# Sage's Video Quality Monitor

Renshou Dai

February 13, 2008

## 1 Introduction

Sage's Video Quality Monitor (SVQM) analyzes the RTP video packets that traverse an IP network and then derives a set of measurement reports that can be used to predict, with high correlation, the end-user-perceived video quality.

SVQM is implemented in Sage's 96x platform with Ethernet interfaces. It is a sub-feature under the "RTP MONITOR" test. With this new addition, a user now has 3 sub-features under this "RTP MONITOR", namely: INMD (In-service Non-intrusive Measurement Device per ITU-T P.561 [1]), Dual-spectrum analyzer and this SVQM. Using Sage's "RTP MONITOR", a user can target a specific pair of RTP streams by specifying the RTP-socket attributes such as source and destination IP address and UDP port number. Or, if left unspecified, the "RTP MONITOR" will seize the first pair of RTP streams that it sees. Stream pairing is based on the symmetricity property of the RTP-socket attributes. More specifically, one stream's destination IP address and port number are the other stream's source IP address and port number, and vice versa. "RTP MONITOR" also allows a user to specify an arbitrary jitter buffer size in unit of milli-second to simulate the true end-terminal's packet-discarding effect due to excessive network jitter.

In order for SVQM to work, the internal algorithm has to determine if an RTP stream is carrying video or voice. Without *a priori* information on a video stream's payload type, SVQM relies on the differential time-stamp detection algorithm that will guarantee SVQM to lock onto a pair of video streams, and ignore the voice streams.

SVQM applies to the RTP-transport of H.264-encoded [2] video slices based on RFC-3984 [3]. RFC 3984 is designed for conversational video applications such as IP video telephony, video conferencing and telepresence, Internet video streaming, video-on-demand and IPTV etc.

The need for SVQM is clear. Simply monitoring and collecting the RTP statistics are not useful enough for predicting the video quality. For H.264-encoded video stream in particular, not every packet is equal. Loss of certain packets are "fatal" to the video quality; yet the loss of some other packets does not pose visible effects at all. The ITU-T G.107-based E-model [4] was designed only for voice stream, and it can not be adapted for video application.

SVQM works by evaluating each packet loss's true impact to the structural integrity of the H.264-encoded video sequence. SVQM's value lies in the fact that it is NOT just a simple collection of RTP statistics. SVQM simulates a true video-terminal by dejittering the incoming RTP packets, re-ordering the potentially-out-of-order packets, discarding the packets whose delay exceed the user-specified jitter buffer size, extracting the NAL (Network Abstraction Layer [2]) unit(s) from each RTP packet, integrating a group of NAL units into an Access Unit [3] (a single video frame, simply speaking), and further integrating a group of Access Units into a Coded Video Sequence. When an RTP packet loss occurs, either due to true network loss or as a result of the delay-and-discarding by the jitter buffer, the RTP loss's true impact is carefully evaluated to see if it only affects a single vide frame, or if it will affect a whole video sequence, and its true impact will be accumulated and normalized and scaled to provide a video quality MOS (Mean-Opinion-Score) number. Details are explained in the following sections.

# 2 Background information and principles

The essential concepts and terminologies from RFC 3984 and H.264 are summarized here in order for you to understand the algorithm behind SVQM.

## 2.1 VCL and NAL

The H.264 codec distinguishes conceptually between a video coding layer (VCL) and a network abstraction layer (NAL). The VCL contains the signal processing functionality of the codec. It follows the general concept of most of today's video codecs, a macroblock-based coder that uses inter picture prediction with motion compensation and transform coding of the residual signal. The VCL encoder outputs slices: a bit string that contains video data for the decoder.

## 2.2 NAL unit, its relationship with RTP packet and its type

The Network Abstraction Layer (NAL) encoder encapsulates the slice output of the VCL encoder into Network Abstraction Layer Units (NAL units), which are suitable for transmission over packet networks or use in packet oriented multiplex environments. A single NAL unit can occupy a single RTP payload, or multiple NAL units can occupy a single RTP's payload in aggregation mode, or a single NAL unit can be segregated into multiple RTP packet's payloads in segregation mode.

Each NAL unit begins with a single-byte header. The header contains a 5-bit "type" field that indicates what type of NAL unit this is. The interpretation of this "type" field is described in Table 1 of RFC 3984 [3] and Table 7-1 of H.264 [2].

## 2.3 Access Unit and video frame

A set of NAL units forms an Access Unit that always contains a primary coded picture. The decoding of an Access Unit always results in a decoded picture. All RTP packets that carry the NAL units within a single Access Unit have the same time stamp. The end of an Access Unit is marked by the "marker bit" inside the RTP packet header. The beginning of a new Access Unit is detected by the sudden jump of time-stamp between two consecutive RTP packets (sequence numbers differ by 1). Inside SVQM, an Access Unit is treated as a single video frame. The inverse of the time-stamp difference between adjacent Access Units (video frames) multiplied by 90KHz sampling rate gives the instantaneous frame rate (number of frames per second). Within a single Access Unit, when one or more NAL units are lost or delayed and discarded, the whole video frame (Access Unit) is considered damaged.

## 2.4 Coded Video Sequence, reference and predicted video frames

A Coded Video Sequence is a sequence of Access Units that consists, in decoding order, of an instantaneous decoding refresh (IDR) access unit followed by zero or more non-IDR access units including all subsequent access units up to but not including any subsequent IDR access unit.

In simpler terms, a Coded Video Sequence starts with a reference video frame (IDR access unit), followed by a sequence of predicted video frames (non-IDR access units). The reference video frame causes the receiving terminal to refresh the whole background picture, and all the following video frames are predicted based solely on this reference video frame. So, the reference video frame is vitally important. If the reference video frame is damaged, the whole Coded Video Sequence will be considered damaged by SVQM. Perceptually, damage to the reference video frame will cause the whole background picture to be either corrupted or frozen. Damage to a predicted video frame only corrupts the "small" moving portion of the picture.

## 2.5 Parameter Set NAL units

For each coded video sequence, the H.264 encoder also generates two parameter sets: sequence parameter set and picture parameter set. These parameter sets are generally also sent in-band with NAL unit encapsulation and transported through RTP. In such case, the NAL units containing these parameter sets are also considered part of the reference video frame (IDR Access Unit). Therefor, damage to any of the parameter sets affect the whole coded video sequence.

# 3 Definitions of SVQM measurement results

## 3.1 RTP MONITOR results

As mentioned before, the SVQM sub-feature works under the "RTP MONITOR" test, so SVQM automatically inherits the statistics reported by "RTP MONITOR". Descriptions of these statistics are as follows:

1. RTP source port number: indicates the UDP source port number from which this RTP packet is coming.
2. RTP source address: indicates the source IP address from which this RTP packet is coming.
3. Total RTP packet count: indicates how many RTP packets should have been received so far based on continuous tracking of the RTP sequence numbers.
4. Actually received RTP packets: indicates how many RTP packets have actually been received so far. The difference between "total count" and this "actual count" gives the RTP packet loss seen at this moment.
5. Out of order packet count: indicates how many RTP packets have arrived out of order.
6. Delayed and discarded packet count: indicates how many RTP packets have arrived later than what the user-specified jitter buffer can tolerate, therefor they have to be discarded and the voice/video decoding layer will detect packet losses.
7. Inter Arrival Jitter in $\mu$s (micro-second): a metric based on the algorithm provided in section 6.4.1 of RFC 3550 [5]. But notice that all RTP packets that carry NAL units within a single Access Unit all have the same time stamp. So, for video packets, this "Inter Arrival Jitter" has to be modified. It is calculated only at the boundary between two adjacent Access Units, where a time stamp jump occurs. For voice stream, it is calculated for each packet.
8. RTP payload or vocoder type: most voice codecs have static mapping as shown in RFC 3551 [6]. For example, PCMU maps to payload type 0; PCMA maps to type 8 and G729 maps to payload type 18. For video packets, the payload type is dynamically assigned. The plain payload type number (such as 99) is shown here to the user.
9. RTP packet arrival jitter: this jitter (in $\mu$s) indicates an RTP packet's arrival time delay from the theoretical time when it should arrive. If such jitter is greater than the jitter buffer size, then this packet will be declared "delayed and discarded".
10. RTP packet's ToS field: ToS (Type-of-Service) field inside the IP header has been used by some network devices as a priority indicator.

## 3.2 Additional SVQM results

In addition, the SVQM offers the following measurement results that are more relevant to the actual user-perceived video quality:

1. Classified NAL unit count: this is a 32-element array of counters. The "type" field of each received NAL unit will be extracted, and the related counter indexed by this "type" value will be incremented by 1. The "type" is a 5-bit field, hence the 32-element array. The definition of the type field is shown in Table 7.1 of H.264 [2] and Table 1 of RFC 3984 [3]. The most relevant type values are: type=1, non-IDR NAL unit; type=5, an IDR NAL unit; type=7, sequence parameter set NAL unit; type=8, picture parameter set NAL unit.

2. Total NAL unit loss: an accumulative counter that records how many NAL units of all types have been lost, including those delayed-and-discarded NAL units.

3. Total video frame count: an accumulative counter that records how many Access Units (video frames) have been received so far.

4. Damaged video frame count: an accumulative counter that records how many video frames have been damaged. When at least one "constituent" NAL unit inside an Access Unit (video frame) is lost, this video frame is considered damaged, and this counter will be incremented by 1.

5. Coded video sequence count: an accumulative counter that records how many coded video sequences have been received so far.

6. Damaged coded video sequence count: an accumulative counter that records how many coded video sequences have been damaged. A video sequence is declared damaged if the leading IDR (reference) video frame is damaged.

7. Effective damaged video frame count: an accumulative counter that records the effective number of damaged video frames. This "effective counter" is a superset (a combination) of the damaged video frame counter and the damaged video sequence counter. When a damaged video sequence is detected, all video frames within this sequence are considered damaged and accumulated onto this "effective counter". This "effective counter" is the best predictor of video quality perceived by the end-users.

8. Frame rate: indicating how many video frames are detected per second.

9. Video MOS: a number between 1 and 5. This is directly mapped from the ratio between "effective damaged video frame count" and "total video frame count". MOS stands for Mean-Opinion-Score.

10. MDI_DF and MDI_MLF: these two fields are based on RFC 4445 [7], a proposed Media Delivery Index (MDI). This standard is targeted at broadcast TV using MPEG Transport Stream (TS) packets over UDP where a constant "drain rate" can be assumed. For conversational video applications, this MDI is not exactly applicable. So the two fields presented

here are SVQM's best effort approximation. MDI_DF (MDI delay factor) is simply a copy of the "inter-arrival jitter" reported under "RTP MONITOR". MDI_MLF (MDI Media Loss Factor) is the ratio of "effective damaged video frame count" divided by "total video frame count".

# References


[1] "In-service non-intrusive measurement device-Voice service measurements," *ITU-T Recommendation P.561*, July, 2002.

[2] "Advanced video coding for generic audiovisual services," *ITU-T Recommendation H.264*, March, 2005.

[3] "RTP payload format for H.264 video," *IETF RFC 3984*, February, 2005.

[4] "The E-model, a computational model for use in transmission planning," *ITU-T Recommendation G.107*, March, 2005.

[5] "RTP: A Transport Protocol for Real-Time Applications," *IETF RFC 3550*, July, 2003.

[6] "RTP Profile for Audio and Video Conferences with Minimal Control," *IETF RFC 3551*, July, 2003.

[7] "A Proposed Media Delivery Index (MDI)," *IETF RFC 4445*, April, 2006.