ON PRESERVING THE QUALITY OF INTERNET STREAMING THROUGH A DOCSIS NETWORK

Matt Haines and Asha Vellaikal
Aerocast, Inc.

Abstract

Internet-based streaming media services can deliver high-quality audio and video to PC users at speeds from 500 kbps to over 1 Mbps. However, delivering a large number of these streams through a DOCSIS network can push the boundaries of the network capacity. The result is increased congestion over the DOCSIS channel that impacts all downstream users and reduces the quality of the real-time streams being delivered through the network.

In this paper we discuss solutions that can be used to preserve the quality of streaming media services through a DOCSIS network. These solutions can be divided into two categories depending on whether or not the underlying DOCSIS network supports Quality of Services (QoS) features. For DOCSIS 1.0, which does not support underlying QoS features, we discuss methods for adapting the streams in response to network conditions. For DOCSIS 1.1, which does support underlying QoS features, we discuss methods for streaming media applications to utilize the underlying QoS capabilities.

INTRODUCTION

Cable providers using the Data Over Cable Services Interface Specification (DOCSIS) now provide broadband Internet service to a growing number of homes and businesses. As of September 2001, 30 Multiple System Operators (MSOs) served 7.6 million cable modem subscribers [1].

In a typical configuration (Figure 1), a single Cable Modem Termination System (CMTS) provides a dedicated 27/38 Mbps downstream data channel that is shared by up to 1000 cable modem homes [2]. If 10% of the homes are equally sharing the bandwidth at any one moment, then each home would receive approximately 270/380 kbps of downstream bandwidth. However as emerging "high bandwidth" applications take hold, such as IP telephony and streaming media, this allocation falls below the threshold of adequate bandwidth.

The Internet Streaming Media Alliance (ISMA) [3] Profile 1 is targeted towards broadband users who want to view entertainment quality Internet media streams over personal computers or set-top boxes. Profile 1 is based on MPEG-4 [4] and supports video encoding rates from 500 kbps up to 1.5 Mbps. This means that a small number of users who are consuming high quality streaming media can dominate the total aggregate downstream bandwidth from the cable head end unless steps are taken to limit their bandwidth usage. However, aggressive bandwidth limiting techniques, such as throttling the cable modems to specific levels, often results in a poor quality streaming media experience. Unlike web traffic, the end-user experience for real-time streaming protocols degrades as bandwidth is limited below the encoding rate.

Thus the problem facing cable operators offering DOCSIS services is how to effectively control downstream bandwidth usage while preserving the quality of streaming media and other real-time, high-bandwidth services. In this paper, we address this problem by presenting a number of possible solutions for preserving the quality of streaming media. These solutions are grouped

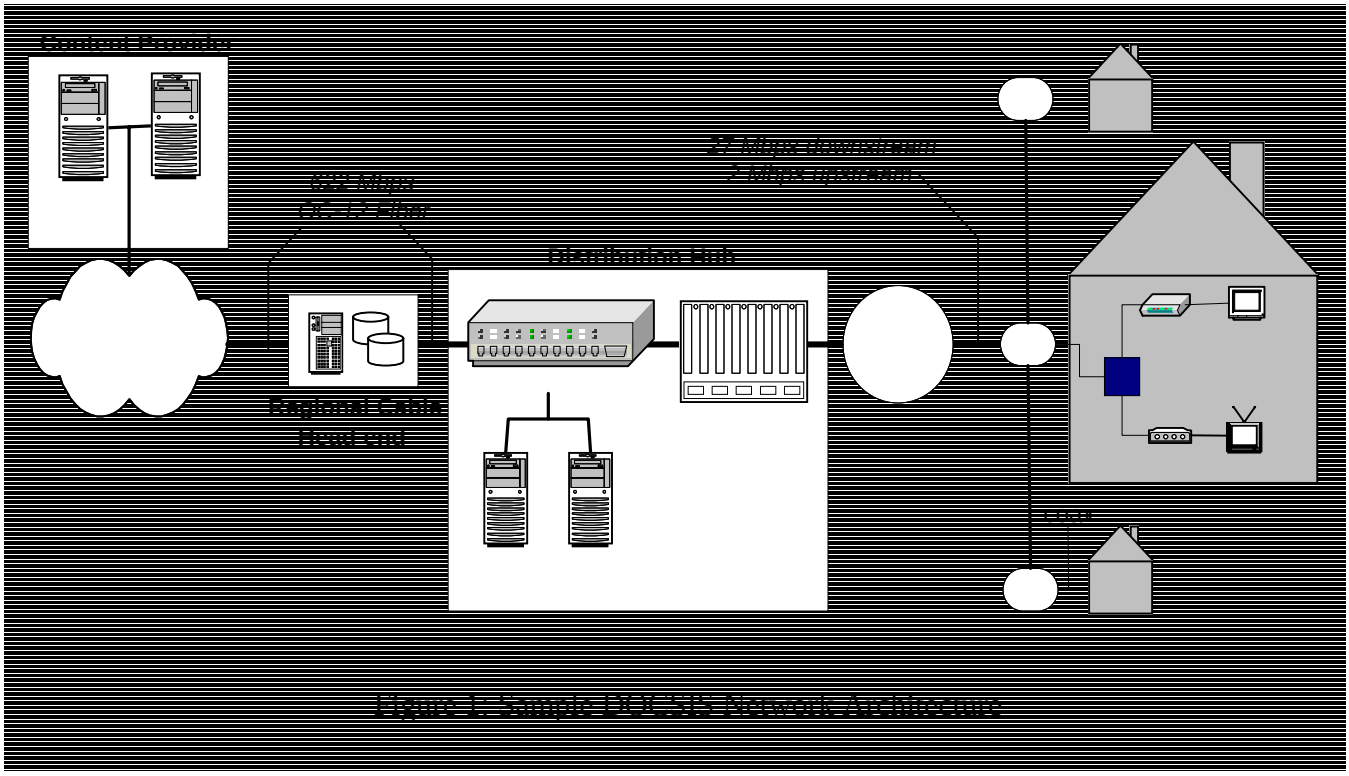

Figure 1: Sample DOCSIS Network Architecture

into two broad categories corresponding to DOCSIS 1.0 and DOCSIS 1.1, where the latter provides dedicated QoS features.

DOCSIS 1.0

In DOCSIS 1.0, there are no provisions for Quality of Service features such as bandwidth reservation. However, there are other means for preserving the quality of streaming media under demanding network conditions.

Co-Location with CMTS

Even with plenty of downstream bandwidth from the CMTS to the user, poor quality streaming will occur if the cable distribution hub (or regional head end) experiences congestion from the content provider origin servers through the Internet backbone. By co-locating streaming media servers within the distribution hub, the cable operator effectively shortens the end-to-end transmission and retains total control of the bandwidth needed to provide high quality streaming media services.

Co-location requires that all origin server content be replicated and sent to the distribution hubs, which may or may not be possible depending on the agreement between the MSO and the content provider. In the event that distributing origin content is not possible, an alternative solution is to co-locate streaming media caching servers [5]. The caching servers act to proxy the user's requests, serving the information from local cache if possible and otherwise requesting the information from the content provider origin server. While the first user to view a streaming media object may experience the Internet bottleneck, subsequent requests for the same content would be served from local cache, just as if the origin server were co-located. It is even possible to eliminate the bottleneck for the first user by pre-loading content that is expected to be popular into the caches.

The upside to co-location is total control over the end-to-end streaming bandwidth, eliminating any Internet backbone bottlenecks. The downside to co-location is the cost for additional servers (origin or proxy) in each distribution hub and the content coherence problem for distributed origin servers.

Adaptive Streaming

Another method to ensure reasonable quality under bandwidth variations is to use adaptive streaming, which modifies the encoding/streaming rate in response to network conditions. This approach requires a feedback mechanism between the client and the server to exchange up-to-date information regarding the bandwidth or network conditions experienced by the client. The server uses this feedback information to adapt the streaming rate appropriately by either reducing the rate when network congestion increases or by increasing the rate when congestion clears, thereby achieving reasonable quality under dynamically varying network conditions.

One simple method of adaptive streaming is to keep a single encoded file at the server and drop frames to reduce the overall rate. This frame dropping technique is often referred to as "stream thinning." However, better quality can be achieved by keeping multiple bit rate encodings of the same stream with dynamic "up-shifting" or "down-shifting" between these encodings at key frames.

Examples of commercial products that support adaptive bit rate encodings are SureStream from Real Networks [6] and IntelliStream from Microsoft Windows Media [7]. Real Networks allows up to eight encoding rates in a single file with video window sizes and audio sample rates fixed for all bit rates.

The upside to adaptive streaming is that the streams can be adapted to changing bandwidth conditions to provide the optimal viewing experience. The downside of adaptive streaming is that the streams are not optimally encoded for a given bit rate since some of the encoding parameters are kept constant over all bit rates, and that encoded file sizes are increased for each encoding rate offered.

Access Limiting

The quality of streaming media applications suffers greatly when the available bandwidth for a session drops below the expected (encoding) rate. This would not be so bad if existing streaming media sessions were protected from new users consuming the last available bandwidth. Unfortunately, this is not the case. Rather, when a new user begins a session that consumes the last percentage of available downstream bandwidth, all existing sessions will suffer and begin "thrashing," or missing packets and sending out more re-transmit requests that further exacerbate the problem.

The solution to thrashing is to not allow the available bandwidth to be completely consumed. This requires changing the access policy so that rather than always granting access to new sessions, it is possible to reject or limit access when the available bandwidth drops below a certain "safe" level.

Limiting access without the cooperation of all applications using the downstream pipe is only a partial solution. This is the principal behind formal QoS solutions that will be addressed in the next section. Still, limiting access for a single class of applications, such as streaming media, can still offer a significant benefit to preserving quality.

A prerequisite to implementing access limiting is to utilize a single funneling device through which all streaming media sessions flow. This provides a single point for gathering information on bandwidth conditions and deciding which requests are

allowed or rejected. Devices which can play this role include a streaming server, proxy server, or intelligent router. In all cases, the device would be co-located close to the CMTS for gathering up-to-date information about the downstream bandwidth conditions through the CMTS's Network Management System (NMS) interface.

The upside to access limiting is the ability to limit the total number of streaming sessions or downstream bandwidth usage. This prevents thrashing and allows admitted users to preserve the quality of their sessions. The downside to access limiting without QoS support is the limitation to a single class of service, such as streaming, and the requirement of gathering real-time bandwidth information from the CMTS.

Excess Bandwidth Utilization

Many streaming applications are not constant with respect to their required bit rate. This provides an opportunity for creative ways for utilizing extra bandwidth that may be available at one moment to compensate for a reduction in available bandwidth at a later moment.

Skip Protection is a technology to improve the quality of playback at the client end. Skipping refers to pauses during playback caused by the need to re-buffer packets due to network congestion. Servers can utilize excess bandwidth available to buffer data faster than real-time (also referred to as "bursting") on the client machine. Thus if a larger buffer is available at the client, servers employing skip protection can utilize information about bandwidth conditions to fill up that buffer. Many servers including the QuickTime Streaming Server and the new Microsoft Windows Media Corona server now provide skip protection.

Forward Error Correction (FEC) techniques add redundancy to the original data stream so as to provide error resiliency to packet loss/corruption in the absence of a feedback channel between the client and the server. The DOCSIS physical layer uses FEC techniques to ensure reliable transmission over a noisy medium. However, it is also possible to use an application-level FEC mechanism to take advantage of excess bandwidth conditions. The Streaming Fountain product offered by Digital Fountain [8] utilizes an application-level FEC technique to encode excess information into the streaming packets. In the event that bandwidth is later constricted, the prior excess information can be used to re-construct lost packets without having to request retransmission.

The upside for excess bandwidth utilization is the ability to preserve the quality of streaming sessions in the presence of downstream bandwidth fluctuations. The downside of excess bandwidth utilization is the requirement of a client-side component to decode and utilize the extra bits being transmitted.

DOCSIS 1.1

Delivering applications with guaranteed quality of service (QoS) requires network-level components that provide end-to-end packet delivery with specified constraints, and QoS mapping to translate application-level quality of experience parameters to network-level QoS parameters.

QoS Network Components

In DOCSIS 1.0, all IP traffic from a single cable modem is grouped together under a single Service Identifier (SID). This means that all traffic types, including data, voice, and video, are treated equally by the Cable Modem (CM) and CMTS. DOCSIS 1.1 introduces the ability to separate different traffic types into different Service Flows and allow for different service parameters to be

applied to each of the flows. In addition to adding service flows, DOCSIS 1.1 introduces new components for service flow management, downstream packet classification, and dynamic MAC messages, which together provide the basis for true QoS capabilities.

Differentiated Services (diffserv) is a QoS mechanism for supporting a limited number of QoS behaviors and aggregating all possible flows into this smaller set of behaviors. Diffserv defines a set of per-hop behaviors (PHB) that are applied to packets as they move through diffserv capable routers, such as a DOCSIS 1.1 CMTS. Though PHB only defines behavior for a single router, it is possible to combine multiple routers with the same PHBs and apply admission control to limit the number of PHB packets entering the system, thereby achieving end-to-end QoS.

An MSO can control the PHBs for all routers within its domain, but to provide true end-to-end QoS that spans multiple Internet domains, the MSO needs to negotiate bilateral agreements at domain boundaries called Service Level Agreements (SLAs). The SLA defines how a PHB from one domain will be carried through another domain.

Now that it is possible to give some traffic preferential treatment over other traffic, a policy system is needed to decide which packets receive the preferential treatment at the expense of other packets. The policy components include a policy database for keeping track of all relevant information; a set of policy decision points (PDPs) for inspecting resource requests and accepting or rejecting them; and a set of policy enforcement points (PEPs), which enforce the decisions made by the PDPs. For MSOs, the DOCSIS 1.1 CMTS will serve as the PEP since it has ultimate control over all packets into or out of the DOCSIS network.

The CMTS can receive policy information in two different ways. The first is "configured QoS," where policy is specified in the form of static classification information (packet IP/port) mapped to corresponding PHBs. In this case, the CMTS acts as both the decision point by performing the classification and the enforcement point by applying the correct PHB. The second method to receive policy information is "signaled QoS," where policy information arrives dynamically in the form of ReSerVation Protocol (RSVP) messages. In this case, the CMTS extracts the resource request from the RSVP message and presents it to the specified PDP for classification. Signaled QoS provides direct feedback to hosts by either rejecting the RSVP message or by accepting the RSVP message, in which case the CMTS is automatically configured to classify and handle the appropriate traffic.

QoS Mapping

End-to-end QoS using diffserv components (PHB, SLA, PDP, PEP, etc.) provides proper end-to-end packet delivery. However, true QoS requires support up and down the protocol stack on each side as well. That is, the streaming media clients and servers need to be able to communicate their quality needs to the underlying QoS network that will deliver the packets. Translating QoS specifications between different levels of the protocol stack is called *QoS mapping*.

Streaming audio/video application users express quality of experience in terms of parameters such as frames per second, resolution, and sampling rate. In addition, highly interactive applications also include delay as an important aspect of the user experience. These application parameters must then be mapped into network parameters such as bandwidth, packet loss, packet latency, and packet delay variation (jitter). For video applications, bandwidth and packet loss are typically more important than latency and jitter. While bandwidth demand is

typically specified by the encoding rate, recent user studies have put absolute packet loss rates for VOD at 5% [9].

Once an application has determined how to map its quality of experience parameters onto network QoS parameters, it must employ a network API that allows these parameters to be specified. The latest version of the Microsoft Windows Socket (winsock2) API [10] allows for QoS parameters that utilize underlying QoS services from the operating system. This includes support for RSVP signaling, QoS policies, and invocation of traffic control over several protocol suites.

However, requiring an application to code directly to an underlying QoS network provides a dependency on the structure of that network. If the same application is to be executed atop various QoS networks, the dependencies become burdensome. To alleviate this mapping problem, the MPEG committee has defined an abstract QoS network within the Delivery Multimedia Integration Framework (DMIF) [11]. DMIF and its API (DAI) hide network-level details from the application programmer, including QoS signaling and transport mechanisms. DMIF-based QoS has already been studied for MPEG-4 streaming over IP networks with RSVP signaling and ATM networks with Q.2931 signaling [12].

Practical Improvements

Similar to the previous section that gave techniques to improve video quality in a non-QoS-enabled network, we now discuss how similar techniques can also be beneficial when DOCSIS 1.1 is available.

Co-locating video servers with the CMTS allows for end-to-end QoS during the transition period when Internet-wide QoS is not available but the user access network is QoS ready (DOCSIS 1.1 enabled). Even after Internet QoS is widely available, co-location allows for end-to-end QoS within the MSO domain, thereby removing the requirement of establishing and maintaining SLAs with Internet backbone providers.

Adaptive streaming techniques such as the availability of multiple bit rate encodings can be incorporated into QoS reservation decisions, thereby allowing for more choices. For example, if bandwidth reservation for the highest quality encoding fails, the server/client can re-negotiate with the QoS management for a lower bit rate version.

Access limiting is a fundamental aspect of any QoS enabled network that manages its resources for competing flows. However, unlike a DOCSIS 1.0 network where access limiting has to be explicitly introduced using specialized servers that regulate access, a QoS enabled network has native support for access limiting across all servers and application types. In addition, QoS-enabled networks allow advance reservations, which are not possible with simple access limiting.

Since QoS generally guarantees a certain constant bandwidth level for admitted applications, excess bandwidth utilization techniques would seem to offer little advantage. However, techniques such as skip protection with large client-side buffers can be used to reduce the dependence of proper QoS mapping for parameters like jitter. Excess bandwidth utilization techniques might also be used for variable bit rate encodings to reserve the “average” bandwidth requirement rather than the “peak” bandwidth requirement.

CONCLUSIONS

As DOCSIS networks continue to add subscribers and services, failure to preserve the quality of bandwidth critical applications will result in network congestion and poor user experience. However, there are several approaches for preserving the quality of streaming media in both a DOCSIS 1.0 and DOCSIS 1.1 network.

For DOCSIS 1.0, which does not support native QoS features, there are a number of techniques to preserve streaming media quality. Co-location places the entire end-to-end delivery route under the MSO domain, allowing for complete control over bandwidth policy decisions. Adaptive streaming and excess bandwidth utilization techniques try to preserve the optimal user experience under shifting bandwidth conditions. Access limiting provides a primitive level of QoS within a single application class, such as streaming media.

For DOCSIS 1.1, native QoS capabilities make it possible to offer differentiated service to streaming media applications, ensuring a certain level of bandwidth and latency tolerance. However, complete QoS requires support from all levels of the protocol stack as well as end-to-end network delivery. As these pieces start to unfold in a DOCSIS network and the wider Internet backbone, co-location, access limiting, adaptive streaming, and excess bandwidth utilization techniques can offer assistance in bridging the gaps and improving the overall user experience.

REFERENCES

1. *Cable Modem Market Stats and Projections*. Cable Datacom News. www.cabledatacomnews.com.
2. *Cable Data Network Architecture*. Cable Datacom News. www.cabledatacomnews.com.
3. *ISMA Specification*. Internet Streaming Media Alliance. www.isma.tv.
4. *Overview of the MPEG-4 Standard*. Moving Picture Experts Group. www.mpeg.telecomitalialab.com.
5. *Inktomi Traffic Server with Media IXT*. Inktomi Corporation. www.inktomi.com.
6. *SureStream*. Real Networks Corporation. www.real.com
7. *IntelliStream*. Microsoft Corporation. www.microsoft.com/windowsmedia.
8. *Streaming Fountain*. Digital Fountain Corporation. www.digitalfountain.com.
9. *QoS Requirements to Support Video and Audio Applications*. Dave Price. JANET QoS Workshop 2001.
10. *Microsoft Windows Quality of Service Platform Development*. Microsoft Corporation. www.microsoft.com/hwdev/tech/network/qos/default.asp
11. *An Overview of the Delivery Multimedia Integration Framework for Broadband Networks*. Jean-Francois Huard and George Tselikis. IEEE Communications Surveys 2(4), 1999.
12. *DMIF based QoS Management for MPEG-4 Multimedia Streaming: ATM and RSVP/IP Case Studies*. Victor Marques, Ricardo Cadime, Amaro de Sousa, and A. Oliveira Duarte. 3rd Conference on Telecommunications, April 2001.