

分类号 TP3

密级

UDC

编号

# 中国科学院研究生院

## 博士学位论文

实时流媒体系统若干关键技术的研究

高奎

指导教师 高文 教授  
中国科学院计算技术研究所

申请学位级别 工学博士 学科专业名称 计算机应用技术

论文提交日期 2005.4 论文答辩日期 2005.6

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院研究生院

## 摘要

近年来随着计算机技术、视频压缩技术、高带宽存储设备和宽带网络的发展，在 Internet 上的实时流媒体应用应运而生。然而，当前的 Internet 网络以尽力而为（Best-effort）的方式提供数据传输，带宽波动、延迟抖动和丢包等现象时有发生。传统的非伸缩性编码方法产生的码流不能适应网络带宽的波动。为此，希望采用可伸缩性编码（Scalable Coding）来提供网络自适应（Network adaptive）的流媒体服务。基于位平面编码的精细可伸缩性编码方法（Fine Granular Scalable, FGS）已经被 MPEG-4 标准接受，提供了细粒度的可伸缩性。该编码方法通常生成两种视频流：基本层码流和增强层码流。基本层码流必须传输，并且码率较低；增强层码流可以根据带宽网络状况任意截取，甚至可以不传。由于可伸缩性码流可以在一个很大的范围内自适应调整，因此能够适应复杂的网络带宽波动。本论文主要研究如何选择和调度数据包，来提供交互式的实时流媒体服务。主要贡献如下：

1. 本论文提出了基于端到端的网络自适应的实时流媒体系统体系结构。在这个框架中，我们引入了“实时”的概念，也就是，每一个数据包应该在一定的最终期限（Deadline）之前达到客户端，否则将不能被解码。这是本论文的出发点。

2. 本论文分析了可伸缩性流媒体的实时特征，并提出了基于分层的实时调度算法。在此基础上，本论文还提出了一个基于不精确计算（Imprecise computation）的负载模型（Workload Model）和实时调度算法来处理可伸缩性流媒体的调度任务。每一个可伸缩性的视频流的调度任务被分成两部分：强制执行的子任务（Mandatory subtask）和可选择执行的子任务（Optional subtask）。强制执行的子任务用于调度基本层的数据，可选择执行的子任务用于调度增强层的数据包，不同的子任务采用不同的优先级。基于不精确计算负载模型和实时调度算法为网络自适应的可伸缩性流媒体服务提供了灵活的调度策略，增强了系统的容错性，提高了资源利用率和客户端播放质量。该算法的低复杂度使得其可以在实际的系统中得以实现。

3. 本论文提出了一个在线平滑的调度算法，在可用网络带宽波动的情况下，能够提供较平滑的视频质量。由于网络带宽的波动和不可预测性，容易产生视频质量的剧烈变化。通常，人们宁愿看到质量稍低但平滑的视频节目，也不愿意看到质量稍高但波动的视频质量。该算法采用了基于失真度的性能评价函数并结合实时调度算法来选择要发送的数据包，使视频质量波动最小的同时，提高网络带宽资源的利用率。

4. 本论文提出了一个动态的资源分配策略为并发的多用户公平地分配服务器资源。由于流媒体服务器资源有限，特别服务器吞吐率（也就是 I/O 带宽）的限制，只能接入一定数量的并发用户请求。不同的用户在不同的时刻带宽的波动不相同，传输码流的码率也不相同。根据每个用户的可用带宽和码流的码率状况，动态、公平地分配服务器资源。该策略不仅可以公平地分配服务器资源，而且还提高了系统资源的利用率和客户端的总的播放质量。

5. 本论文提出了一种支持 VCR (Video Cassette Recoder) 功能的新框架。在流媒体应用中, 用户希望能够提供交互式的 VCR 功能, 如倒放、快进、快退、随机访问等操作。为了支持 VCR 功能, 在服务器端存储正放和倒放的 FGS 的基本层码流。两个码流中的预测编码的帧采用在变换域中进行预测, 并对预测误差进行量化, 即使采用不同的预测图象, 也可以得到相同的重建图象, 从而两个基本层可共用一个相同的增强层码流。在此基础上, 来实现面向网络自适应的 VCR 功能的支持。该框架不仅灵活支持 VCR 功能, 而且还大大节约了存储空间。

**关键词:** 视频流, 精细可伸缩编码, 实时调度, 平滑, VCR, 网络自适应

## Research on Key Techniques of Real-time Streaming System

GAO Kui (Computer Application Technology)

Directed by Professor GAO Wen

Recent developments in computing technology, video compression technology, high bandwidth storage devices, and high-speed networks have made it feasible to provide real-time streaming media applications over the Internet. However, the current Internet is a best-effort network where loss, variable delays and bandwidth fluctuation occur during data delivery in a streaming session. Traditional non-scalable streaming cannot adapt well to variation of the available network bandwidth. Scalable/layered encoding has been believed to be promising to provide network adaptive streaming service. FGS (Fine Granular Scalable) coding schemes is accepted as scalable video coding by the MPEG-4 standard. FGS video streaming is to code a video sequence into a base layer and multiple enhancement layers. The base layer uses non-scalable coding to reach the lower bound of the bit-rate range and must be transmitted. The enhancement layer may be truncated into any number of bits or not transmitted. Therefore, the FGS streaming can adapt a wide range of data rate variability by distributing enhancement layers over a wide range of bit rates. In this thesis, we investigate how to schedule and send the packets by real-time scheduling scheme and how to provide an interactive real-time streaming service over network with packet losses and variable delay. The main contributions of this thesis are as follows:

1. A real-time network-adaptive architecture for end-to-end streaming system is proposed. The important concept “real-time” is introduced in the framework. All the packets must be transmitted before their deadline, otherwise they can not be decoded on time. This is the foundation of this thesis.

2. This thesis analyses the real-time characteristic of scalable streaming system and proposes a layer-based real-time scheduling algorithm. And then, a real-time imprecise computation workload model and an imprecise computation scheduling algorithm on scalable media stream delivery are proposed. The scheduling task of each stream is divided into two parts: a mandatory subtask and an optional subtask. The mandatory task is for the base layer substream and the optional task is for the enhancement layer substreams. Different subtask adopts different priority. The imprecise computation scheduling algorithm enables the use of imprecise computation workload model as a means to provide scheduling flexibility in scalable streaming systems and enhance their

fault tolerance and improve utility of the system resource and the playback quality in client. The low complexity of the proposed algorithm also enables them to be applied in real-time applications.

3. This thesis proposes an online smoothing algorithm for a single of scalable media with the variable network bandwidth. Without smoothing, the playback quality would be variation along with network bandwidth fluctuation at the client, which is annoying to human ears and eyes. However, it is generally agreed that it is visually more pleasing to watch a video with consistent, albeit lower, quality than one with highly varying quality. This algorithm adopts an optimized framework based on the distortion and a real-time scheduling scheme to select and schedule the packet to the client according to the network bandwidth.

4. The thesis proposes an on-line resource allocation algorithm to allocate video streaming server resource for multiple concurrent scalable video streaming. Since the server resources, especially the I/O bandwidth or throughput of the server, are limited, only a limited number of concurrent clients with the requested QoS can be served. Different user has different variation of network bandwidth and bit rate of playback stream. The algorithm fairly allocates the server resource to multiple concurrent scalable video streams according to the network status and bitrate of every streaming session. The utility of the server resource and the total playback quality can be improved by the algorithm.

5. This thesis proposes a novel coding framework of jointing the FGS video streaming and VCR function supporting. In video streaming applications, it is highly desirable to provide digital VCR(video cassette recording) interactive functions, e.g. reverse play, fast forward/reverse play, random access, etc. Both the forward and reverse bit-streams of base layer are stored at the server end in order to support VCR funducation. The proposed technique is to take advantage of the fact that the residual difference signal resides in a quantized space in the transform domain. The prediction frame in the reverse bitstream and forward bitstream are generated with different reference, but their reconstructions are identical. Therefore, the same enhancement layer can be used for both base layer. Based on the modified MPEG-4 FGS coding framework, the full VCR functionality can be effortlessly supported and a lot of storage space can be saved.

Keywords: Video streaming, Fine granular scalability (FGS), Real-time scheduling, Smoothing, VCR, Network adaptive