A Latency Analysis on H.264 Video Transmission Systems

Ralf M. Schreier [1], *Member, IEEE*, Albrecht Rothermel [2], *Senior Member, IEEE*
[1] *Vienna University of Technology, Austria,* [2]*University of Ulm, Germany*
email: schreier@ims.tuwien.ac.at

***Abstract*-- Controlling and minimizing the delay of real-time video transmission systems is a key issue in latency sensitive applications. Examples for such kind of systems include classical video conferencing applications covered by H.263 implementations, and remote controls or RF studio cameras which require much lower delays. This document gives an overview of the delay sources in compressed video transmission systems and discusses algorithmic and buffer delay effects of different CBR compression modes.**

I. INTRODUCTION

Designing systems for low delay video transmission requires basic implementation knowledge as well as fundamental knowledge of the video codec algorithms and buffer management. As the standards only describe the algorithmic decoding procedure and profile features, the system designer has many options in selecting coding modes, coding parameters and buffer sizes to meet the requirements of the specific target application. It is the intention of this paper to give guidelines for the delay estimation of complete video transmission systems considering simulation results of H.264 video test sequences.

Concerning low-delay applications, the available literature mainly discusses network, rate control and frame-skipping [1], [2] ,[3] optimizations with a focus on video telephony or variable bit rate (VBR) streaming applications. Our analysis does not implement and evaluate a specific rate control or frame-skipping algorithm. In this paper, we evaluate the inherent system behavior without influence of rate control to determine the bounds and requirements for video transmission with minimum quality impacts in constant bitrate (CBR) transmission mode.

Selecting the coding parameters for low-delay applications is a trade-off between coding efficiency, limitations on rate fluctuations, re-synchronization capability and computational requirements. The following analysis includes the relevant coding modes for broadcast or point-to-point applications with a guaranteed re-synchronization time. For achieving low buffer delays at CBR it is generally favorable to generate a constant amount of data in a short time interval with minimum intrusion of a rate control algorithm. We denote this time interval as frame CBR or GOP-CBR according to the number of frames which are used for averaging rate variations

II. CODING MODE DELAY ANALYSIS

A. System Delays

The components which are critical with respect to the delay of a video compression system are illustrated in Fig. 1. According to this model, the overall system delay can be calculated by:

$$D_{sys} = D_{cap} + D_{reorder} + D^{e}_{proc} + D^{e}_{buff} + D_{net} + D^{d}_{buff} + D^{d}_{proc} + D_{out}\,.$$

Throughout this document all delays are counted in multiples of frame periods T_{frame}. Implementation-specific delays mainly depend on the selection of the basic processing block size and the pipelining strategy of the encoder/decoder. To our knowledge the reference codecs as well as many software based implementations use a frame-based processing scheme which introduces large delays. On the other hand, hardware-optimized ASIC/FPGA implementations frequently use a macroblock pipeline which reduces the on-chip memory requirements as well as delays. Due to the nature of continuous time video sampling and system load balancing, we propose to use a row of macroblocks similar to the MPEG-2 slice size as the basic processing block [4]. This results in small capture and processing delays of $D_{cap} = D_{proc} = T_{slice} \approx 0.05\ T_{frame}$.

The encoder and decoder buffer delays can be analyzed with the leaky bucket model proposed by [1], [5]. In contrast to other publications, we performed the buffer analysis on slice level rather than on frame level. Taking into account the rate fluctuations within a frame has the potential for significant reductions of buffer sizes, but it can also result in quality losses in cases when the video material is more critical than video sequences which were considered for buffer adjustment.

It should be noted that real-time video compression systems

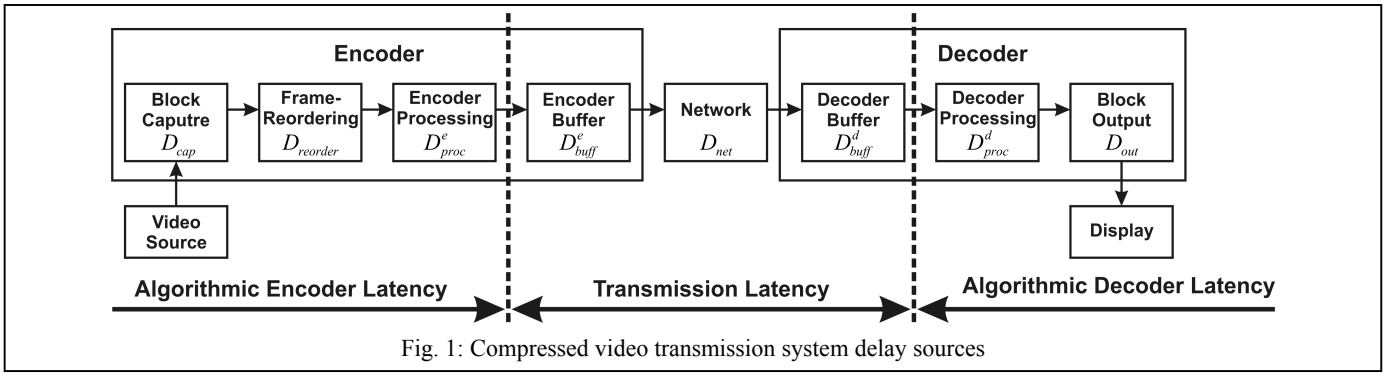

Fig. 1: Compressed video transmission system delay sources

usually operate in a constant delay mode. Hence, the algorithmic as well as the buffer delays are adjusted according to worst-case operating conditions and it is not possible to adjust the delay during continuous operation.

B. Intra Coding Mode

The intra coding mode can achieve very low delays at the cost of a low coding efficiency. D_{cap}, D^{e}_{proc} and D^{d}_{proc} can be reduced to T_{slice}. As the intra coding mode generates approximately the same amount of data for each frame it is a safe assumption to introduce a one T_{frame} buffer delay for the encoder and decoder buffers which allows unrestricted distribution of data within each frame. Further reduction of the buffer delays is possible if the rate control algorithm limits the rate variations within a frame to reasonable bounds. Video sequences with very unbalanced vertical texture (e.g. "flowergarden") require the largest encoder and decoder buffer delays up to $D^{e}_{buff} = D^{d}_{buff} \approx 0.4\,T_{frame}$.

C. IP Coding Mode

The predictive coding mode can achieve the same processing and capture delays as the intra coding mode but the minimum buffer delay is defined by the size ratio of the I-frame compared with the P-frames in the GOP. In order to achieve full image quality at a constant bit rate, the additional data rate of the I-frame is averaged out over following P-frames resulting in a CBR interval of one GOP. The resulting decoding buffer delay can be calculated by

$$D_{buff,IP} = \frac{1}{1+(N_{GOP}-1)\cdot p\_ratio}\cdot N_{GOP}\cdot T_{frame}, \text{ with}$$

N_{GOP} = number of frames in GOP

p_ratio = average size ratio of predicted frames vs. intracoded frames.

For the delay adjustment of this coding mode, reasonable buffer limits can be derived from the H.264 simulation results given in Fig. 2. Especially in low-bitrate applications the data rate is concentrated in the I-frames, resulting in a *p_ratio* below 0.1 and delays above 6 T_{frame} (N_{GOP} = 12). A realistic lower bound for the encoder buffer is 0.25 T_{frame}.

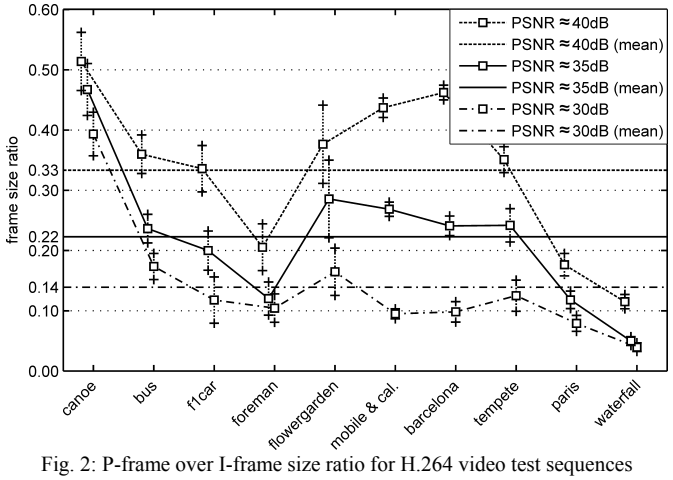

Fig. 2: P-frame over I-frame size ratio for H.264 video test sequences (CIF, no deblocking filter)

D. IPB Coding Mode

The delay analysis of the IPB coding mode follows the IP coding mode analysis, with an additional frame reordering delay of 2 T_{frame} (IBP-mode), and 3 T_{frame} (IBBP-mode). The buffer delays are quite similar to the IP coding mode. These characteristics prohibit using the IPB coding modes in delay sensitive applications.

E. Frame-CBR Intra Refresh Coding Mode

A region-based intra refresh method as proposed in [6] can reduce the buffer delays effectively while maintaining the coding efficiency of the IP coding scheme. The intra refresh can be combined with frame or GOP-level CBR operation. The frame CBR operation can achieve smaller delays but 9 % to 14 % additional data rate must be allocated on the average for large frames with high amounts of intra texture. This is caused by the fact that the amount of intra information can vary significantly within a refresh cycle. For some critical video sequences (e.g. "foreman", "flowergarden"), frames with more than 30 % overhead in data rate were observed. Realistic delay bounds for the encoder and decoder buffers are $D^{e}_{buff} = D^{d}_{buff} = 0.2 \ldots 0.4\,T_{frame}$.

F. GOP-CBR Intra Refresh Coding Mode

Increasing the CBR interval to a whole GOP eliminates the overhead data rate of the frame CBR mode at the cost of higher buffer latencies. An analysis of 30 intra refresh strategies as proposed in [6] indicates, that encoder and decoder buffers delays of $D^{e}_{buff} = D^{d}_{buff} = 0.6 \ldots 1.2\,T_{frame}$ can be achieved.

III. Conclusions

This paper described the basic delay sources of compressed video transmission systems. In an optimized implementation, the largest delays are introduced by the encoder and decoder stream buffers in CBR operation. Based on H.264 coding experiments bounds for the buffer delays are given.

References

[1] A. Ortega, "Variable bit-rate video coding", in *Compressed Video over Networks,* M.-T. Sun and A. R. Reibman, Eds, M. Dekker, New York, NY, 2000, pp. 343-382.

[2] J. Ribas-Corbera, S. Lei, "Rate control in DCT video coding for low-delay communications", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 9, No. 1, Feb. 1999, pp. 172-185.

[3] P. Navakitkanok, S. Aramvith, "Improved rate control for advanced video coding (AVC) standard under low delay constraint", *Proc. Int. Conf. on Information Technology: Coding and Computing,* ITCC 2004, 5-7 April 2004, Vol. 2, pp. 664-668.

[4] R. M. Schreier, A. M. T. I. Rahman, G. Krishnamurthy, A. Rothermel: "Architecture analysis for low-delay video coding", *IEEE Int. Conf. on Multimedia and Expo (ICME),* Toronto, July 9-12, 2006, pp. 2053-2056.

[5] J. Ribas-Cobera, P. A. Chou, S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC", *IEEE Trans. on Circuits and Systems for Video Technology,* Vol. 13, No. 7, July 2003, pp. 674-687.

[6] R. Schreier, A. Rothermel, "Motion adaptive intra refresh for the H.264 video coding standard", *IEEE Tr. on Consumer Electronics,* Vol. 52, No. 1, Feb. 2006, pp. 249-253.