

# 1

## Introduction

In answer to the increasing popularity of the Internet and especially one of its major applications, the World Wide Web, not only in research and professional environments but also among the wider public, the provision of new types of content is increasing rapidly. In the beginning of the World Wide Web, the available content was mainly based on text (hyperlink) documents and still images. With the increasing popularity for private users, content which is usually provided by the entertainment industry becomes more and more interesting. Such content is usually characterized as multimedia data [1] such as audio, video and the combination of both. With the need for such types of content, new services and applications are offered on the Internet. An example of new services is radio station programmes which are offered, in addition to the traditional terrestrial broadcast, via the Internet.

Nevertheless, it is obvious that these services are only popular where no alternative exists to obtain this content or where the offered quality is comparable to existing alternatives. For example, listening to a radio station via the Internet can be compared to listening to the same station on a car radio. The quality is definitely not equivalent to CD quality: on both Internet and car radio, interruptions and changes in the quality can occur. A client who is used to listening to the radio in his or her car accepts these quality degradations and, thus, is willing to use a service such as Internet radio. The explicit provision of live events on the Internet is an example of where there is a lack of alternatives. In this case users accept a degradation in quality, since there is only the one possibility for receiving the live event. The broadcast of live events, such as pop concerts or sports events, can even lead to partial collapse of the service caused by the high user demand.

The situation is completely different in cases in which alternatives in a much better quality are available. Since the mid-1990s several video on

demand (VoD) trials have been performed but none of them has resulted in a major success. In fact, there are only a few VoD services available in the Internet. Despite the failure of these trials a huge effort has been put into overcoming the problems that prevent VoD and video streaming from becoming a successful service in the Internet. Recent developments show that VoD is, at least in some areas, gaining popularity. This tendency is certainly supported by new technologies that allow users at home to receive data at a higher bandwidth and, thus, better quality. Yet there are still open issues that have not been solved so far. Two of these issues, which are considered in this book, are the absence of quality of service (QoS) in the Internet and the heterogeneity of the clients. Both require new mechanisms that allow an adaptation of the streaming rate to available network and client resources. Furthermore, an integration of these new mechanisms with a video distribution architecture is necessary to increase the overall scalability of VoD services. In this book, these new mechanisms and their integration in a video distribution architecture are presented.

## 1.1 WHY SCALABLE INTERNET VIDEO ON DEMAND SYSTEMS?

In the last few years, the Internet has been used for an increasing amount of traffic stemming from the emergence of multimedia applications which use audio and video streaming [2]. This increase is expected to continue and be reinforced since access technologies such as Asymmetric Digital Subscriber Line (ADSL) and cable modems enable residential users to receive high-bandwidth multimedia streams. One specific application which will be enabled by future access technologies is video on demand (VoD). True VoD (TVoD) [3] is a subtype of VoD which allows users to watch a certain video at any desired point in time while also offering the same functions as a standard VCR (i.e., fast-forward, rewind, pause, stop). The challenges of providing TVoD in the Internet are manifold and require the orchestration of different technologies. Some of these technologies, such as video encoding (for example, MPEG-1), are fairly well understood and established. Other technologies such as the distribution and caching of video content and the adaptation of streaming mechanisms to the current network situation and user preferences are still under investigation.

Existing work on TVoD has shown caches to be extremely important with respect to *scalability*, from the network, as well as from the video servers' perspective [4]. Scalability, of course, is an important issue if a TVoD system

is to be considered for use in the Internet. Yet, simply reusing concepts from traditional Internet Web caching is not sufficient to suit the special needs of video content since, for example, popularity life cycles can be very different [5].

In addition to scalability, it is very important for an Internet TVoD system to take the ‘social’ rules implied by Transmission Control Protocol’s (TCP) cooperative resource management model into account, i.e., to be *adaptive* in the face of (incipient) network congestion. Therefore, the streaming mechanisms of an Internet TVoD system need to incorporate end-to-end congestion control mechanisms to prevent unfairness against TCP-based traffic and to increase the overall utilization of the network. Note that traditional video streaming mechanisms rely on open-loop control mechanisms, i.e., on explicit reservation and allocation of resources. As it is debatable whether such mechanisms will ever be used in the global Internet, e.g., in the form of RSVP/IntServ [6], mechanisms presented in this book build upon the current best-effort service model of the Internet which is based on closed-loop control exerted by TCP-like congestion control. Yet, since video transmissions need to be paced at their ‘natural’ rate, adaptiveness can only be integrated into streaming mechanisms in the form of quality degradation and not by delaying the transfer as is possible with elastic traffic such as File Transfer Protocol (FTP) transfers. An elegant way of introducing adaptiveness into streaming is to use scalable video [7] formats as they allow dropping segments (the transfer units) of the video in a controlled way without high computational effort of, for example, adaptive encoding as described in reference [8]. Thus, it overcomes the inelastic characteristics of traditional encoding formats such as MPEG-1 or H.261. In addition, adaptive streaming in combination with an adaptive encoding format like layer-encoded video can avoid uncontrolled losses and, thus, increase the perceived quality of a video in contrast to an uncontrolled streaming. A side-effect of adaptive streaming is the fact that heterogeneous clients and access networks can be supported more efficiently.

Little work has been performed so far on the aspect of combining both, scalability for VoD systems and adaptive streaming. Thus, the focus of this book is on new mechanisms that combine the benefits of both approaches in order to maximize the quality of the video stream that is delivered to the client. However, while the combination of caching and adaptive streaming promises a scalable and TCP-friendly TVoD system, it also creates new design challenges. One drawback of adaptive transmissions is the introduction of quality variations during a streaming session. These variations affect both the viewer’s perceived quality and the quality of the cached video and, thus, the acceptance of a service that is based on such technology.

The overall question this book tries to answer is: Can the benefits of system scalability and adaptive streaming be combined to create new systems that can increase the performance of VoD services?

## 1.2 WHAT IS THE GOAL OF THIS BOOK?

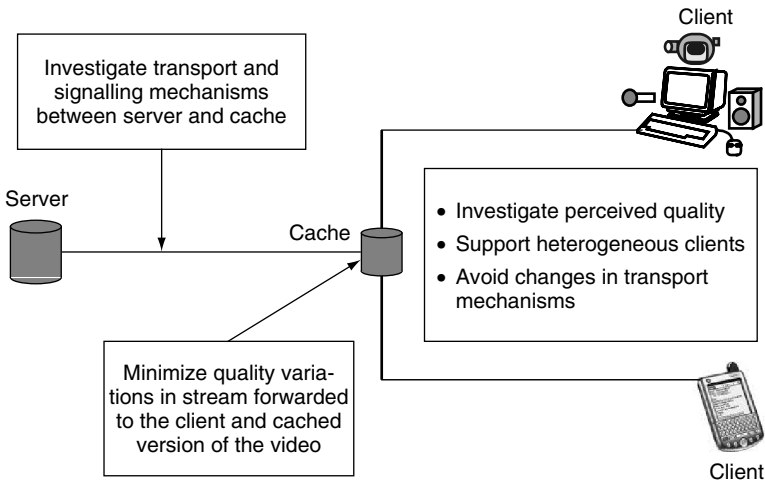
The goal of this book is to answer the aforementioned question by extending existing mechanisms and creating new ones to increase the performance of VoD systems in today's Internet (an Internet without quality of service support). The validity and applicability of these mechanisms are proven through investigations based on assessment, simulation and a prototype implementation which in combination lead to the final results. These new mechanisms should be usable as building blocks for scalable Internet VoD systems. Next to the development of the individual mechanisms it should also be shown how these mechanisms can be orchestrated to build a well-suited distribution infrastructure for VoD services which is in contrast to approaches where only isolated parts of the distribution infrastructure are investigated. Nevertheless, the mechanisms should also be usable independently of each other to allow VoD operators to tailor a service based on these mechanisms according to their specific needs. For example, the mechanism that reduces quality variations, which is located on the cache, should be independent of the transport mechanisms between server and cache. It is certainly not the goal of this book to demonstrate how a specific VoD application is built.

It is a fact that in an Internet without QoS support quality variations and data loss during a streaming session cannot be avoided. Therefore, those quality variations should be kept to a minimum in order to increase the acceptance of VoD services. The minimum of quality variations can also be seen as the maximum number of variations that the viewers tolerate. As a consequence, an intolerable number of variations would lead to the fact that users do not accept the offered service. Investigations on the subjective impressions of quality variations in layer-encoded videos which might reveal such information have not till now been performed. Therefore, this book offers better insight into how variations in scalable video affect the viewers' perceived quality by conducting such a subjective assessment.

Based on this newly gained knowledge we investigate whether caches can be used to improve the quality of a layer-encoded video stream that is transported from or through the cache to the client. In other words, how can these layer variations, with the aid of caches, be kept to a minimum. As a constraint, the mechanisms at the client used to receive and display the

video should be kept unchanged. This decision is based on the fact that it is far easier to establish new mechanisms on a manageable number of servers and caches (compared to the enormous number of uncontrollable clients). Next, the new mechanisms should be designed in a way that allows for heterogeneous clients and access networks.

Since scalability in a VoD service means, among other things, to increase the number of simultaneously served clients, it is important that server load is reduced by streaming data directly from caches to the client. Thus, it is also important not only to minimize layer variations in the stream delivered to the client but also to minimize these variations in the cached version of the video to allow the delivery of a high-quality video from the cache. Figure 1.1 shows the elements of a VoD service and the systems that are focused on in this book.



**Figure 1.1** Conceptual overview.

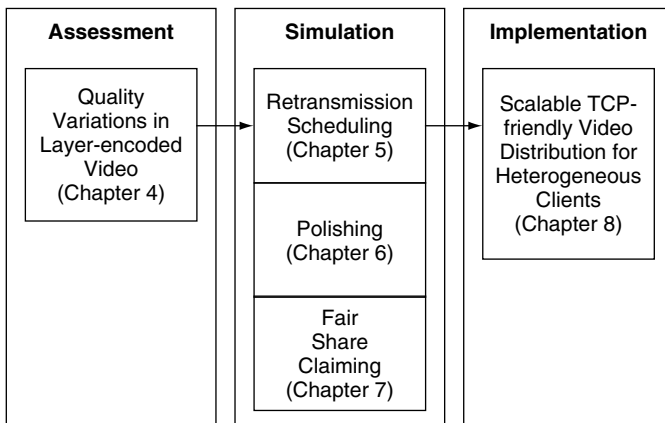
### 1.3 OUTLINE OF THIS BOOK

Chapter 2 gives an overview of the scalable adaptive streaming (SAS) architecture which combines system and content scalability in order to allow VoD services in a best-effort Internet in combination with heterogeneous access networks and clients. After an overview of the SAS architecture the building

blocks of such an architecture, which repeatedly occur throughout the book, are introduced.

In Chapter 3 an overview of related work in the area of video streaming and distribution is given. It is shown which solutions are already available as commercial products and what their shortcomings are. Activities in several standardization organizations are briefly mentioned in order to give the interested reader better orientation, since it is not just one organization that is involved. In the remainder of this chapter related work from the research area is presented that has served as the basis for the work presented in this book.

A survey on related work in the area of retransmission scheduling revealed a lack of subjective investigations on how layer variations in layer-encoded video influence the viewer's perceived quality. Existing work on retransmission scheduling is based on speculative assumptions. Therefore, as a first consequence a subjective assessment is performed in the scope of this book to get better insight into how layer variations affect the perceived quality. The subjective assessment is presented in Chapter 4.



**Figure 1.2** Organization of this book.

The results of this investigation are used to confirm the applicability of an objective metric which is developed in order to evaluate heuristics for retransmission scheduling. The investigations on the latter (presented in Chapter 5) show that an optimal solution, given reasonable computing power, is computationally infeasible.

Additionally, the results of the subjective assessment reveal that an increase in the amount of stored data for a cached layer-encoded video object does

not necessarily increase its perceived quality (see section 4.5). This means that dropping certain segments of a layer to reduce the amount of variation can increase the perceived quality. Based on this knowledge, reducing layer variations by dropping certain segments seems to be an additional option to improve the perceived quality, leading to a new mechanism called *polishing* which is presented in Chapter 6. Polishing can be used either for cache replacement or during playout from the cache to the client. In the first case, segments are deleted from the cache, based on the polishing algorithm, in order to free storage space for new video objects, while in the second case certain segments are not streamed from the cache to the client.

With a simulative environment solely built to investigate the newly created mechanisms for retransmission scheduling and polishing, a series of simulations are performed. The goal of the simulations is to show the applicability of both mechanisms and their dependence on certain parameters (e.g., the available bandwidth for retransmissions). The results obtained by the simulations were satisfying and showed, in the case of retransmission scheduling, a significant improvement compared to already existing mechanisms.

In subsequent work, which is presented in Chapter 7, a new mechanism allowing the transport of segments requested for retransmission is developed, leading to a combination of TCP-friendly streaming and retransmission scheduling. This approach has a side-effect allowing a TCP-friendly transport stream to claim its fair share on the network path, although layer-encoded video is transmitted. Based on the mechanism for fair share claiming an implementation design for an already existing streaming platform is made.

In Chapter 8, this design is extended to allow scalable TCP-friendly video distribution for heterogeneous clients. Therefore, the cache is extended by gateway functionality enabling standard clients in the SAS architecture. Based on this extended platform, experiments are performed to demonstrate the applicability of the newly created mechanisms which are building blocks of the SAS architecture.

Finally, a summary of the contributions created in this book is given and final conclusions are drawn.

## 1.4 WHO IS THIS BOOK FOR?

- Students: This book is certainly not created as classic textbook. So, it is not meant for students who are about to learn the basics of video streaming and video distribution systems in the Inter-net. It should rather be seen as introductory literature for students working on streaming-related projects.

In addition, it can serve as complementary literature for students who are highly interested in this topic.

- **Lecturers:** As already mentioned above, this book does not have textbook character and was not written with this goal in mind. Nevertheless, this book can serve lecturers as additional material to update or extend their existing lectures. For example, parts of this book fit very well in a content distribution lecture, while other parts can be useful in a networking class.
- **Designers:** This book should be very helpful for software engineers who are involved in the design of a video streaming architecture which is based on existing Internet technology. Many of the results presented in this book can be used as guidelines for making decisions for a future system.
- **Implementors:** The overall architecture for scalable and adaptive streaming systems presented in this book might be a good starting point for someone who has taken on the task of implementing such a system. It should be mentioned here that most of the results presented in this book are based on an existing implementation which is available as open source (<http://komssys.source-forge.net>).