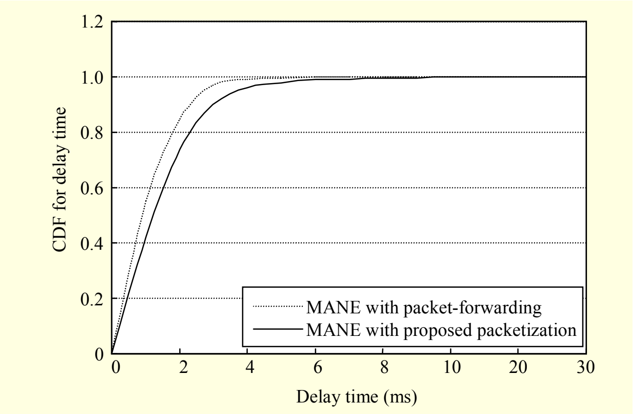


Fig. 12. Cumulative distribution function (CDF) for delay time of City1 sequence.

scheme. In addition to the MANE-based multicast employing the proposed RTP packetization scheme, a simple packet-forwarding MANE was implemented to serve as a reference. This packet forwarder simply takes the RTP payload of an incoming packet and fills this payload into an outgoing packet.

The following quantitative evaluation shows the impact of in-network adaptation and packetization on the transmission delay. The video sequences used for the evaluations and their layer configurations are described in Table 2.

We compared MANE with simple packet-forwarding and MANE with proposed packetization algorithm. For all measurements, 20 clients were receiving the same content from the streaming server via the two MANEs. To compare the processing delay time of the two MANEs, a representative video sequence was selected from Table 2, which is City1. Figure 12 shows the induced delay for the City1 sequence with an overall bit-rate of about 13 Mbps for all 20 clients. By comparing the two MANEs, it becomes evident that the additional processing effort due to the proposed packetization algorithm is very low. CPU usage results shown in Fig. 13 show the proposed packetization induces a little more computational load and a significant number of parallel streams (20 connections) can be handled in real-time.

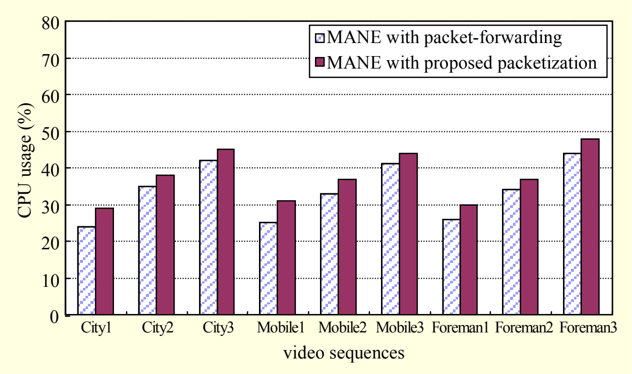


Fig. 13. Comparison of average CPU usage.

VII. Conclusion

In this paper, we discussed the suitability of non-interleaved mode RTP packetization to provide very low end-to-end latency and timely layer synchronization among scalable enhancement layers of SVC video. Then, we proposed a practical non-interleaved mode RTP packetization scheme for SVC video unicast/multicast transport over IP networks. With the proposed packetization, we could observe a significant packet reduction ratio ranging from 10% to 50%, depending on the resolution of the test sequences and given network bandwidth. Also, the amount of additional delay time and CPU usage caused by the proposed packetization can be regarded as manageable. Considering that no research has provided a result yet on a concrete RTP packetization scheme for SVC video, the proposed method can be a practical solution in real implementation of RTP packetization to transport SVC video over IP networks.

References

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,"

IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 9, Sep. 2007, pp. 1103-1120.

[2] J. Reichel, H. Schwarz, and M. Wien, "Joint Scalable Video Model (JSVM)," *Joint Video Team*, Doc. *JVT-X202*, Geneva, Switzerland, July 2007.

[3] J. Kang et al., "Development of QoS-Aware Ubiquitous Content Access (UCA) Testbed," *IEEE Trans. Consum. Electron.*, vol. 53, no. 1, Feb. 2007, pp. 197-203.

[4] X. Wu, S. Cheng, and Z. Xiong, "On Packetization of Embedded Multimedia Bitstreams," *IEEE Trans. Multimedia*, vol. 3, no. 1, Mar. 2001, pp. 132-140.

[5] H. Schulzrinne, S. Casner, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," *IETF RFC 3550*, *STD 64*, July 2003.

[6] S. Wenger, Y. Wang, and T. Schierl, "RTP Payload Format for SVC Video," IETF Internet Draft: *draft-ietf-avt-rtp-svc-18.txt,* Mar. 2009.

[7] S. Wenger, Y. Wang, and T. Schierl, "Transport and Signaling of SVC in IP Networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, Sep. 2007, pp. 1164-1173.

[8] K. Seo et al., "Media Synchronization Framework for SVC Video Transport over IP Networks," *Int. Conf. Computational Science*, LNCS, vol. 4490, May 2007, pp. 621-628.

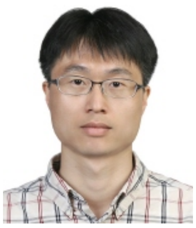
[9] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-Driven Layered Multicast," in *Proc. ACM SIGCOMM*, Stanford, CA, Aug. 1996, pp. 117-130.

[10] B. Cain et al., "Internet Group Management Protocol, Version 3," *IETF RFC 3376*, Oct. 2002.

[11] H. Choi, J. Kang, and J. Kim, "Dynamic and Interoperable Adaptation of SVC for QoS-Enabled Streaming," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, May 2007, pp. 384-389.

Kwang-deok Seo received the BS, MS, and PhD degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1996, 1998, and 2002, respectively. From Aug. 2002 to Feb. 2005, he was with LG Electronics. Since March 2005, he has been a faculty member in the Computer and Telecommunications Engineering Division, Yonsei University, Gangwon, Korea, where he is an associate professor. His current research interests include digital video broadcasting, mobile IPTV, scalable video coding, and protocol design for scalable video transport. He is a member of KICS, IEEE, and IEICE.

Jin-soo Kim received the BS degree in electronics engineering from Kyungpook National University, Daegu, Korea, in 1991, and the MS and PhD degrees in electrical engineering from KAIST, Korea, in 1993 and 1998, respectively. From 1998 to 2000, he was with the Business Division of System LSI at Samsung Electronics, where he was involved in the development of MCU chipsets. Since March 2000, he has been a faculty member in the School of Information Communication and Computer Engineering, Hanbat National University, Korea, where he is an associate professor. His research interests include scalable video coding, networked video-rate shaping and adaptation, media convergence, and multimedia re-multiplexing.



Soon-heung Jung received the BS degree in electronic engineering in 2001 from Pusan National University, Pusan, Korea. He received the MS degree in electronic engineering in 2003 from KAIST, Daejeon, Korea. From 2003 to 2005, he was a research engineer with LG Electronics, Korea. Since 2005, he has been a senior member of engineering staff in the IPTV Research Department of ETRI, Korea. His research interests are in the areas of visual communications, video signal processing, video coding, and digital broadcasting.

Jeong-ju Yoo received the BS and MS degrees in telecommunications in 1982 and 1984, respectively, from Kwangwoon University, Seoul, Korea. He received the PhD degree in computing science from Lancaster University, United Kingdom in 2001. Since 1984, he has been a principal member of technical staff in the IPTV Research Technology Department of ETRI, Korea. He has been a team leader of IP Broadcasting Media Research Team at ETRI from 2007 and he was a Head of MPEG Korea delgates from 2007 to 2009. His research interests are in the area of QoS, video coding, and IPTV.