Designing a Scalable Video On Demand System

Jason But, *Professor Greg Egan*

CTIE - Centre for Telecommunications and Information Engineering
Monash University, Australia

Abstract: As network access costs, especially over the last mile, decrease and broadband access becomes more available to the home user, the idea of Video-on-Demand (VoD) as an application has made a comeback. Previously, VoD systems could only support a local Intranet, however when considering deploying a global VoD service, the issue of scalability becomes important. Indeed many low-bitrate Internet streaming services have led the way in distributed video server design. At this point we need to re-examine the issues of designing a global VoD service that is not only scalable, but also cost effective – the most important driver in deciding whether VoD will survive its latest incarnation, or whether it will fade away to bide its time again.

Keywords: Video-on-Demand, scalability, distributed server, global service

I. Introduction

Video-on-Demand (VoD) has suffered many false starts in the past. Originally, implementations didn't proceed beyond trial systems due to technical difficulties in streaming high bit-rate video over existing networks. More recently, these technical limitations have been overcome, only to find trials failing to become economically viable – the service could not be made to run at a profit. As networking technology improves, the costs of implementing a video streaming service continue to drop. Indeed, low bit-rate video on the Internet has boomed in popularity due to its improved quality – made possible by better encoding algorithms and improved networks. As a result of this, it is obvious that high quality streaming digital video services will become available in the Internet in the near future.

Given this scenario, it is interesting to re-examine lessons learnt in previous trials, and to outline design rules that must be adhered to if designing a scalable video streaming service. An ideal service would be able to service all Internet users globally, or at least nationally, and must be economically feasible while being competitive to local video/DVD hire stores. In this paper we will discuss the original central streaming server design and contrast it with the more recently favoured distributed server design. We will also discuss scalability issues and implementation costs involved. Finally, we will show that while a distributed server streaming video system will scale to service a large and widespread customer base, the costs involved are still too high for consideration today and in the near future. This results in a new question: *Is there a system design such that the technical advantages of the distributed server design are available, yet implementation costs are reasonable?*

* Corresponding author: jason.but@eng.monash.edu.au

II. Entertainment Quality Streaming Video

With the recent successful introduction of Digital Home Video in the form of DVD, many people are again considering the concept of streaming high-quality digital video over a public network infrastructure (such as the Internet) into the customer home. This idea is not new and has been both proposed and trialled many times with partial success[1-3]. There are many reasons why trials of this nature have not succeeded in the past, but these boil down to two major problems. The first of these problems relate to the network or Internet, even today, the Internet is not capable of supporting a true VoD application, the expense of bandwidth coupled with limited quality of service conspire to affect video streaming – often rendering the received video un-viewable. The second of these problems, often not encountered in all but the largest trials, is the scalability of the server complex. The question often asked is not only whether the server can support a large number of users, but also whether these users can still be supported when spread over a large geographical area.[1, 2, 4-6]

Many trials were so intent on proving whether or not video could be streamed that they forgot about whether the service would be viable. When considering if a service will be viable, we must consider it from two viewpoints, the first and foremost is whether operating the VoD service will be profitable. Considering the service will be in competition with the local Video Hire Store, the service will only be viable if it can operate profitably whilst charging essentially the same rate for video hire ($6 per day). Secondly, a video service will only be viable if customers are willing to view the videos on offer, this means that popular hires and new releases must be available on a digital streaming service. This brings us directly to the issue of copyright and its protection on streaming media. While consumers can pirate an analogue tape from the video store, the copy is of inferior quality and further copies even more so. The fear with digital video is not only that a copy is a perfect reproduction of the original, but also that repeated copies do not degrade in quality. As such, copyright protection is an important issue that must be resolved before a commercially viable streaming video system can be implemented.[5, 7]

In summary, to build a viable streaming video system we must consider the technical issues of streaming a video to a customer site, the economic issues of running a profitable service while maintaining user costs at competitive levels, and the security issues of ensuring that the digital video stream is protected against theft and tampering. In this

0-7803-7547-5/02/$17.00 © 2002 IEEE

paper we will explore the issues of building a technically viable streaming service as well as investigate the costs involved in providing this service. The costs explain why most VoD trials have failed and will continue to do so in the near future if the same system design models are used.

A. Networking Issues

Technically, the existing customer access network is the major hurdle when considering providing entertainment quality digital video over the Internet. Current low bit-rate video, while acceptable for small-screen playback, is not of sufficient quality to be viewed on a television screen in the users lounge room. Since the service must compete against video/DVD hire stores, the quality of the streamed video must be comparable. As such, digital video should be encoded in either MPEG-1 or MPEG-2 format. MPEG-1 video encoded at 2Mb/s is of a better quality than VHS but inferior to DVD. MPEG-2 video encoded at 6-8Mb/s is equivalent to the digital media on a DVD disk.[3, 7]

Video compressed at these rates cannot be transmitted over a standard modem connection to the Internet. The first requirement is that the customer have a broadband connection to the Internet, this means either an ADSL/cable modem connection today, or a 3G mobile phone/direct fibre connection in the future. Even so, a broadband connection only supplies a high-speed link from the customer to the ISP. In order for a third party to stream high bit-rate video to the client, there are more technical hurdles to overcome.

Streamed video is delivered across the network at the average compressed rate of the video, while downloading a video involves transferring the data across the network at the maximum available rate. The act of streaming at the encoded rate implies that this bandwidth must be available at all times during streaming, this requirement is often referred to as a guarantee of the Quality of Service. QOS is a network feature commonly available in the public phone network, where the bandwidth required for a call is reserved for the use of the communicating parties. However, QOS is unavailable in the Internet and similar data networks where a best-effort service is usually provided. The common solution to providing sufficient QOS for video streaming is to over-dimension the network so that sufficient bandwidth is always available. Even this can be difficult due to bursty data traffic occasionally flooding available bandwidth on a given link. Another QOS problem is due to the random amount of time a datagram spends in router queues.[8, 9]

Much work is being done in the aim of improving the quality of service available over the Internet, this work forms part of the Internet2 or Next Generation Internet project which seeks to improve service by:

- Increasing link and therefore available bandwidth.
- Faster router processing – reducing delays in queues.
- QOS features (IntServ and DiffServ) being made available throughout the network.
- Faster customer access rates due to broadband access technologies.

While improved broadband access technologies will increase the maximum throughput to a level capable of supporting entertainment quality video streams, it does not guarantee that it will be possible to stream a video to the customer. To do this, we must be able to ensure that the necessary bandwidth is available within the network between the server streaming the content and the customer. This is not a major problem when considering a server servicing a local site, or a server located within the ISP network. It does become an issue when the server is located some distance (network-wise) from the customer. When designing a video streaming system, it is necessary to consider the size and location of the potential client base and to ensure that both the server and network design will be able to stream video to the end customer.

B. Previous VoD Trials

There have been many VoD trials in the past, all of which have failed for varying reasons. There were two major factors in the failure of these trials – the first is lack of consideration of bandwidth costs, which at the time were expensive. The second was lack of consideration of scale, in that systems were designed to service a small number of clients in the local area only. While none of these systems remained commercially operational, a great deal was learnt from these early trials[3, 6-8, 10]

Now that the cost of bandwidth has dropped – unlimited broadband access in Australia is available at rates of about $2 per day – the first layer of economic viability has been addressed and users can connect with sufficient 'last-mile' bandwidth to stream video.

Considering the issue of scale, a content provider will desire to establish a service that can stream video on a large scale – if not globally, at least nationally. Previous trials failed this test as they generally consisted of a single large streaming server located at the ISP. Since bandwidth was plentiful between ISP and customer, it was technically possible to implement a functional service. These systems were also expensive and could not scale to provide service to a larger group of users, the expenses in expanding such a trial system were prohibitive.

C. Cinemedia SWIFT Trial

CTIE was involved in two digital streaming video trials, the first – McIVER – was internal to the University, while the second – SWIFT – was a collaborative effort between CTIE, Cinemedia (Film Centre) and SGI. The purpose of the trial was to digitally encode a substantial amount (200 hours) of Australian content and to make it available to secondary schools and other educational institutions throughout Victoria. The initial design chosen for this trial involved a large SGI streaming server installed at Monash University configured to stream to a number of clients connected via the Internet. Since Internet bandwidth was

insufficient, an 8Mb/s ATM link between Monash University and Cinemedia was established to enable Cinemedia to view up to three separate video streams from the server. Despite the expense involved, Cinemedia were still limited to a maximum of three consecutive streams.[5]

When connecting other remote sites, CTIE first considered the use of cable modem to stream video – unfortunately, we discovered that cable modem could only support one high quality stream and even then only if the shared medium was not congested and there was sufficient buffering at the client[11]. The difficulties and costs involved in streaming video to remote locations led CTIE to change the system design to include a series of distributed video servers. One small streaming server was installed at each remote location and content was delivered from the central server by either slow network transmission, or CD-ROM if network costs were too high. In this setup, the Monash University and Cinemedia sites were serviced directly by the central server, while one local and one rural school were serviced by locally installed streaming servers. Content delivery to, and installation onto, the remote servers was done manually rather that automatically.[5]

This design showed the beginnings of a system that could service a large area and a large number of clients. A local streaming server services all customers in the immediate vicinity whilst content management remains at the central server. Content can be delivered to remote locations over slow or high-speed links using the available bandwidth, and then installed. A single transfer to a remote server allows many copies of that asset to be streamed remotely at lower cost. The system was not ideal, both content management and payment need to become fully automated to keep running costs to a reasonable value.

Copyright and payment issues were also addressed in the SWIFT trial. At the behest of the content owners, a system was developed to track viewing of each asset and to distribute payment to all parties as required. It was also important to consider the issue of protecting the digital asset whilst it was being streamed. To this end, CTIE considered the issues involved in encrypting the streaming video.

III. User Needs

There are two different sets of users of a video streaming system. The first of these is the content provider – these people own content and wish to use a streaming system to allow customers to access and view this content. The second group of users are the customers – these people pay to view a particular video asset and wish to use the streaming system to access video entertainment in place of a video or DVD. In this section we discuss the needs and requirements of these two groups of people.

A. Content Provider Requirements

A content provider requires three major conditions to be met before agreeing to utilise a particular video streaming system. The first is customer reach – the provider will wish to be able to deliver their video material to as many users as possible, over as wide an area as possible. The second is protection of material – the provider will wish to be secure in the knowledge that digital copies of their assets cannot be made and that copyright on their material will be protected. The final condition is guarantee of payment – the provider will want to ensure that the money transfer is secure, and that it is impossible to view a video stream without paying.

The first condition is a technical issue for the system design engineers to consider. A content provider will desire this condition not only because they wish as large a customer base as possible, but also because advertisement and management of the delivery system will be simplified. If a system is available everywhere, advertising can be global and generic, with less complaint from users who have seen the service advertised but are unable to access it. Also, management costs are reduced if a single system can service an entire nation rather than employing a different system to service each population centre. The last two conditions are an absolute requirement governing the economic viability of provision of a digital video service. Without meeting these requirements, a trial system will never become commercial.

B. Customer Needs

Customers of a streaming video service also have a number of requirements that must be met before they would consider utilising the service. The first is ease of use – the service must be as simple to use as Web browsing is today, if the service is complicated and difficult to use, patronage will be low. The second is freedom of choice – a user should not be tied to a single content provider and must be free to use other service providers. Also, use of a particular streaming service provider should not require usage of a particular ISP, nor should patronage of an ISP mandate a particular streaming service provider. The final condition relates to payment – the user must be certain that personal details and resultant money transfers will be kept secure, and provision of service for payment must be guaranteed.

The most important conditions are the first two, computer applications consistently fail if the interface is complicated and difficult to use. Also, as the World Wide Web doesn't mandate which browser to use, which sites to visit, or which sites are visible from a given ISP: similarly a video streaming service should not place restrictions on the user such as geographical location or utilisation of an ISP.

Both the customer and content provider conditions are equally important – if either user group is unhappy, they will not use the service. Absence of either user group leads to service failure, a lack of customers will minimise returns to content providers who would then operate at a loss – causing the eventual failure of the streaming service. No content providers means a lack of content – customers will cease to use the service due to a lack of interesting material.

IV. Central vs. Distributed Streaming Servers

There are two competing video streaming system designs; a monolithic central streaming server versus a number of smaller distributed streaming servers. Opinion has swung away from the central server design in favour of a distributed server design. In the following sections, we will discuss the implications of both designs and outline their major problems. While a distributed server design meets the technical issues of streaming video better than a central server, the design has problems that ensure that while the system is technically feasible, it is not economically viable.

A. Central Server Design

If an organisation is interested in providing a VoD service over the Internet, the most obvious solution would be to purchase a powerful computer equipped with a large disk array and video streaming software coupled with a high bandwidth connection to the Internet. As long as the network provides the required QOS between the server and the intended customers, the system will function adequately. This design is shown in Fig. 1, in which a single large video server streams video over a nationwide network. The problems with this design are fourfold:

- The number of potential customers is limited due to the high cost of adequate bandwidth.
- High system upgrade costs to cope with the load of an increased customer base.
- The design cannot scale to service a widespread user base such as an entire nation, high costs in QOS provision over large areas prohibit this.
- A system failure results in a loss of service.

The major limitation of this system is the high cost of bandwidth, which involves the connection of the server to the Internet, as well as the available bandwidth between the server and customers. Whether the customer or the service provider directly pays the cost, it will eventually be passed on to the customer. In today's network environment, high bandwidth costs limit the size of the customer base. In the SWIFT trial mentioned earlier, an 8Mb/s broadband connection was purchased between the University and Cinemedia sites. Despite the high cost of this link, only three concurrent video streams were possible. To extend the system to cover a larger customer base, including many secondary schools statewide, the overall cost of providing the necessary bandwidth was prohibitive for a complete implementation. However, these costs will decrease as the gradual introduction of the Next Generation Internet provides not only greater bandwidth, but also better management of that bandwidth through QOS.

The second problem inherent in a single server design is the upgrade costs involved when the customer base exceeds the server's capabilities. In this case, the server must be replaced by a larger server able to service the increased requirements of a larger customer base. These costs involve hardware, licenses for the video streaming software, and a more expensive – faster – connection to the Internet. These costs do not increase linearly with increased customer numbers, thus as the service becomes more popular, the service provider faces increased costs per customer. Another major problem with a single server solution is that there is a single point of failure. Should the video server suffer a system failure, all customers would face a loss of service, at cost to the service provider.

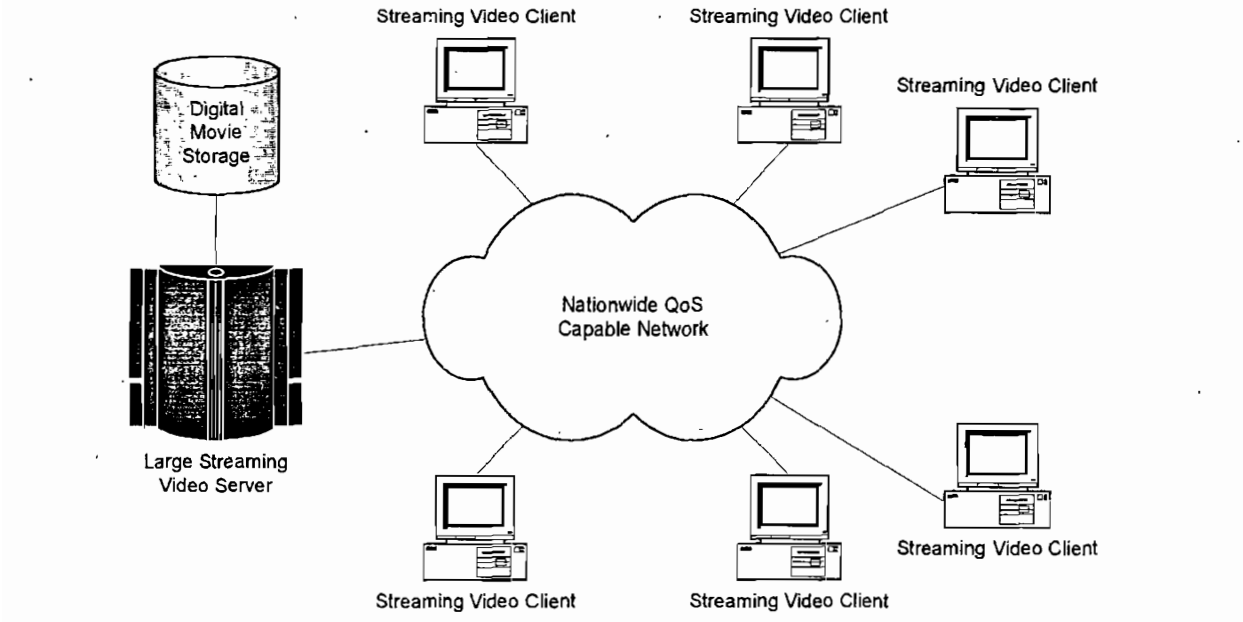

Fig. 1. Monolithic Single Server VoD System Design

The final, and most severe, problem occurs when considering provision of a service to a widespread customer base, such as nationally. The service provider, and therefore the customer, must pay for a high bandwidth, QOS enabled link between the server and each customer. The cost of providing and managing a QOS enabled link over a large-scale network is extremely high. As such, it is economically unfeasible to site a central server in a capital city such as Melbourne and expect to provide a video service to the entire nation. The costs of streaming a single movie to a remote location, like Western Australia, is prohibitive, let alone streaming to multiple customers at this location.

In Conclusion, a single server design is not economically feasible when streaming entertainment quality video to a large and/or widespread user base. It effectively limits entrance to VoD service provision to large corporations with money, smaller film distributors could not afford to provide streaming video to a large user base. Given the inherent problems with scaling such a service, a single large VoD server is unlikely to be used in a real-world implementation. Indeed, lower quality streaming Internet Video developers (such as Real Networks, Microsoft Media Services, etc.) are already developing distributed servers and caching systems to overcome the problems inherent in a single server design, changes which will filter through to high-quality video streaming products. I will discuss the design of a distributed server system in the next section.

B. Distributed Streaming Server Design

A distributed VoD server design overcomes many of the limitations inherent in a monolithic single server system. In this configuration, multiple smaller servers are configured at remote locations to service the customer base in the immediate local area and requested videos are transmitted at the available bandwidth without QOS guarantees between the distributed servers on the network. This system design is shown in Fig. 2, where two large companies are operating competing nationwide VoD services. Since the servers only stream video to the local area, the number of customers that each services is lower, and cheaper, less powerful hardware can be used. Also, costs in providing a QOS guaranteed connection are lower between a client and a local server as compared to a remote server. An increase in the user base within a local area can be handled by either increasing the capacity of the distributed server servicing that area, or by installing a second distributed server within the same area. Increasing the system coverage is a matter of installing a new server in the new remote location. The only major drawback remaining in this design is the large costs of a single company implementing a large enough system to cover a wide area, meaning that smaller video distributors are locked out of the networked video service industry.

The distributed server design was also used in the SWIFT trial, whereby two smaller servers were installed at remote locations (secondary schools in Victoria). Assets were now either transferred overnight using slower, cheaper, network connections, or copied to CD and physically transferred to the remote server. The asset was then installed on the remote server, which serviced multiple clients at the site without the need for an expensive broadband connection back to the central server. Indeed, a remote server with a low speed connection back to the central server was able to serve more concurrent streams at lower cost than the central server could over the broadband connection to Cinemedia. This proved the viability of the distributed server design, and that costs were lowered. Even though the streaming servers were of the same brand, interoperability between servers was still a problem, as was the increased complexity in asset management and transfer between servers.[5]

The operating principles of a distributed server design lead to a better allocation of resources, even if a cheap QOS capable nationwide network becomes available. The main reason for this lies in the network requirements for video streaming – each user requires QOS guarantees between themselves and the streaming server. In a single server design, this guarantee must be provided from the central site to all clients currently streaming video, no matter their location. Not only does this potentially require more data being transferred over greater distances (for multiple concurrent streams), but that each stream only utilise its required bandwidth for viewing the video. In a distributed server arrangement, video assets are copied between servers at the current available bandwidth (which could be faster than the required streaming bandwidth), and only streamed using QOS guaranteed connections from the local server to the user. As a result, bandwidth management is simpler as it is only required between the local server and the client. Nationwide data transfer drops, as a single transfer to one distributed server will then service all clients in that area.

Video assets are transferred between distributed servers, at the available bandwidth. If the nationwide backbone has capacity for an 8Mb/s file transfer, then a 2Mb/s encoded video can be transferred to the remote server four times quicker than if the asset was being streamed. The remote server can immediately commence streaming as it is receiving and installing the video asset. Other clients within the same area can access the video on the server without a second transmission from the central server. In a situation where there are popular movies, such as new releases, they can be pre-delivered to the distributed servers during off-peak time to take advantage of cheaper network premiums.

The biggest advantage of a distributed server design over a single server design is its scalability. The main reasons why a distributed server configuration is more scalable are:

- Lower Network Infrastructure Requirements and Costs – A nationwide QOS capable network is not required as streaming takes place from a local server, a high speed backbone is required for content transfer between servers. QOS guarantees are only required between the client and the closest streaming server, network costs are reduced due to lower QOS requirements.

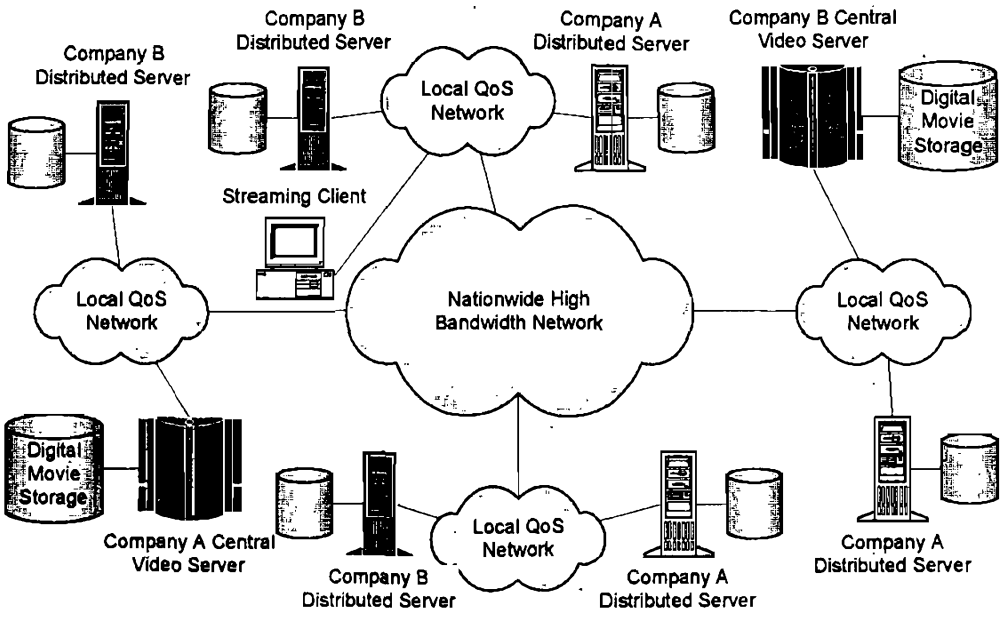

Fig. 2. Distributed Server VoD System Design

- Server Infrastructure Costs – The service provider no longer needs an expensive streaming server. A file server with a large disk array can be used as a central store and smaller, cheaper video servers can be distributed around the network. These servers are cheaper to both buy and maintain. In the event of a system failure of a distributed server, the load can be shifted to a nearby server without loss of service to customers. Replacement of a failed server is also simplified.
- Growth of Customer Base – If the customer base grows, the system can be scaled to support these customers by either the addition of further distributed servers or upgrading an existing distributed server to a larger model.
- Extending Range of Service – Extending the range of service requires a local QOS capable network, the addition of a single distributed server to this area, and adequate average bandwidth from the central server to the new distributed server. This is cheaper than a nationwide QOS capable network.

There are still some problems with the distributed server design, these involve system management as well as overall implementation costs. When comparing with a single server design, we have increased the implementation complexity, and therefore the management of such a large system. With a single server, asset management involves keeping track of assets installed on the server, in a distributed server design, we must keep track of assets installed on each server as well as the location and status of each server. Whilst this added complexity would certainly make an implementation more difficult, it is an obstacle that must be overcome in order to provide a scalable VoD service. These features are already being integrated into existing Internet streaming video products (Real Networks, Windows Media Services) as these products move from single to distributed server implementations. Even so, these products require a single brand of video streaming software communicating using proprietary protocols, forcing operators to select a single brand to minimise interoperability issues.

The other problem is the overall cost required in setting up a nationwide VoD service using the distributed server model. While the costs are lower than those required for a working single server solution, they are still prohibitive for all but the largest companies. Also, if two competing companies set up such a system, there would be a large duplication of hardware in order to for both companies to provide a similar service to all users nationwide.

In conclusion, a distributed server design will readily scale to provide not only a nationwide, but also a global service if a capable Internet backbone is available. The cost however, of providing this service is still prohibitive for smaller companies and effectively limits access to the industry to large companies. Also, competing VoD services leads to duplication of equipment within the network and possibly some areas being serviced by only one company. A distributed server design has solved the scalability issues of a single server design, but hasn't addressed the issue of costs and the economic feasibility of implementation.

V. Conclusion

In conclusion, recent and expected future improvements in both Internet infrastructure and customer access technologies mean that the network will be capable of supporting an entertainment quality streaming video

application in the near future. Given this, it is a good idea to review the lessons learnt from past video streaming trials. The most remarkable problem with past trials is that the major goal had been in overcoming the technical difficulties in streaming video to a customers place of residence, but scant thought was given to designing a system that would scale to service a large number of users over a widespread area. Also, in the rush to implement a functional system, no effort was made to cover other requirements of both content providers and users such as payment, protection of content, ease of use, competitive rates, and probability of profit.

Having looked at the original central server design and examined its problems, we conclude that this system design is not scalable and therefore not useful to consider when designing a video streaming system. Instead, a distributed server design offers scalability and better management of existing network resources. Just as importantly, this design makes no QOS demands over long network hauls, requiring the network provide adequate service only from a local server to the customer. The major result to come from this review is that a distributed server design is the preferred option when designing a streaming video service. Indeed, many low bit-rate streaming products now allow for distributed servers and caching of video streams. It is only a matter of time before these features become predominant and widely spread through all video streaming products.

There are still problems to be solved however, the overall cost in providing a true distributed streaming server system in prohibitive for all but the largest companies. Even if the system were implemented in a small area and expanded slowly, the costs mean small content owners cannot offer their product to the market. Also, much of this cost will be duplicated if two or more corporations decide to offer similar services in competition. Finally, this outlay of costs makes it more difficult to build and operate a service that can successfully compete with the local video hire store. This means is that more issues have emerged which must be addressed before a video streaming system will be economically viable. These issues include lowering the implementation costs, allowing smaller content owners to operate their own streaming service, protection of content, securely handling money transactions, and providing this in a single application that can be easily used and understood by a non-technologically oriented home user.

References

[1] A. Viña, J. L. Lérida, A. Molano, and D. delVal, "Real-Time Multimedia Systems," presented at 13th IEEE Symposium on Mass Storage Systems, Annecy, 1994.

[2] Y.-H. Chang, D. Coggins, D. Pitt, D. Skellern, M. Thapar, and C. Venkatraman, "An Open-Systems Approach to Video on Demand," IEEE Communications Magazine, vol. May 1994, pp. 68-80, 1994.

[3] S. Fist, "Dial M For Movie: Video-on-Demand," Australian Communications, vol. August 1994, pp. 65-72, 1994.

[4] H. J. Chen, A. Krishnamurthy, T. D. C. Little, and D. Venkatesh, "A Scalable Video-on-Demand Service for the Provision of VCR-like Functions," presented at 2nd International Conference on Multimedia Computing and Systems, Washington D. C., 1995.

[5] T. Cornall, B. Pentland, and P. G. Egan, "Digital Media Library Project. Video on demand for schools.," presented at International Symposium on Intelligent Multimedia and Distance Education, Baden-Baden, Germany, 1999.

[6] P. Branch, A. Newstead, and R. Kaushik, "Design of a Wide Area, Video-On-Demand User Interface," presented at ATNAC 1996, Melbourne, 1996.

[7] S. Leask, "Video on Demand - Service Possibilities," presented at Australian Broadband Switching and Services Symposium '92, Monash University, Melbourne, 1992.

[8] M. Jadoul, B. Voeten, and W. Verbiest, "Interactive Video-on-Demand: Design and Implementation of an End-to-End System," presented at ATNAC'94, Melbourne, 1994.

[9] K. H. Tseng and K. S. Lim, "User's QoS Requirements and Guarantees for Real-Time Digital Video Communication," presented at ATNAC'94, Melbourne, 1994.

[10] P. Branch and B. Tonkin, "Multicampus Video On Demand at Monash University," Australian Journal of Educational Technology, vol. 13, pp. 85-97, 1997.

[11] T. Cornall, "CTIE-TR-1998-03: Evaluation of Optus Cable Network for 2Mbit/s Real-Time Services," CTIE - Monash University, Melbourne CTIE-TR-1998-03, May 1998 1998.