Characteristics of Streaming Media Stored on the Web

MINGZHE LI, MARK CLAYPOOL, ROBERT KINICKI, and JAMES NICHOLS
Worcester Polytechnic Institute

Despite the growth in multimedia, there have been few studies that focus on characterizing streaming audio and video stored on the Web. This investigation used a customized Web crawler to traverse 17 million Web pages from diverse geographic locations and identify nearly 30,000 streaming audio and video clips available for analysis. Using custom-built extraction tools, these streaming media objects were analyzed to determine attributes such as media type, encoding format, playout duration, bitrate, resolution, and codec. The streaming media content encountered is dominated by proprietary audio and video formats with the top four commercial products being RealPlayer, Windows Media Player, MP3 and QuickTime. The distribution of the stored playout durations of streaming audio and video clips are long-tailed. More than half of the streaming media clips encountered are video, encoded primarily for broadband connections and at resolutions considerably smaller than the resolutions of typical monitors.

Categories and Subject Descriptors: C.2.0 [**Computer Communications Network**]: General

General Terms: Measurement, Documentation

Additional Key Words and Phrases: Apple QuickTime, long-tailed, Microsoft Windows Media Player, multimedia, RealNetworks RealPlayer, self-similarity, streaming

1. INTRODUCTION

Improvements in the connectivity levels of today's computers have enabled Web users who cross cultural and national boundaries to stream multimedia applications from far away Web servers to browsers on their desktops. Whether it is news, sports, or entertainment clips, the newest generation of Web users expect the convenience of being able to initiate audio and video streams by simply clicking on a browser link. In 2001, Real Networks estimated that 350,000 hours of online entertainment was being broadcast each week over the Internet, and this statistic does not include the volume of additional hours downloaded on-demand by Web users around the world.

Authors' address: M. Li, M. Claypool, R. Kinicki, J. Nichols, Computer Science Department, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609; email: {lmz,claypool,rek}@cs.wpi.edu.

CAIDA (Cooperative Association for Internet Data Analysis) emphasized in 2002 the significant fraction of Internet link capacities that were being allocated to support streaming media applications. Announcements such as RealNetworks' 2003 press release to support the advancement of streaming multimedia applications over wireless cellular networks add to the concern among Internet experts about the ability to support access to streaming media clips that are readily available on the Web. This anxiety over future streaming media applications significantly restricting performance for other Web users has translated into a variety of research papers that propose new network protocols [Floyd et al. 2000; Rejaie et al. 1999] or more sophisticated network router algorithms that seek to lessen the anticipated effect of streaming media [Mahajan et al. 2001; Feng et al. 2001; Cao et al. 2000; Stoica et al. 1998] on Internet performance. Several recent research efforts [Chesire et al. 2001; Wang et al. 2001; Li et al. 2002; Mena and Heidemann 2000; Veloso et al. 2002; Chung et al. 2003; Kuang and Williamson 2002a] have focused on capturing the characteristics of current streaming application behavior to better understand its impact. Only by knowing the relative frequency of commercial streaming products and how they typically stream multimedia traffic can researchers begin to prepare for the next generation of Web users.

Unfortunately, there is little recent published work on specific characteristics of streaming media clips stored on the Web. While there have been studies characterizing Web content measured at the client side [Bray 1996; Woodruff et al. 1996], there have been no recent studies of the general attributes of streaming media clips stored at Web servers. In 1997, Acharya and Smith [1998] studied video content stored on the Web by analyzing every video available in the (then popular) Alta Vista search engine. However, the nature of streaming media has changed considerably since that time. For example, Acharya and Smith [1998] found that the Internet could not support real-time streaming given the encoded bitrates and last-mile connection capacities available in 1997. Today, RealNetworks 'RealPlayer and Microsofts' Media Player, two popular streaming media products [Jupiter Media Metrix 2001] that did not even exist in 1997, RealNetworks' have significantly improved a Web user's ability to stream multimedia to home computers.

The papers by Ousterhout et al. [1985] and by Baker et al. [1991] proved to be influential in the design of new file systems and distributed file systems because they provided fundamental research on the nature of data stored in file systems and how these files were likely to be accessed. Accessibility to media clips on the Web through a variety of commercial streaming media products has reached such a state that similar studies on the characteristics of streaming media stored on the Web are needed to appreciate the future impact of millions of Web users around the world concurrently streaming freely available stored multimedia clips from remote Web servers to media clients in their homes.

This investigation built customized tools to address the following questions about the characteristics of streaming media content currently stored on the Web.

—*What are the most popular streaming media products used to store freely available audio and video on the Web?* Previous research [Li et al. 2002] has shown that proprietary encoded media products utilizing the same network bitrates differ in their impact on streaming network traffic performance. Similar to the situation in 1997 when the large user base for MPEG, AVI, and QuickTime was an obstacle for incoming streaming technologies, quantifying the current dominant technologies used to create streaming media clips can uncover new obstacles for future media applications.

—*What is the ratio of streaming audio clips to streaming video clips freely available on the Web?* The type of media, whether audio or video, stored on the Web gives researchers indications as to current users' bitrate expectations when streaming over the Internet. Streaming audio often requires only modest bitrates but typically has very discrete encoded bitrate levels. Video, on the other hand, is often bitrate-hungry and can stream over a wide range of encoded bitrates.

—*Are the media playout durations stored in media clips long-tailed?* Self-similar traffic is difficult to manage and there have been a number of studies of Internet traffic patterns that suggest self-similarity (see Park and Willinger [2000] for a survey). Long-tailed distributions of transfer times [Paxson and Floyd 1995; Willinger et al. 1995; Feldmann et al. 1995] may contribute to the self-similarity of Internet traffic. If the distribution of playout durations stored within media clips can be shown to be long-tailed, then this provides evidence to support the conjecture that the distribution of streamed media traffic on the Internet is self-similar.

—*What are typical streaming media target bitrates?* When encoded, streaming media clips use a target bitrate that has a direct impact on the network traffic rate the media will experience when streamed. Video target bitrates are influenced by such parameters as frame resolution, frame rates, and color depth. Knowledge of stored target bitrates provides insight into the strategies that media content providers use to deal with limited capacities encountered at last-mile connections.

—*What fraction of the streaming media codecs available are being used?* Innovative compression technologies in new codecs have the potential to deliver higher quality video with lower bitrates. Moreover, new codecs incorporate technologies that yield more sophisticated behaviors that adapt to network conditions to improve quality and performance. Understanding the percentage of older codecs that persist on the Web provides information as to the speed at which new codec technologies are deployed.

This article provides detailed information to answer these questions about streaming media stored on the Web today. Since commercial products by their sheer volume have had a strong influence on streaming traffic, our analysis focuses on commercial streaming products such as Microsoft's Media Player, Real Networks' RealPlayer, and Apple QuickTime. Unlike other measurement studies that have tended to view real streaming traffic by monitoring behavior near clients or servers [Chesire et al. 2001; Mena and Heidemann 2000;

Veloso et al. 2002; Merwe et al. 2002; Merwe et al. 2000], this investigation seeks the broder perspective of reviewing streaming content on media servers world-wide. While there is substantial audio and video content stored on peer-to-peer (p2p) file sharing systems [Saroiu et al. 2002; Saroiu et al. 2003], p2p content is typically not streamed at a target bitrate that takes into account viewer perceptual quality. P2p applications first download the streaming content as fast as capacity will allow and then subsequently play it out locally. Thus, network traversal behavior for p2p file sharing systems is similar to bulk file transfers and has quite different effects on the network than remotely accessed streaming media clips. Since this study focuses on the characteristics of streaming media that is played out in real-time, analysis of the content characteristics of audio and video stored on p2p systems is left outside the scope of this investigation.

A specialized Web crawler was built and launched from 17 carefully selected starting points across the Web. The crawler traversed over 17 million URLs with the objective of efficiently identifying unique URLs specific to streaming media. Other custom-built tools were used to extract information from nearly 30,000 media URLs and record media parameters that have previously been indicated as potentially impacting the perceived quality of media content streamed over the Internet. Analysis on the number of starting points and the number of URLs crawled from each starting point suggests that characterizations based on these 30,000 sampled clips are representative of streaming media stored on the Web at large.

The results of this data gathering indicate that the volume and relative amount of streaming media stored on the Web has increased significantly since 1997. Proprietary content is the most prevalent with RealNetworks and Microsoft Media having the most encoded media stored on the Web today. Most streaming media clips are relatively short. Application of proposed long-tailed distribution tests [Downey 2001] lends credence to the belief that stored streaming media playout durations are long-tailed. Analysis of the stored video clips shows that many videos on the Web are encoded for significantly lower resolution than can be supported by typical monitors. This suggests the potential for the Internet to see significant increases in video bitrates as last hop connections improve.

The results from this work can be useful when selecting representative streaming media clips for empirical Internet measurement studies that attempt to mimic the behavior of commercial media streaming traffic over the Internet, such as in Wang et al. [2001]; Li et al. [2002]; Chung et al. [2003]; Kuang and Williamson [2002a]; Wang et al. [2003]; and Kuang and Williamson [2002b]. Moreover, the results from this work can also be used to generate more detailed traffic models of streaming media for large scale Internet simulations.

This article is organized as follows. Section 2 discusses the crawling methodology used and the custom tools developed to measure the characteristics of streaming audio and video available on the Web. Section 3 analyzes the results of the crawler's search of the Web for stored streaming media and provides insight concerning the overall characteristics of streaming audio and video clips on the Web today. Section 4 discusses sampling issues related to this

investigation. Section 5 puts forth conclusions, and Section 6 proposes possible future work.

2. METHODOLOGY

The following methodology was used to collect extensive information on the nature of streaming media currently stored on the Web.

—Media Crawler, a customized Web crawler, was developed to search for and collect the URLs of freely available audio and video clips (see Section 2.1).

—A strategy was devised for selecting the starting points for initiating the Web Crawler such that a representative sample of available Web streaming audio and video clips could be obtained in a reasonably efficient manner (see Section 2.2).

—Tools were developed to extract the characteristics of the streaming audio and video content from the URLs collected by Media Crawler (see Section 2.3).

—Each of the streaming media clips in the complete set of unique streaming media URLs was started in order to perform packet header analysis and to record accessible attributes from the audio and video clips stored on the Web (see Section 3).

2.1 Media Crawler

To facilitate retaining the identity of audio and video URLs while crawling the Web, Larbin,[1] an open source, general purpose Web crawler, was modified to create *Media Crawler*. Starting from a specified root URL, Media Crawler recursively traverses embedded URLs and determines, based on protocol type, those URLs that refer to streaming audio and video content. For example, Microsoft Media Services (MMS) uses `mms://` as the protocol type and RealPlayer, QuickTime, and the newest version of Media Player use `rtsp://` to indicate that they are using RTSP, the Real Time Streaming Protocol.[2]

Due to current firewall restrictions [Merwe et al. 2002], audio and video are sometimes streamed over HTTP. Thus Media Crawler also examines URL extensions to find streaming media clips. Table I itemizes the set of URL extensions that Media Crawler uses as an indicator of streaming media content. This set of extensions was created by extracting the list of standard file type extensions that appear in file operation drop-down list boxes in most commercial media players.

Since the objective was to obtain a list of unique streaming media URLs and to avoid crawling loops, Media Crawler maintains a data structure that holds previously crawled URLs. Each time a new URL is reached, Media Crawler must search the data structure to determine if this new URL has already been encountered. Hence, the time to determine URL uniqueness grows with the number of previously identified unique URLs within a single crawl. This factor necessitated a strategy of launching Media Crawler serially from multiple

[1]http://larbin.sourceforge.net
[2]http://www.rtsp.org/

Table I. Audio and Video URL Extensions

Media Type	Extension
AVI	.avi
AU	.au, .snd
MP3	.mp3, .m3u
MPEG	.mp(e)g, .mpv, .mps, .mpe, .m2v, .m1v
MPEG-4	.mp4, .m4e
MPEG Audio	.mpega, .mpa, .mp1, .mp2
QuickTime (QT)	.mov, .qt
Real Media (RM)	.ra, .rm, .ram, .rmvb, .smil
WAV	.wav
Windows Media (WM)	.asf, .asx, .wma, .wmv, .wax, .wvx

starting Web pages rather than crawling more extensively from a single starting point.

2.2 Starting Pages

The growth of streaming media over the Web is tightly coupled with the availability of high bitrate Internet connections. Consequently, in selecting multiple starting points for Media Crawler, the strategy was to pick Web pages that were both popular and likely to be accessed by well-connected users. Since another goal of this investigation was to not only consider stored streamed media readily accessed from clients in the US, starting points were chosen from Web sites hosted in the ten most-wired countries (excluding the US) based on a market analysis report on broadband penetration [Topic 2002]. This scheme provides a more representative set of characteristics for streaming media stored throughout the World Wide Web. Secondarily, geographically dispersed starting pages reduces the overlap in the search space among the individual crawl instances.

A report by Nielsen,[3] the television and Internet ratings company, was consulted to determine the top ten Web sites in each country and to guide the selection of crawler starting points both inside and outside of the US. In those cases where Nielsen provided no information about the most popular sites within a country, a popular domestic newspaper or news portal was selected as the starting page. Since the United States is the most wired country, seven US Web pages were included in the set of starting points. These seven Web pages were selected from the most popular Web sites that cut across the following specific Web page types: news, sports, entertainment, Internet portal, search engine, and streaming media technology. Table II lists, in alphabetical order and by country, the 17 starting Web pages used in this research. A discussion of the impact of the number and specific choices for starting locations on the statistical validity of the sample population is given in Section 4.

Beginning from distinct starting pages, each of the 17 Media Crawler instances search URLs until a threshold of one million unique URLs has been reached, where-upon an output file is created that lists all URLs that refer to streaming media objects. While Media Crawler records unique streaming

[3]http://www.nielsen-netratings.com/

Table II. Media Crawler Starting Pages

Domain	Starting Page	URL
Canada	Canadian Government	canada.gc.ca
China	Sina.com	sina.com.cn
France	Free.fr	free.fr
Germany	T-Online	t-online.de
Italy	Republica Daily	republica.it
Japan	NTT Communications	ntto.co.jp
Korea	Empas Search Engine	empas.com
Spain	Grupo Intercom	grupointercom.com
Taiwan	China Times	news.chinatimes.com
UK	British Telecom	bt.com
US	America Online	aol.com
US	Alta Vista	altavista.com/video
US	ESPN Sports	espn.com
US	Hollywood Online	hollywood.com
US	New York Times	times.com
US	RealNetworks	real.com
US	Windows Media Home	windowsmedia.com

media URLs within a single crawl, the output from the crawls will overlap and include the same streaming media URL on multiple files. Thus, a separate program was run to create the final set of unique streaming media URLs across the 17 one-million URL data sets. Section 3.1 discusses the amount of overlap in streaming media URLs between pairs of data sets.

An additional problem in gathering stored Web pages for this study is the fact that references to specific Web content can become invalid for many reasons including content relocation, content removal, content damage, server failure, routing failure, and other errors. To minimize the number of invalid URLs caused by relocation or removal of Web content, the second stage of this study that included starting the stream of each of the available streaming media clips was conducted less than 24 hours after the final set of unique streaming media media URLs was produced.

2.3 Measurement of Content Characteristics

Once the set of unique valid media URLs was obtained, the next step was to use specialized tools to individually access each of the media content objects to collect from the audio and video clips information that included encoding format, target bitrate, playout duration, frame size, codec type, and other relevant properties. To automate this data gathering process, customized tools were built from a variety of commercial application Software Development Kits (SDKs), open source programs, and custom built components.[4]

Two new tools were used to analyze Real Media content. *RealAnalyzer* was custom-built using Microsoft Visual C++ and the RealNetworks SDK[5] provided

[4]The complete set of tools, including source code, can be downloaded from http://perform.wpi.edu/-downloads/#video-crawler.

[5]http://www.realnetworks.com/resources/sdk/index.html

by RealNetworks for customized RealPlayer development. The SDK comes with documentation, header files, and samples that expose the interfaces used in the RealPlayer streaming core and enable development of new tools and applications that can stream Real media. Real Analyzer gathers content description information such as URL, encoded bitrate, duration, resolution, live or pre-recorded, title, and copyright.

TestPlay, the second custom-built tool, gathers RealPlayer content statistics. An original version of TestPlay is available with the RealPlayer SDK under the directory `sdk/samples/intermed/testplay`. TestPlay allows the measurement of content encoding information including the number of sources, encoded bitrates, and codec information. With the modifications to RealAnalyzer and TestPlay that enable them to use a playlist of URLs, the combination of TestPlay and RealAnalyzer provides a means of automated measurement of the major characteristics of Real Media content.

To analyze Windows Media content, two tools similar to those for Real Media were built. The first custom tool, *Windows Media Analyzer*, uses Microsoft Visual C++ and the Windows Media Encoder 9 Series SDK[6] provided by Microsoft for customized Media Player development. Windows Media Analyzer gathers content information including-URL, encoded bitrate, duration, resolution, live or pre-recorded, title, and copyright. The second customized tool, *Wmprop*, extracts Windows Media Player content statistics. An original version of the tool is available with the Windows Media SDK under the directory `WMSDK/WMFSDK9/samples/`. Wmprop allows the measurement of content properties analogous to those recorded by TestPlay.

Finally, *MPlayer*,[7] an open source tool that runs on the Linux operating system, was used to analyze Apple QuickTime content. When playing QuickTime content, MPlayer produces resolution and codec information. However, MPlayer did not provide the encoded bitrate of QuickTime content.

3. ANALYSIS

The first phase of this investigation consisted of initiating 17 distinct Media Crawler runs from Worcester Polytechnic Institute (WPI)[8] between February 13, 2003 and March 18, 2003. Table II lists the individual starting points for each of the 17 Crawler instances. Each execution of Media Crawler searched the Web until one million distinct URLs were reached.[9] The total execution time for a Crawler instance depends upon the Web starting point. The crawl beginning from sina.com.cn in China, the starting point with the largest round-trip time from WPI,[10] took approximately 24 hours to traverse one million distinct URLs, while several of the crawls begun at closer Web sites only took about four hours to complete (see Li et al. [2003] for more details).

[6]http://www.microsoft.com/windows/windowsmedia/create.aspx

[7]http://www.mplayerhq.hu/homepage/design6/info.html

[8]WPI network configuration data can be found at http://www.wpi.edu/Admin/Netops/-infrastructure.html.

[9]The complete set of URLs obtained can be downloaded from http://perform.wpi.edu/downloads/-#video-crawler.

[10]WPI is physically located in Worcester, Massachusetts, USA.

Analysis of the data collected by Media Crawler is divided into four components, three of which are discussed in this section, while the fourth topic concerning the significance of our sample size is taken up in Section 4. The first stage, aggregate analysis, studies the clustering of multimedia URLs per server and presents information about the popularity of commercial streaming products. Using the customized tools described in Section 2, the second analysis stage collected content information from each of the available media clips to provide data on the relative content created by the major commercial streaming products and to test whether the audio and video playout durations are long-tailed. The third phase of the analysis drills down to a lower level and looks into encoded bitrates and other media clip attributes that impact streaming media transmission rates.

3.1 Aggregate Analysis

Prior to aggregate analysis, duplicate URLs from the 17 distinct 1-million URL data sets were removed, resulting in 11,533,849 unique URLs (see Li et al. [2003] for details on the overlap of URLs from each set). From the unique URLs, a set of 54,762 URLs were identified as streaming media by using standard indicators of media player types and the set of URL extensions described in Section 2.1 . In 1997, Acharya and Smith [1998] reported finding 22,600 media URLs out of 25 million Web pages [Sullivan] indexed by Alta Vista at that time. Thus, the percentage of audio and video objects stored on the Web has increased more than five-fold from about 0.09% in 1997 to about 0.47% in 2003. Moreover, given that the Google search engine currently indexes more than 3 billion Web pages,[11] one can make a rough estimate that nearly 15 million freely available streaming audio and video clips are stored on the Web today. Note, nothing in this investigation takes into account Web pages that a client pays to reach.

The complementary cumulative distribution function (CCDF) of the number of URLs found per server is given in Figure 1. The 11,533,849 million unique URLs came from 712,104 different Web servers and the median number of URLs per server over the total set of servers crawled is only one URL. The 54,762 unique audio and video URLs came from 4678 different servers, and the median number of URLs per server for the set of servers that had streaming media was 4. However, the graph indicates that about 1% of the streaming media servers provides 100 or more media URLs per server.

Figure 2 depicts the average percentage of URLs for each media type within a set of one million URLs, coming from one instance of Media Crawler. The average count (out of one million URLs) for each media type is indicated by the number above each bar. The error bars represent the standard deviation across the 17 sets of Media Crawler data. In the sampled population, Real Media accounts for almost half of all the streaming media URLs and more than doubles the count of Windows Media content. QuickTime, MPEG, and AVI, cited as the most popular video types in 1997 [Acharya and Smith 1998], make up only a combined 10% share of the multimedia content in 2003. MP3, a popular streaming audio format, is the most popular nonproprietary format

[11]http://www.google.com/, searching 3,307,998,701 Web pages as of December 18, 2003.

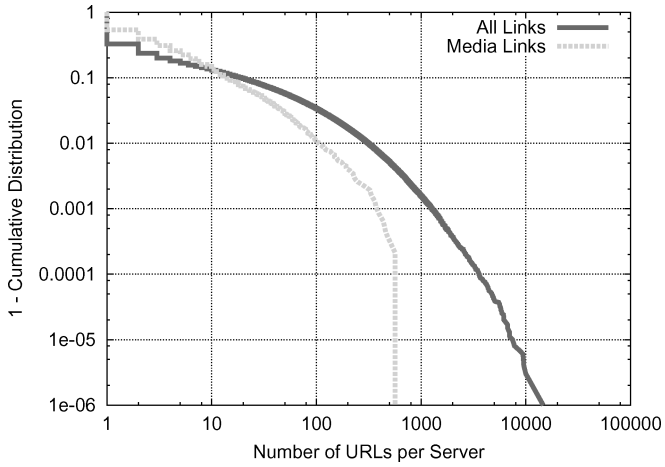

Fig. 1. URLs per Web server and media URLs per Web server with streaming media.

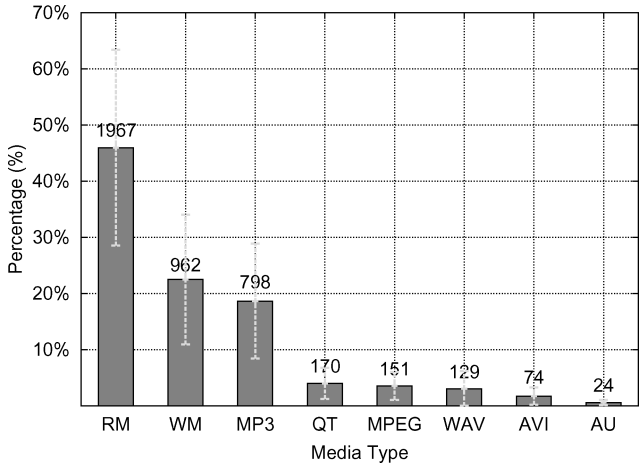

Fig. 2. Percentage of each media type.

in the sampled population, and MP3 is more prevalent than Apple QuickTime Media.

3.2 Commercial Product Analysis

The RealNetworks Real Media, Microsoft Windows Media, and Apple QuickTime Media commercial products account for about 72% of the URLs in the complete list of unique media URLs collected by Media Crawler. Given the dominance of these three products, the decision was made to focus further detailed analysis only on the characteristics of these three streaming products. Real Media, Windows Media, and Apple QuickTime Media support both audio and video, and they all can stream both prerecorded and live audio and video over the Internet.

Table III. Number of Streaming Media Clips Analyzed

Media Type	Audio	Video	Total	Percent
Real	9863	8504	18367	63
Windows	2591	6567	9158	32
QuickTime	28	1474	1502	5
Total	12482	16545	29027	100

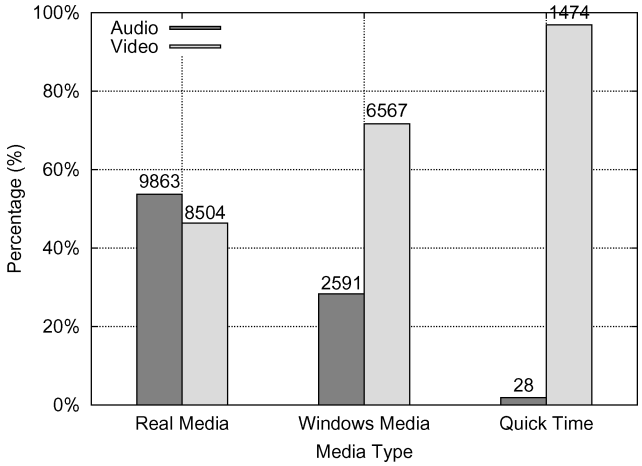

Fig. 3. Percentage of audio and video for each media type.

Of the 39,490 unique Real Media, Windows Media, and Apple QuickTime Media URLs recorded by Media Crawler, only 29,027 (about 74%) were available URLs. The remaining unique media URLs collected by Media Crawler were classified as unavailable when the data analysis phase was unable to reach a URL previously recorded by the crawler. Further analysis (see Li et al. [2003]) with our tools to determine why these clips were unavailable produced three primary reasons: "cannot find the specified file" (50% of errors), "cannot connect to the server" (25%), and "authorization failure" (10%). Table III shows a breakdown of the count of accessible streaming media clips. All subsequent analysis in this article is based on the data obtained from the 29,027 accessible multimedia URLs.

While in principle each media URL can be a playlist with many streaming media clip entries, the data analysis implies this occurs infrequently. Over 97% of the playlists refer to only one streaming media clip and only about 1% of the playlists refer to 3 or more streaming media clips (see Li et al. [2003] for detailed analysis on the playlists).

Figure 3 graphs the percentage of audio and video for each of the three major media types. Overall, 43% of the media clips are audio only. 54% of the Real Media clips are audio only. Combining information from Figure 2 and Figure 3, it is clear that in the collected URLs there is more Real Audio stored on the Web than MP3 audio. Comparatively, less than a third of Windows Media is audio only and virtually no Apple QuickTime is audio only. Due to the insignificant

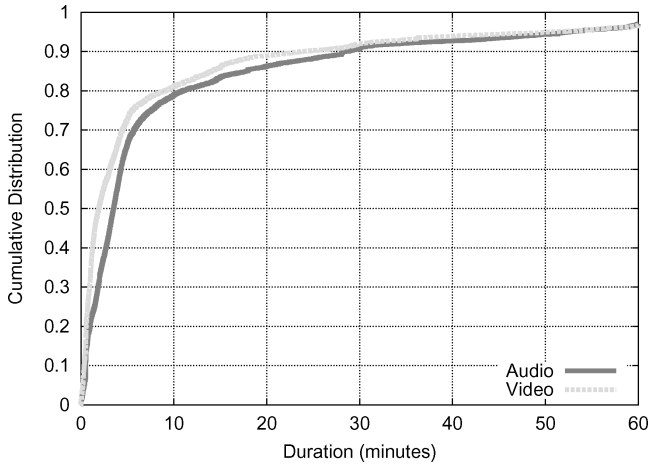

Fig. 4. CDF of streaming media duration.

amount of QuickTime audio, subsequent analysis considers only QuickTime video.

Our tools examine attributes in the streaming media header to determine if the media is live or prerecorded. For Windows Streaming Media, the types identified were broadcast, streamed, or downloaded. Broadcast indicates live streams while the other two types imply prerecorded content. For Real Media, the header indicates either live or prerecorded. For QuickTime, the duration is a very large (over 40 days) fixed integer for live media. While all three media formats support both live and prerecorded streaming content, 98% of the available streaming clips collected are prerecorded. In the sample population of available clips, about 2% of the Real Media clips were live; about 3% of the Windows Media clips were live; and less than 1% of the QuickTime clips were live.

During the duration analysis, three outlier clips with a duration of 10 days (roughly an order of magnitude longer than the next shortest streaming clip) were uncovered. Closer inspection revealed these clips were actually a form of streaming text (RealText) that continually looped the same short message. These three clips were removed from all subsequent analysis.

The CDF of the duration of available audio and video clips is presented in Figure 4. The main body of the distribution of audio and video durations are similar. Most stored audio and video clips are relatively brief. The median duration of all clips is about 3 minutes with the median for video and audio clips about 2 minutes and about 4 minutes, respectively. 10% of the audio and video clips have a duration of less than 30 seconds, while 10% have a duration over 30 minutes. This data indicates that the durations of videos stored on the Web today are significantly longer than in 1997 when 90% of video clips lasted 45 seconds or less [Acharya and Smith 1998]. However, the median clip duration in Figure 4 is much shorter than that of a typical TV program or movie. This suggests the potential for a major increase in clip durations in the near future.

Self-similar traffic is difficult to manage and a long-tailed distribution of network flow transfer times may contribute to the self-similarity of Internet traffic.

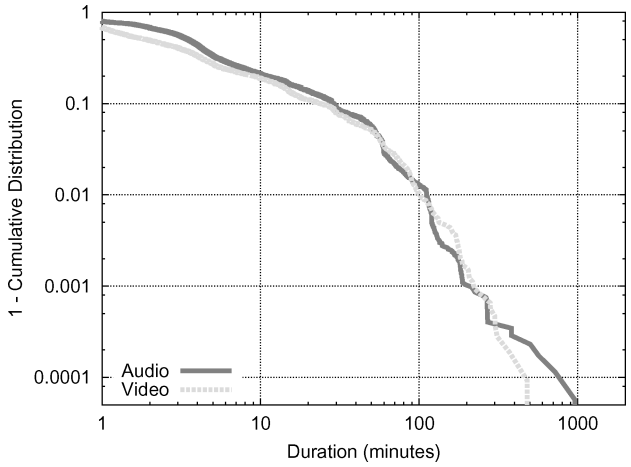

Fig. 5. CCDF of streaming media duration.

If the distribution of the durations stored in streaming clips is long-tailed, it increases the likelihood that the distribution of transfer times for prerecorded streaming transfers over the Internet are also long-tailed. Note, this discussion excludes live streaming events that have an undetermined duration. To allow for clearer examination of the distribution tails, Figure 5 graphs the CCDF of the stored audio and video duration distributions.

The definitive test for a long-tailed distribution is that the steepness of the slope in the CCDF does not increase in the extreme tail but continues with constant slope (the line may become jagged as the number of samples becomes sparse, but the slope stays the same). Visual inspection of the duration distributions in Figure 5 implies that the durations of the stored audio and video clips may be long-tailed. However, as discussed by Downey [2001], certain distributions, such as lognormal, appear visually similar to long-tailed when, in fact, they are not. The characteristic difference between a long-tailed distribution and one that is not long-tailed is the curvature. A long-tailed distribution does not have a curved tail. To determine whether the distribution of durations for the streaming media URLs collected in this study is long-tailed, the curvature test proposed by Downey [2001] was applied. The details from the five steps in this process include the following.

(1) Measure the curvature of the tail of the sample distribution, where the tail is defined as $P(X > x) < 1/16$.[12] Curvature is quantified by taking three-point estimates of the first derivative and fitting a line to the estimated derivative. For the crawled media clips, the curvature of the audio distribution tail is 0.0378 and the curvature of the video distribution tail is 0.0505.

(2) Estimate the Pareto slope parameter, α, that best models the tail behavior of the sample using a program developed by Crovella and Taqqu [1999] called

[12]We also tested $P(X > x) < 1/32$ and $P(X > x) < 1/64$ and our overall results were the same.

aest.[13] For the media clips, the estimate of α given by aest is 1.006975 for the audio distribution and 1.000161 for the video distribution.

(3) Generate 1000 samples from a Pareto distribution with slope parameter α, where each Pareto sample has the same number of points that are in the data sample, n, and calculate μ, the mean curvature of the 1000 samples. There are $n = 11,836$ samples in the audio distribution with $\mu = 0.004845$, and there are $n = 16,282$ samples in the video distribution with $\mu = 0.003722$.

(4) Calculate d, the difference between the curvature of the original sample and μ. For the set of crawled media clips, the audio distribution curvature differs from μ by 0.032958, while the video distribution curvature differs from μ by 0.046778.

(5) Count the number of samples out of 1000 that have a curvature that differs from μ by as much as d. This count is the p-value for the null hypothesis that the samples come from a long-tailed distribution. For the audio durations, 498 differ from μ by d or more, and the p-value is 0.498. For the video durations, 495 differ from μ by d or more, and the p-value is 0.495.

Thus, the relatively high p-values in step 5 means the null hypothesis that the samples come from a long-tailed distribution cannot be rejected. This implies that the distribution of playout durations for the streaming media clips encountered by Media Crawler may be long-tailed.

3.2.1 *Video.* Video can operate over a wide range of bitrates. Based on H.261 and MPEG-4, MPEG-1 and MPEG-2 standards, video conferences and low-bitrate videos stream at about 0.1 Mbps; VCR quality videos stream at about 1.2 Mbps; broadcast quality videos stream at about 2–4 Mbps; studio quality videos stream at about 3–6 Mbps; and HDTV quality videos stream at about 25–34 Mbps. Uncompressed video can require hundreds and even thousands of Mbps. Thus, video applications potentially can demand enormous streaming data rates that are greater than the available network capacity.

Figure 6 provides CDFs for the encoded video bitrates for Windows Media and Real Media (as explained in Section 2, QuickTime Media encoding rates could not be captured). The median encoded bitrate is around 200 Kbps with the Windows Media median encoded bitrate slightly higher than the median encoded bitrate for Real Media. Approximately 29% of the videos are encoded to stream over a 56 Kbps modem. This is a substantial increase from 1997 [Acharya and Smith 1998] when fewer than 1% of videos were encoded for modem bitrates. Nearly 70% of the videos are targeted for broadband (56 Kbps–768 Kbps). This is appreciably higher than the 50% recorded in 1997. Approximately 1% of the videos have bitrate targets above typical broadband connections (768 Kbps–1500 Kbps), and less than 1% have bitrate targets above T1 rates (1540 Kbps). These encoded bitrate percentages are down dramatically from about 20% in 1997.

[13]Downloadable from http://www.cs.bu.edu/faculty/crovella/aest.html.

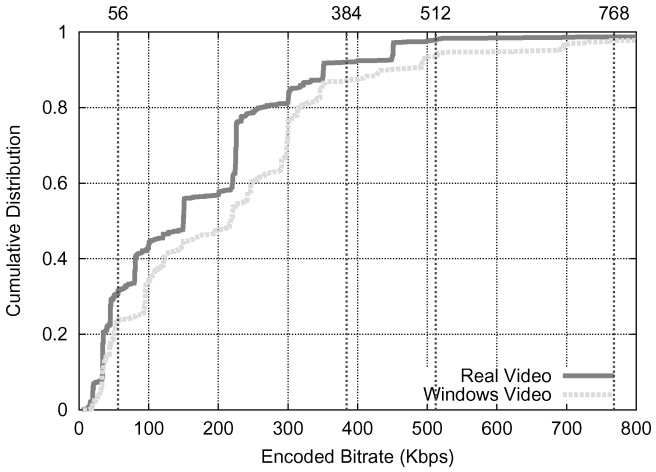

Fig. 6. CDF of streaming video encoded bitrate.

The general shift in target encoded bitrates that includes a larger percentage of streaming videos targeted towards lower bitrates even while end host bitrates have increased, suggests that improvements in streaming technologies make it possible to effectively send streams at lower bitrates. The predominance of videos targeted towards broadband connections suggests end users in the home are the typical target audience and that encoded bitrates will increase as last-mile home connections increase.

Techniques where multiple target bitrates are encoded into one video (such as with Windows Media "Intelligent Streaming" and RealNetworks "SureStream") are designed to provide better quality when a streaming media server scales down due to the bitrate restrictions and network congestion. A typical video stream will have two encoded streams, one for the video and one for the audio. If there are more than three streams in one clip, the assumption is that this clip has multiple encoded bitrate levels. Only our customized Windows Media tool was able to determine the number of encoded bitrate levels. Figure 7 depicts the cumulative distribution of the number of encoded streams per Windows Media clip for the clips in the sample population. From the measurements, one sees that approximately 12.1% of the Windows Media clips have multiple bitrate encoding levels. The lack of encoded bitrate choices for a media server has important ramifications on network quality of service. If these videos are streamed over UDP during constrained bitrate conditions, their lack of scaling options implies these multimedia flows will be unfair and severely affect competing TCP traffic. Note the distribution of Windows Media encoded bitrate levels in Figure 7 is in direct contrast to previously reported results for Real Media in Chung et al. [2003] where 65% of the Real Video clips had multiple encoded bitrate levels.

Figure 8 focuses on the CDFs of the video clip resolutions. The resolutions shown were obtained by multiplying frame width by frame height for each video clip. Approximately 70% of the videos have a standard aspect ratio of 4/3. The remaining 30% of the video clips have aspect ratios slightly above and

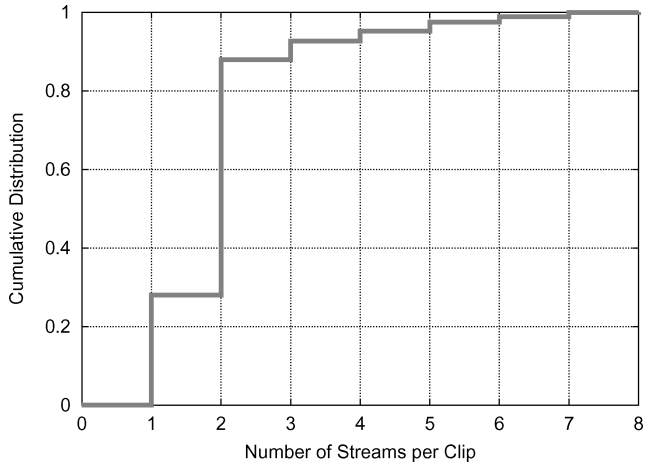

Fig. 7. CDF of number of Windows Media streams encoded in one clip.

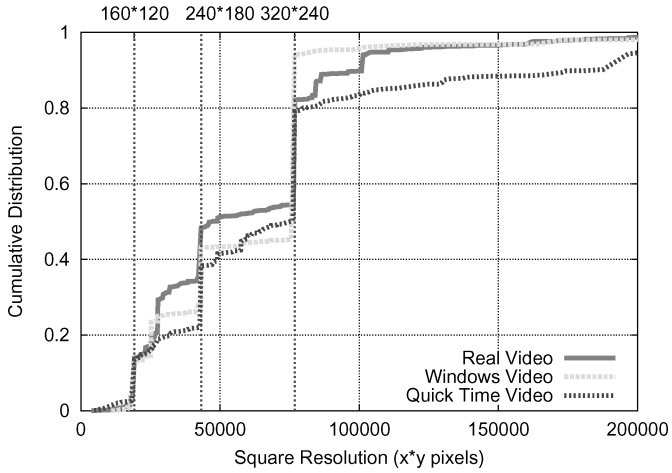

Fig. 8. CDF of video resolution (length × width).

slightly below 1.3 (see Li et al. [2003] for more details). The vertical lines in Figure 8 indicate commonly used video resolutions: 160 × 120 (quarter-screen), 240 × 180 (three eighths-screen), and 320 × 240 (half-screen). The steps in the distributions correspond roughly to different resolution choices available in commercial media encoding products. Commercial media encoding applications provide default choices for resolution and other encoding parameters that are typically guided by common practices.

Nearly half of the videos in Figure 8 have less than half-screen resolution and less than 1% of the videos provide full-screen resolution. These small window sizes relative to the resolutions of typical desktop monitors are likely chosen because of the relationship between resolution and required streaming bitrate. A video with a resolution of 320 × 240 will typically result in bitrates on the order of hundreds of Kbps (the target bitrates shown in Figure 6). Given current

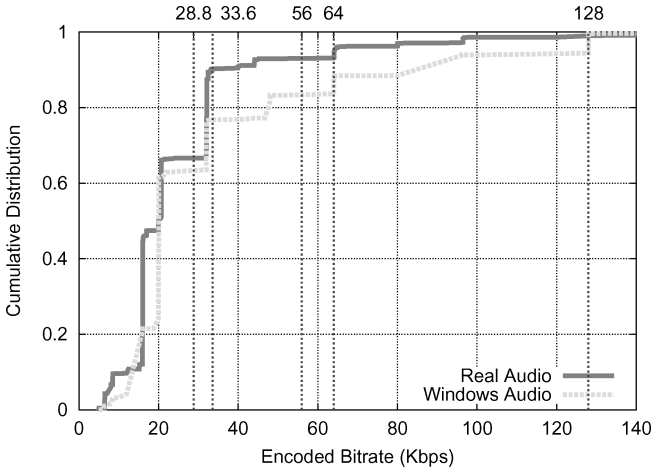

Fig. 9. CDF of streaming audio encoded bitrate.

typical desktop resolutions of at least 640 × 480 coupled with continual end-user demand for higher quality video, there is enormous potential for increasing the sizes of today's streaming video frames. Additionally, future advances in codec compression algorithms will facilitate larger frame sizes for the same encoding rates. One can also expect improvements in network bitrates to provide increased available bitrates to streaming flows. This implies future streaming traffic with larger frame sizes and higher bitrate demands on the Internet.

3.2.2 *Audio.* Figure 9 presents the CDFs for the encoded bitrates of the streaming audio clips for both Windows Media and Real Media. The streaming audio encoded bitrates are low compared with the encoded bitrates for streaming video shown in Figure 6. About 90% of streaming audio shown is targeted for modems, and the median encoded audio bitrate is suitable for streaming over older 28.8 Kbps modems. In 1999, an empirical study of streaming audio at a popular Internet audio server [Mena and Heidemann 2000] found 100% of the playout rates targeted at modem bitrates. Approximately 10% of the streaming audio in Figure 9 is specifically targeted at users with broadband or higher connections. Given that playout of CD quality audio requires hundreds of Kbps, it is likely that the fraction of high streaming audio encoded bitrates will increase. However, given the compression rates and listening quality of technologies such as MP3 (which typically streams at 128 Kbps), it is unlikely that audio encoding bitrates will increase above those required by broadband connections.

3.3 Media Codec

The codec has a large impact on the network performance of streaming media. For example, as an improvement to the Windows Media video version 8 codec (WMv8), version 9 supports fast streaming to smooth out changes in the available bitrate during streaming. While beneficial to users, the network impact of newer codecs is not always positive. For example, WMv8 fills the playout buffer

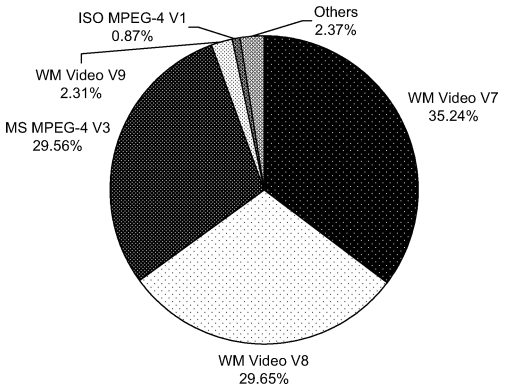

Fig. 10. Breakdown of Windows video codecs.

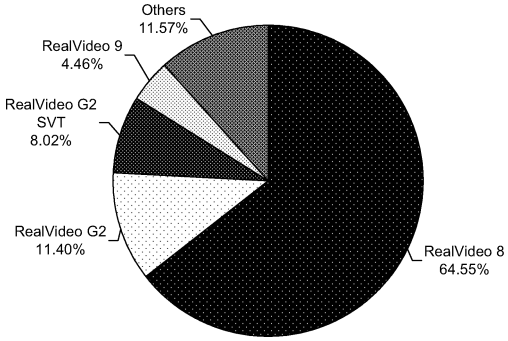

Fig. 11. Breakdown of Real video codecs.

at the target playout rate [Li et al. 2002], while WMv9, in a manner similar to RealPlayer [Chung et al. 2003], buffers at a significantly higher data rate.

Figure 10 and Figure 11 capture the breakdown of the codecs used to create Windows and RealNetworks streaming videos in the set of clips gathered by the crawler. The actual share of codec space occupied by a specific codec implementation in Figure 10 is not particularly significant except as a snapshot in time (e.g., WMv9 had only 2.31% of the recorded codecs in May 2003). However, future studies may find this data valuable in tracking the acceptability and change in market penetration over time of innovations such as WMv9.

Figure 10 shows the prevalence of different versions of the codecs for Windows Media video. Of the codecs shown, MS MPEG-4 v3 and WM Video 7 are the oldest. The latter is Microsoft's implementation of the MPEG-4 standard which is similar to the H.263 standard. MS MPEG-4, renamed WM Video 7 and released in May 2001, uses discrete cosine transform and motion prediction to encode and compress video content. WM Video 8, released soon after in September 2001, was the only Microsoft codec until the most recent version, 9, was released in January 2003.

Figure 11 shows the distribution of the different versions of the Real Video codec. RealVideo 8 dominates in the space of codecs that operate with

RealPlayer. Similar to WMv9, Real Video 9 is still not yet deployed in significant numbers relative to RealVideo 8.

4. SAMPLING ISSUES

In collecting data for large scale measurement studies on the Web, there are important issues related to the number of samples compared to the size of the overall population. In 1997, researchers were able to locate and download all videos found on the Web [Acharya and Smith 1998], but today that is impractical. Crawling the 17 million URLs and analyzing the media clips used in this study took over one month. At this pace, it would take more than 16 years to crawl over the roughly 3 billion pages currently on the Web and analyze the media clips.

This section considers issues related to the sampling and data gathering approach used in searching 17 million URLs with Media Crawler. To ascertain whether this set of URLs is an adequate sampling of the Web, the strategy was to evaluate the effects of smaller sample sizes on the quality of the resultant analysis. Four specific questions were considered.

—Is it possible to obtain a sufficiently large number of samples with fewer crawler starting points?
—Is it possible to obtain a sufficiently large number of samples while searching fewer than one million unique URLs per crawl instance?
—How does the sampling in terms of the number of URLs and the number of starting points affect the overall distribution shapes?
—How does the choice of starting points in terms of different cultural locations affect the overall distribution shapes?

For each of the 17 crawler starting points, virtual experiments with fewer than one million URLs were considered. Five separate data-gathering plateaus were reviewed, beginning with 200,000 URLs and proceeding in increments of 200,000 URLs, up to the full one million URLs. So the smallest data set had 3.4 million URLs ($17 \times 200,000$), and each subsequent data set increased by 3.4 million until the full 17 million URL set was reached.

Figure 12 demonstrates that, at the 10.2 million URL plateau and beyond, all the percentages for the various media product types remain constant. Given that the number of media URLs encountered in each of the 3.4 million URL data sets are approximately the same (see Li et al. [2003]), the data suggests that, at least for this statistic, crawling beyond 17 million URLs is not likely to change the results. Data on the absolute number of media URLs found as the crawler reaches the five plateaus yields very similar results.

To drill down further, the impact of the data set size on the distribution of several important media clip characteristics was also analyzed. Figure 13 presents five CDFs of video playout duration. Each CDF is for one crawler plateau from 3.4 million to 17 million URLs. The similarity in the distribution of video playout durations further suggests that there is little quantitative benefit in the reliability of the CDF to be gained by crawling longer to find larger sets of unique Web URLs.

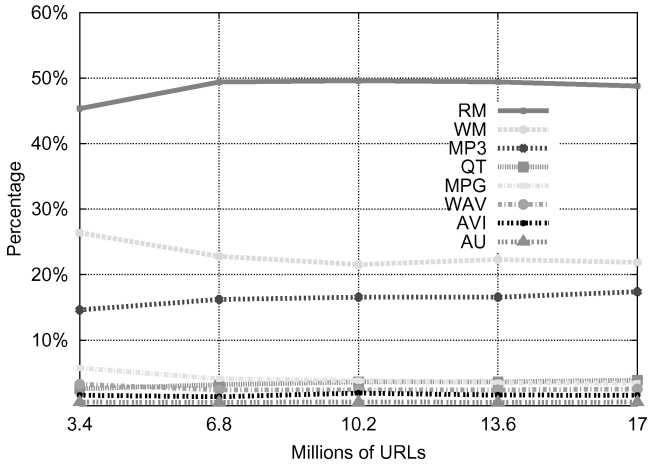

Fig. 12. Percentage of media types versus the number of URLs crawled.

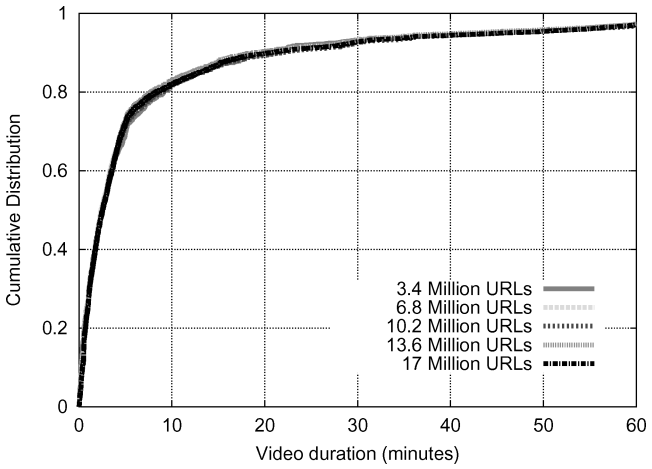

Fig. 13. CDF of video duration for different sample set sizes.

The next issue considered was whether the number of starting points would have a significant effect on the results obtained. From the 17 original starting points, data from 5 separate subsets of randomly picked starting points were analyzed. In this case, all 1 million URLs from each of 3, 6, 9, 12, and 15 randomly selected starting points were evaluated with respect to the media composition and the video playout duration distributions. Figure 14 depicts the video composition versus the number of starting points. For sets with 9 or more starting points, the percentage of each media type remains relatively constant. Given that the number of media URLs crawled is approximately the same for each group of three starting points (see Li et al. [2003]), crawling from a larger number of starting points is not likely to change the results.

Figure 15 graphs the video playout duration CDFs for the same starting point subsets used in Figure 14. The playout distributions are similar for all

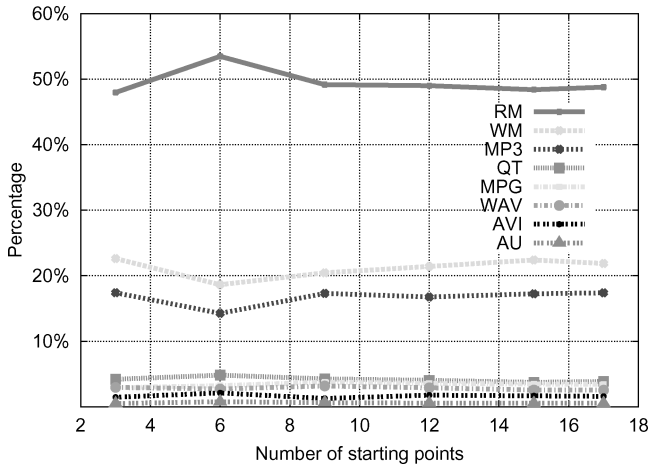

Fig. 14. Percentage of media types versus the number of starting points.

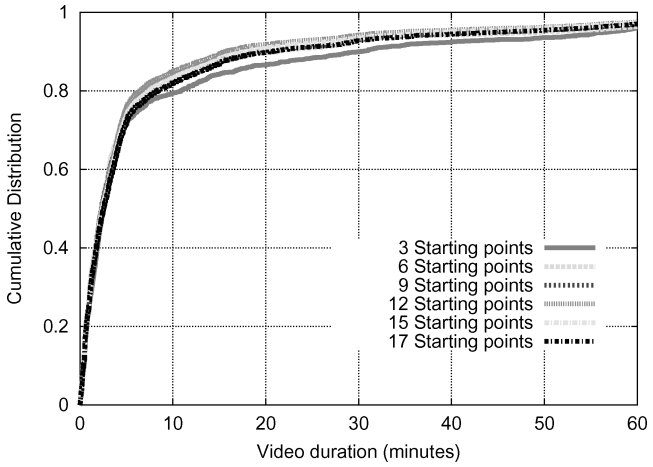

Fig. 15. CDF of video duration for different numbers of starting points.

numbers of starting points except for a slight separation for the distributions having only 3 starting points. This suggests that having more than 6 starting points will not significantly change the nature and shape of the CDF.

To ascertain the effects of different cultural starting points, the URL data was divided into the set of URLs obtained by beginning the crawl from the 7 US starting points and the set of URLs obtained from the 10 starting points outside the US. Figure 16 depicts the composition of media URLs for each data set and Figure 17 depicts the duration distribution of video playouts for each data set. While the composition and playout durations are nearly the same for each data set, there are some slight differences. For example, the US starting points have slightly more Windows Media clips and fewer MP3 clips, but they have an equivalent percentage of Real Media clips.

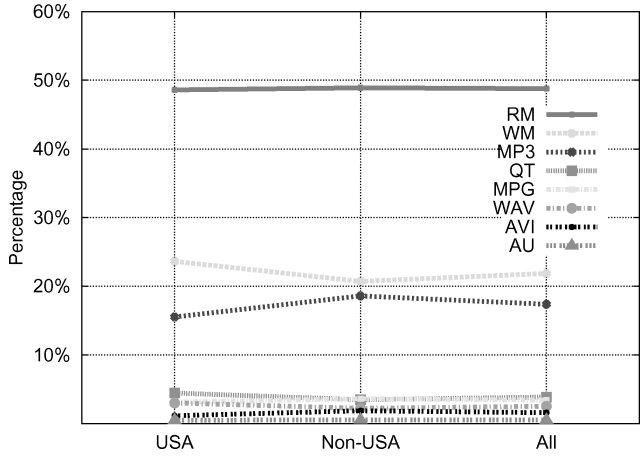

Fig. 16. Media types of US and Non-US starting points.

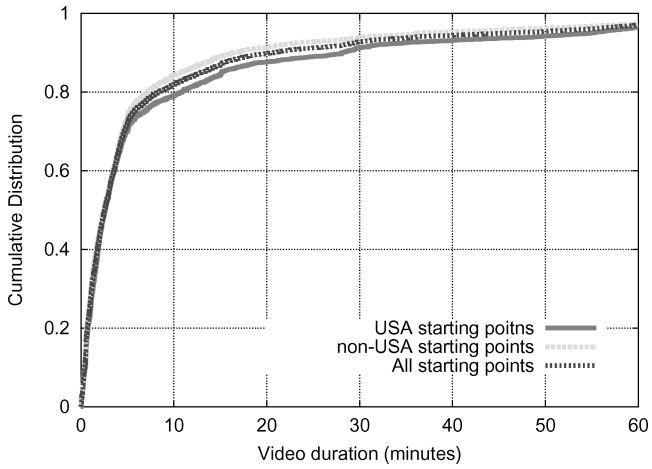

Fig. 17. CDF of video duration of US and Non-US starting points.

Combining the analysis of the number of URLs, the number of starting points, and cultural locations, one could argue that using nine or more starting points with 600,000 URLs per starting point provides a sample space large enough to effectively represent the streaming media stored on the entire Web.

5. CONCLUSIONS

Many researchers worry about the anticipated large increase in the volume of streaming media that will be sent over the Internet in the near future. Without data on the current state of streaming media available on Web pages, it becomes difficult for streaming media performance experts to predict both the short-term and the long-term impact of this expected increase in network traffic on the state of the Internet. Assumptions are often made in network models about the nature of streaming media traffic based on studies that are several years

old. However, significant changes in user access capabilities and improvements in the techniques employed by commercial media players make it risky to use outdated characterizations to represent the current behavior of audio and video Internet traffic.

The goal of this research is to provide the results of extensive data collection of streaming media content available across the Web. Armed with custom-built media player analysis tools, 17 million Web URLs were crawled and checked for availability to yield nearly 30,000 unique audio and video clips. In depth analysis of the freely-accessible stored clips was carried out by downloading initial segments of each clip to extract header and other useful characteristics about the media clips. These downloads originated from 4080 distinct media servers on which audio and video clips were located.

By comparing work in past studies, we find that the total volume of streaming media stored on the Web has increased by over 600% in the past five years. Moreover, the fraction of streaming media objects stored on the Web relative to other objects has increased by more than 500%.

The aggregate data analysis shows streaming audio and video content is dominated by proprietary streaming products, specifically RealNetworks Media is the most widely used with Microsoft Media second. There are relatively the same number of freely available audio clips compared to video clips. Given that video availability is likely to be more constrained than audio availability because last-mile connections are not (yet) all broadband, one should expect a shift in the future towards higher numbers of video sites relative to audio sites storing multimedia on the Web. The vast majority of streaming audio and video URLs are prerecorded, with only a very small fraction live. Most stored streaming media clips are relatively brief, lasting several minutes for both audio or video. However, the 3-minute median duration time is substantially longer than in 1997 when typical video clips were under 1 minute in length.

Despite the growth of broadband connections, the fact that the majority of audio encoded bitrates are still targeted to be acceptable for modem connections is significant. Moreover, the distribution of video bitrates implies that modems can also be used for streaming some video clips. The ability to have streaming content suitable for modems is a useful niche given that it is estimated that half of all US Internet subscribers will still use modems by the year 2005 [Brown 2001]. However, the majority of video target bitrates are broadband. Since current video resolutions used by servers are small relative to typical monitor resolutions, it can be expected that as network bottleneck bandwidths increase, video target bitrates will rise proportionally.

The data in this investigation indicates that current media providers tend to adhere to standard picture dimensions (such as 320×240) and aspect ratios (such as 4/3) when creating videos. There are similar steps in the distribution of audio encoding rates along typical encoding standards.

6. FUTURE WORK

While the advertised target streaming bitrates presented in this report provide insight as to possible network impact, the actual streaming rates experienced

by media flows sent over the Internet are likely to be quite different. The level of responsiveness of streaming media flows to Internet congestion and perceived available bitrate is expected to have a large impact on future network performance. Technologies such as Windows Media intelligent streaming and RealNetworks SureStream can take advantage of multiple target bitrates stored in a single media object. Previous work [Chung et al. 2003] suggests that such multiple bitrate technologies occur in many video clips, and media players can effectively choose the most effective bitrate to use in response to current network conditions. Thus, one valuable extension of this work could involve devising a technique to determine bitrate levels for stored streaming media clips. A more difficult challenge is to determine these bitrate levels and how they should be used under network congestion.

While the results of this study include details on the storage of audio and video on the Web, they do not provide information on the actual streaming of the stored audio and video over the Internet. Future work could complement these results with measurements of actual streaming behavior. Such efforts would be especially useful if a media server with many audio and video encoding rates and choices were specifically studied. Existing techniques that actively query DNS caches such as in Wills et al. [2003] could provide complementary information about the popularity of Web sites with stored audio and video.

Our crawling methodology is specifically targeted at locating and analyzing streaming media, namely, media that is played as it is sent over the network and not completely downloaded ahead of time before playing. There is also considerable audio and video content available on peer-to-peer file sharing systems. Tools to crawl peer-to-peer file sharing systems and analyze the multimedia content found may provide valuable insights into the use and support of such file sharing systems.

REFERENCES

ACHARYA, S. AND SMITH, B. 1998. An experiment to characterize videos stored on the Web. In *Proceedings of the ACM/SPIE Multimedia Computing and Networking (MMCN)*. San Jose, CA. 166–178.

BAKER, M., HARTMAN, J., KUPFER, M., SHIRRIFF, K., AND OUSTERHOUT, J. 1991. Measurements of a distributed file system. In *Proceedings of the 13th Symposium on Operating System Principles (SOSP)*. Pacific Grove, CA. 198–212.

BRAY, T. 1996. Measuring the Web. In *Proceedings of the 4th International World Wide Web Conference*. Paris, France, 994–1005.

BROWN, E. S. 2001. Broadband walks the last mile. *Tech. Rev*. (Online). Available at http://www.technologyreview.com/articles/print_version/brown060501.asp.

CAIDA (Cooperative Association for Internet Data Analysis). 2000. www.caida.org.

CAO, Z., WANG, Z., AND ZEGURA, E. 2000. Rainbow fair queuing: Fair bandwidth sharing without per-flow state. In *Proceedings of IEEE Infocom*. Tel-Aviv, Israel, 922–931.

CHESIRE, M., WOLMAN, A., VOELKER, G., AND LEVY, H. 2001. Measurement and analysis of a streaming media workload. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS)*. San Francisco, CA. 1–12.

CHUNG, J., CLAYPOOL, M., AND ZHU, Y. 2003. Measurement of the congestion responsiveness of RealPlayer streaming video over UDP. In *Proceedings of the Packet Video Workshop (PV)*. Nantes, France.

CROVELLA, M. E. AND TAQQU, M. S. 1999. Estimating the heavy tail index from scaling properties. *Methodol. Comput. Appl. Probab. 1*, 1, 55–79.

DOWNEY, A. B. 2001. Evidence for long-tailed distributions in the Internet. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*. San Francisco, CA. 229–241.

FELDMANN, A., GILBERT, A., HUANG, P., AND WILLINGER, W. 1995. Dynamics of IP traffic: A study of the role of variability and the impact of control. In *Proceedings of ACM SIGCOMM*. Cambridge, MA. 301–313.

FENG, W., KANDLUR, D., SAHA, D., AND SHIN, K. 2001. Stochastic fair blue: A queue management algorithm for enforcing fairness. In *Proceedings of IEEE Infocom*. Anchorage, AK. 1520–1529.

FLOYD, S., HANDLEY, M., PADHYE, J., AND WIDMER, J. 2000. Equation-based congestion control for unicast applications. In *Proceedings of ACM SIGCOMM Conference*. Stockholm, Sweden, 43–56.

FOR INTERNET DATA ANALYSIS (CAIDA), C. A. 2002. Characterization of Internet traffic loads, segregated by application (Online). Available at http://www.caida.org/analysis/workload/byapplication/.

JUPITER MEDIA METRIX. 2001. Users of media player applications increased 33 percent since last year. Press Release. Available at http://www.jup.com/company/pressrelease-.jsp?doc=pr01040.

KUANG, T. AND WILLIAMSON, C. 2002a. A measurement study of RealMedia audio/video streaming traffic. In *Proceedings of ITCOM*. Boston, MA. 68–79.

KUANG, T. AND WILLIAMSON, C. 2002b. RealMedia streaming performance on an IEEE 802.11b wireless LAN. In *Proceedings of IASTED Wireless and Optical Communications (WOC)*. 306–311.

LI, M., CLAYPOOL, M., AND KINICKI, R. 2002. MediaPlayer versus RealPlayer—A comparison of network turbulence. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop (IMW)*. Marseille, France, 131–136.

LI, M., CLAYPOOL, M., KINICKI, R., AND NICHOLS, J. 2003. Characteristics of streaming media stored on the web. Tech. Rep. WPI-CS-TR-03-18, CS Department, Worcester Polytechnic Institute (May).

MAHAJAN, R., FLOYD, S., AND WETHERALL, D. 2001. Controlling high-bandwidth flows at the congested routers. In *Proceedings of the 9th International Conference on Network Protocols (ICNP)*. Mission Inn, Riverside, CA. 192–201.

MENA, A. AND HEIDEMANN, J. 2000. An empirical study of real audio traffic. In *Proceedings of IEEE Infocom*. Tel-Aviv, Israel, 101–110.

MERWE, J. V. D., CACERES, R., HUA CHU, Y., AND SREENAN, C. 2000. mmdump—A tool for monitoring Internet multimedia traffic. *ACM Comput. Comm. Rev. 30*, 5 (Oct.), 48–59.

MERWE, J. V. D., SEN, S., AND KALMANEK, C. 2002. Streaming video traffic: Characterization and network impact. In *Proceedings of the 7th International Workshop on Web Content Caching and Distribution*. Boulder, CO.

OUSTERHOUT, J., DACOSTA, H., HARRISON, D., KUNZE, J., KUPFER, M., AND THOMPSON, J. 1985. A trace-driven analysis of the Unix 4.2 BSD file system. In *Proceedings of the 10th Symposium on Operating System Principles (SOSP)*. Orcas Island, WA. 15–24.

PARK, K. AND WILLINGER, W. 2000. Self-similar network traffic and performance. In *Self-Similar Network Traffic: An Overview*. (Chapter 1) John Wiley Interscience.

PAXSON, V. AND FLOYD, S. 1995. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Trans. Netw. 3*, 226–244.

REAL NETWORKS INCORPORATED. 2001. RealNetworks facts. URL: http://www.reanetworks.com/gcompany/index.html.

REALNETWORKS. 2003. RealNetworks and major media companies launch streaming news, sports and entertainment content to mobile devices. Press Release. Available at http://www.realnetworks.com/company/press/releases/2003/mediaguides.html.

REJAIE, R., HANDLEY, M., AND ESTRIN, D. 1999. RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the Internet. In *Proceedings of IEEE Infocom*. New York, NY. 1337–1345.

SAROIU, S., GUMMADI, K. P., DUNN, R. J., GRIBBLE, S. D., AND LEVY, H. M. 2002. An analysis of Internet content delivery systems. In *Usenix Operating Systems Design and Implementation (OSDI)*. Boston, MA. 315–327.

SAROIU, S., GUMMADI, P., AND GRIBBLE, S. 2003. Measuring and analyzing the characteristics of Napster and Gnutella hosts. *Multimedia Syst. J. 9*, 2 (Aug.), 170–184.

STOICA, I., SHENKER, S., AND ZHANG, H. 1998. *Core*-stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks. In *Proceedings of ACM SIGCOMM Conference*. Vancouver, British Columbia, Canada, 118–130.

SULLIVAN, D. Search engine sizes. Available at http://searchenginewatch.com/reports/sizes.html.

TOPIC, P. 2002. DSL Passes 30m lines worldwide. Available at http://www.point-topic.com/-analysis.htm.

VELOSO, E., ALMEIDA, V., MEIRA, W., BESTAVROS, A., AND JIN, S. 2002. A hierarchical characterization of a live streaming media workload. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*. Marseille, France, 117–130.

WANG, Y., CLAYPOOL, M., AND ZUO, Z. 2001. An empirical study of RealVideo performance across the Internet. In *Proceedings of the ACM SIGCOMM Internet Measurement Workshop (IMW)*. San Francisco, CA. 295–309.

WANG, Z., BANERJEE, S., AND JAMIN, S. 2003. Studying streaming video quality: From an application point of view. In *Proceedings of ACM Multimedia*. Berkeley, CA. 327–330.

WILLINGER, W., TAQQU, M., SHERMAN, R., AND WILSON, D. 1995. Self-similarity through high-variability: statistical analysis of eithernet LAN traffic at the source level. In *Proceedings of ACM SIGCOMM*. Cambridge, MA. 100–113.

WILLS, C. E., MIKHAILOV, M., AND SHANG, H. 2003. Inferring relative popularity of Internet applications by actively querying DNS caches. In *Proceedings of the Internet Measurement Conference (IMC)*. 78–90.

WOODRUFF, A., AOKI, P., BREWER, E., GAUTHEIR, P., AND ROWE, L. 1996. An investigation of documents from the World Wide Web. In *Proceedings of the 4th International World Wide Web Conference*. Paris, France, 963–979.

Received December 2003; revised July 2004; accepted July 2004