WPI-CS-TR-03-18 May 2003

Characteristics of Streaming Media Stored on the Web

by

Mingzhe Li
Mark Claypool
Robert Kinicki
Jim Nichols

Computer Science
Technical Report
Series

WORCESTER POLYTECHNIC INSTITUTE

Computer Science Department
100 Institute Road, Worcester, Massachusetts 01609-2280

Characteristics of Streaming Media Stored on the Web

Mingzhe Li, Mark Claypool, Robert Kinicki and James Nichols
{lmz,claypool,rek,jnick}@cs.wpi.edu
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA, 01609, USA

Abstract— **The increasing power and connectivity of today's computers have spurred the growth in streaming audio and video available on the Internet through the Web. While there is substantial research characterizing the performance of streaming media and characterizing documents stored on the Web, there have been few studies characterizing streaming audio and video stored on the Web. We crawled over 17 million Web pages from key geographic locations and extracted nearly 30,000 streaming audio and video clips for analysis. Using custom built tools, we analyzed the characteristics of these multimedia objects, determining such information as media type, encoding format, playout duration, bitrate, resolution, and codec. We find proprietary audio and video formats dominate all multimedia content, primarily content by RealNetworks followed next by Microsoft and with Apple following just behind MP3. The playout durations of streaming audio and video clips are long-tailed, suggesting streaming media may contribute self-similar traffic on the Internet. More than half of all streaming media clips on the Web are video, with 90% of videos targeted for broadband connections. Video resolutions are considerably smaller than typical monitor resolutions, implying that video bitrates, which are directly related to resolutions, have enormous potential to increase. The detailed results from this study should be useful for future studies characterizing the performance of streaming media on the Web and also valuable for those interested in generating more accurate Internet traffic simulations.**

Keywords— **Multimedia, Streaming, RealNetworks RealPlayer, Microsoft Windows Media Player, Apple QuickTime, Self-similarity, Long-tailed**

I. Introduction

Improvements in the power and connectivity of today's computers have enabled the growth in 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 have come to expect audio and video streams at their fingertips by simply clicking on a browser link to automatically start playing streaming media. In 2001, Real Networks [1] 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.

CAIDA [2] emphasized in 2002 the significant fraction of Internet link capacities that were already being allocated to support streaming media applications. Announcements such as RealNetworks' [3] press release to support the advancement of streaming multimedia applications over wireless cellular networks have added to the concern among Internet performance experts about being able to support even more readily available access to streaming media clips through the Web. This anxiety over future streaming media applications significantly restricting performance of other Web users has translated into a variety of research papers that propose new network protocols [4], [5] or more sophisticated network router algorithms that seek to lessen the impact of streaming media [6], [7], [8], [9]. Several recent research efforts [10], [11], [12], [13], [14], [15], [16] have focused on capturing the characteristics of current streaming application behavior to better understand its impact. Only by knowing the nature 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 the exact characteristics of streaming media stored on the Web. While there have been studies characterizing Web content [17], [18] measured at the client side, there have been no recent studies on the general characteristics of streaming media stored at the Web server. In 1997, Acharya and Smith [19] characterized 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 [19] found that in 1997 the Internet could not support real-time streaming given the encoded bitrates and available last-mile data capacities. Today, RealNetworks RealPlayer and Microsoft

Media Player, two popular streaming media products [20] that did not even exist in 1997, have significantly improved Web users' ability to stream multimedia to home computers.

Two papers, one by Ousterhout *et al* [21] and the other by [22], provide good examples where fundamental research on understanding the nature of data stored in file systems and studying how these files were likely to be accessed, proved to be influential in the design of new file systems and distributed file systems. The accessibility to media clips on the Web through RealNetworks and Windows Media Players has reached such a state that similar fundamental research on the nature of streaming media stored on the Web is critical to understanding the impact of streaming traffic on future Internet performance.

This investigation built customized tools to answer the following questions about the characteristics of streaming media stored on the Web today:

- *What are the most popular streaming technologies?* Previous research [12] has shown that proprietary encoded media products significantly differ in their impact on streaming network traffic, even when the products utilize the same network bitrates. Similar to the situation in 1997 when the large user base for MPEG, AVI and QuickTime was an obstacle for incoming streaming technologies, by quantifying today's dominant technologies one can uncover current obstacles for future media applications.
- *What are the relative amounts of streaming audio clips versus streaming video clips?* The type of media, whether audio or video, has a large impact on performance requirements. Audio often requires only modest bitrates but typically has very discrete encoded bitrates. Video, on the other hand, is often bitrate-hungry and can stream over a wide range of encoded bitrates.
- *Are streaming media playout durations 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 [23] for a survey). Long-tailed distributions of transfer times [24], [25], [26] may contribute to the self-similarity of Internet traffic. Similarly, if the distribution of streaming media playout durations is long-tailed, then streaming media may contribute to Internet traffic self-similarity, especially as the fraction of streaming media grows.
- *What are typical streaming media target bitrates?* When encoded, streaming media clips have a target bitrate that has a direct impact on the network traffic rate the media will have when streamed. Video target bitrates, in turn, are influenced by such parameters as frame resolution, frame rates and color depth. Knowledge of typical target bitrates provides insight into the strategies that media content providers use to deal with limited capacities at last-mile connections.
- *What fraction of the many 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 Internet provides information as to the speed at which new codec technologies are deployed.

This paper provides detailed information to answer these questions on streaming media characteristics on the Web today. Since commercial products have had a significant 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 viewed real streaming traffic by monitoring behavior near clients or servers [10], [13], [14], [27], [28], this investigation seeks the wider perspective of reviewing streaming content at media servers world-wide. While there is substantial audio and video content stored on peer-to-peer (p2p) file sharing systems [29], [30], this content is not typically streamed at a target bitrate, but is typically first downloaded as fast as capacity will allow and subsequently played. Thus, the network traversal behavior for p2p file sharing systems is more similar to bulk file transfer than to streaming. Since this study is focused 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 file sharing systems is left as a future project.

We built a specialized crawler that launched from 17 carefully selected starting points across the Web and then traversed over 17 million URLs, extracting unique URLs specific to streaming media. Our custom-built tools captured information from nearly 30,000 of the media URLs, recording specific media parameters that have a direct impact on Internet performance. 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 enormously since 1997. Proprietary content is the most prevalent, with RealNetworks having by far the most encoded media stored on the Web and Microsoft Media alone in second place. Most streaming media clips are relatively short, but have a distribution with a long tail. Application of proposed long-tailed distribution tests [31] lends credence to the belief that streaming media playout durations are long-tailed and thus may contribute to the self-similar nature of Internet traffic. 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 are useful to provide guidelines for choosing representative samples of streaming media clips in empirical Internet measurement studies that desire to characterize the behavior of commercial media streaming traffic over the Internet, such as in [32], [16], [33], [11], [12], [15]. Moreover, the results from this work can also be used to generate more accurate traffic models of streaming media for large scale Internet simulations.

This paper is organized as follows: Section II discusses the crawling methodology used and the custom tools developed to measure the characteristics of streaming audio and video available on the World Wide Web; Section III analyzes the results of an extensive effort to search the Web for streaming media, including insight into the overall characteristics of streaming audio and video clips on the Web today; Section IV discusses Internet sampling issues related to this investigation; Section V considers the application of these findings to future research; Section VI puts forth conclusions; and Section VII proposes possible future work.

II. Methodology

We used the following methodology to collect extensive information on the nature of streaming media currently stored throughout the World Wide Web:

- We created a customized Web crawler, Media Crawler, to search for and collect the URLs of freely available audio and video clips. (see Section II-A)
- We devised a strategy for obtaining a representative sample of available streaming audio and video clips stored on the World Wide Web. (see Section II-B)
- We developed tools and techniques for collecting characteristics of the streaming audio and video content from the URLs extracted by Media Crawler. (see Section II-C)
- We started streaming for each URL in the complete set of unique streaming media URLs to perform packet header analysis and uncover those characteristics of streaming audio and video that impact performance of multimedia streamed directly from Web pages. (see Section III)

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	.mov, .qt
Real Media	.ra, .rm, .ram, .rmvb, .smil
WAV	.wav
Windows Media	.asf, .asx, .wma, .wmv, .wax, .wvx

TABLE I
AUDIO AND VIDEO URL EXTENSIONS

A. Media Crawler

We modified Larbin[1], an open source, general purpose Web crawler, to create a specialized crawler, *Media Crawler*, designed to extract audio and video URLs while crawling the Web. Starting from a specified root URL, Media Crawler recursively traverses embedded URLs and identifies by 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 [27], audio and video are sometimes streamed over HTTP. Thus Media Crawler was designed to also examine URL extension to find streaming media clips. Table I itemizes the set of URL extensions that Media Crawler uses as an indicator of streaming media content. We chose this set of extensions by extracting the list of standard file type extensions that appear in file operation drop-down list boxes in most commercial media players.

Since our objective was to create a list of unique streaming media URLs and to avoid crawling loops, Media Crawler maintains a data structure of 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 whether a newly encountered URL is unique grows with the number of previously identified unique URLs. This was a factor in choosing a strategy of serially launching Media Crawler from multiple starting points rather than crawling more extensively from one starting point.

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

B. Starting Pages

The growth of streaming multimedia over the Web is tightly coupled with the availability of high bitrate Internet connections. Consequently, in selecting multiple starting points for Media Crawler the objective we picked popular Web pages that are likely to be accessed by well-connected users. Since another goal of this investigation was to not only consider stored streamed media that is readily accessed from clients in the USA, starting points were selected from Web sites hosted from the ten most-wired countries (excluding the USA) based on data from a market analysis report on broadband penetration [34]. This scheme provides for a more representative set of characteristics for streaming media stored throughout the World Wide Web. Secondarily, geographically dispersed starting pages reduced the overlap in the search space between the individual crawl instances.

We consulted a report by Nielsen,[3] the television and Internet ratings company, to determine the top ten Web sites in each country and to guide the selection of crawl starting points both inside and outside of the USA. For any top ten wired countries where Nielsen provided no information, we selected a popular domestic newspaper or news portal as the starting page. Since the Unites States is the most wired country, we also chose seven USA Web pages as starting points. These seven Web pages were selected from the most popular Web sites that included the following specific Web page types: news, sports, entertainment, Internet portal, search engine, and streaming media technology. Table II lists the 17 starting pages used in this research, listed in alphabetical order by country. In Section IV, we analyze the impact of the number and specific choices for starting locations on the statistical validity of this investigation.

Beginning from each distinct starting page, Media Crawler proceeded to crawl URLs until it discovered one million unique URLs whereupon it created an output file containing a list of URLs that refer to streaming media objects. While Media Crawler records unique multimedia URLs within a single crawl, it is possible that crawls starting from different places on the Web will overlap and produce the same multimedia URL on multiple output files. Thus, we wrote a separate program to create the final set of unique multimedia URLs across the 17 one-million URL data sets. Section III-A presents a discussion of the amount of overlap in multimedia 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, our second stage analysis that including starting up each of the audio and video URLs was conducted less than 24 hours after the final set of multimedia URLs was produced.

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

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

TABLE II
MEDIA CRAWLER STARTING PAGES

C. Measurement of Content Characteristics

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

To analyze Real Media content, we built two new tools. First, we build a custom tool, *RealAnalyzer*, using Microsoft Visual C++ and the RealNetworks SDK[5] that is provided 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 enables development of new tools and applications that can stream Real media. Real

[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

Analyzer gathers content description information such as: URL, encoded bitrate, duration, resolution, live or pre-recorded, title, and copyright.

Second, we developed a custom tool, *TestPlay*, to gather RealPlayer content statistics. An original version of Test-Play is available with the RealPlayer SDK under the directory `sdk/samples/intermed/testplay`. Test-Play allows the measurement of content encoding information including the number of sources, encoded bitrates, and codec information. With further modifications to RealAnalyzer and TestPlay to 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, we developed two other custom tools, similar to those for Real Media. 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. We created a second custom tool, *Wmprop* to gather 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, we used *MPlayer*,[7] an open source tool that runs on the Linux operating system, to analyze Apple Quick-Time content. When playing QuickTime content, MPlayer produced resolution and codec information. However, MPlayer could not determine the encoded bitrate of the QuickTime content.

III. 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 different URLs were reached.[9] The total execution time for each Crawler instance depends upon the starting point. The crawl beginning from the starting point with the largest round-trip time from WPI[10] (namely, sina.com.cn in China) took approximately 24 hours while several of the closer sites took about four hours (see [35] for more details).

The analysis of the collected streaming media information proceeded in four stages:

- First, we performed aggregate analysis on the complete list of media URLs produced by Media Crawler from several perspectives. Coarsely, we studied the distribution of multimedia URLs clustered per server and then conduced finer grain analysis in determining the most popular streaming technologies. See Section III-A.
- Second, the next phase of the data analysis focused on relative proportions of content created by the major commercial streaming products: Real Media, Windows Media, and Apple QuickTime Media. Using custom tools described in Section II, we collected content information from each of the streaming media clips. We determined the relative amount of audio and video for streaming and used content duration information to test the hypothesis that audio and video playout durations are long-tailed. See Section III-B.
- Third, we considered lower level streaming media characteristics that impact streaming transmission rates. This analysis included encoded bitrates, resolutions, and the media codecs used to encode both streaming video and streaming audio. See Section III-B.1 and Section III-B.2, respectively.
- Fourth, we considered the significance of the sampling size on the representativeness of the data gathered from the 17 crawler executions. We evaluated the impact of such experimental decisions as the number of URLs crawled, the number of starting points, and the geographic location of starting Web pages on the results. See Section IV.

A. Aggregate Analysis

Prior to aggregate analysis, we removed duplicate URLs from the 17 distinct 1-million URL data sets, resulting in 11,533,849 unique URLs (see [35] 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, as described in Section II-A . In 1997, Acharya and Smith [19] reported finding 22,600 media URLs out of the 25 million Web pages [36] indexed by Alta Vista at that time. Thus, the percentage of audio and video objects stored on the Web has more than five-fold from about 0.09% in 1997 to about 0.47% in 2003. Moreover, given that the Google search engine currently

[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, MA, USA.

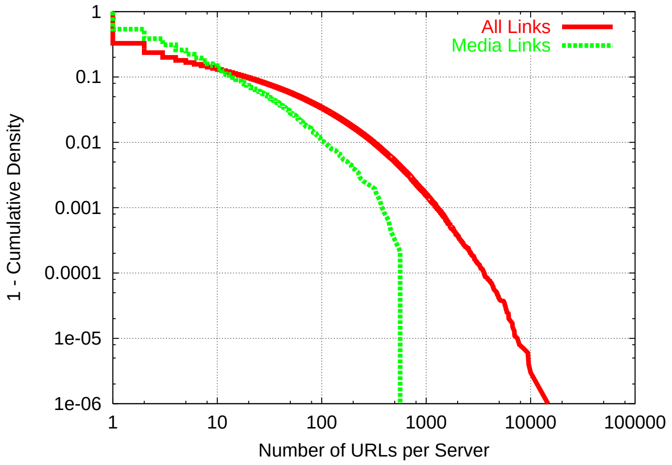


Fig. 1. URLs per Web Server and Media URLs per Web Server with Streaming Media

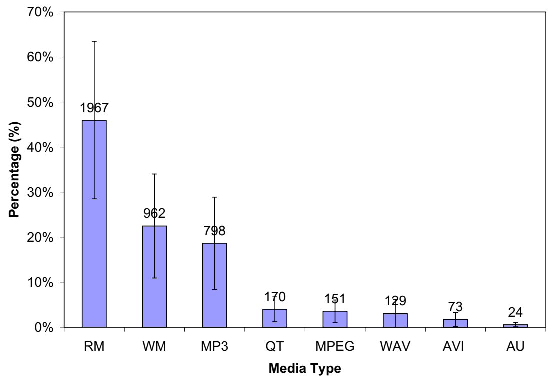


Fig. 2. Percentage of Each Media Type

indexes more than 3 billion Web pages,[11] one can project that there exist nearly 15 million freely available streaming audio and video clips stored on the Web today. Note, nothing in this investigation addresses pages on the Web that a client has to pay to reach.

The complementary cumulative density 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, with the median number of URLs per server over the set of all servers crawled being only one URL. The 54,762 unique audio and video URLs came from 4678 different servers, with the median number of URLs per server for the set of servers that had streaming media being about 4 media URLs per server. Note the graph indicates that about 1% of the set of streaming media servers provide 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, with the error bars depicting the standard deviation across the 17 sets of crawler data. Real Media is the most popular media type stored on the Internet today, accounting for almost half of all streaming media URLs and being twice as abundant as Windows Media. QuickTime, MPEG, and AVI, the most popular video types in 1997 [19], make up only a combined 10% share of the videos in 2003. MP3, a popular streaming audio format, is the most popular non-proprietary format and is more prevalent than Apple QuickTime Media.

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

B. *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 pre-recorded and live audio and video over the Internet.

Of the 54,762 unique Real Media, Windows Media and Apple QuickTime Media URLs recorded by Media Crawler only 29,056 (about 53%) were valid 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 [35]) with our tools as to why these clips may have been unavailable provided 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 paper is based on the data obtained from these 29,056 accessible multimedia URLs.

While in principle, each media URL can be a playlist with multiple streaming media clip entries, the data analysis implies this occurs infrequently on the Web. Over 97% of the playlists refer to only one streaming media clip and only about 1% of playlists refer to 3 or more streaming media clips (see [35] for more detailed analysis on the playlists).

Media Type	Audio	Video	Total	Percent
Real	9863	8504	18367	63
Windows	2591	6567	9159	32
QuickTime	28	1474	1521	5
Total	12482	16545	29056	100

TABLE III
NUMBER OF STREAMING MEDIA CLIPS ANALYZED

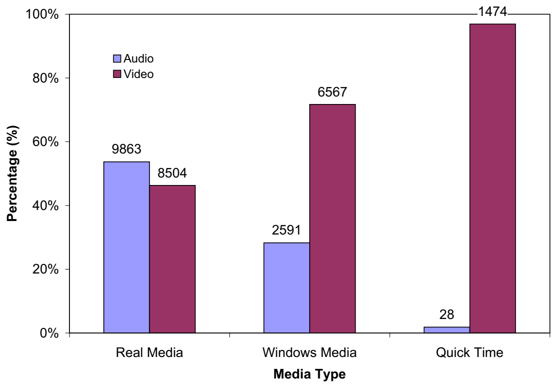


Fig. 3. Percentage of Audio and Video for Each Media Type

Figure 3 graphs the percentage of audio and video for each of the three major media types. Overall, 43% of the media clips were audio only, and 54% of the Real Media clips are audio. Combining information from Figure 2 and Figure 3, it is clear that in the collected URLs there is more Real Audio stored on the Internet 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 amount of QuickTime audio, subsequent analysis only considers QuickTime video.

Our tools use attributes in the streaming media header to determine if the media is live or pre-recorded. For Windows Streaming Media, the types identified are broadcast, streamed, or downloaded where broadcast indicates live streaming and the other two are are pre-recorded. For Real Media, the header indicates either live or pre-recorded. 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 pre-recorded streaming content, the vast majority of the available streaming clips are pre-recorded. 98% of all streaming media clips are pre-recorded, with Real Media having about 2% live clips, Windows Media having about 3% live clips and QuickTime having less than 1% live clips.

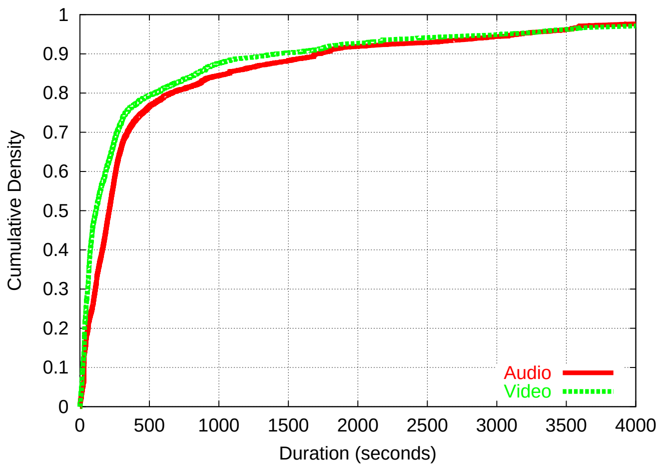


Fig. 4. CDF of Streaming Media Duration

During the duration analysis three outlier clips with a duration of 10 days (roughly an order of magnitude longer than the next longest streaming clip) were uncovered. Closer inspection revealed these clips were erroneous text streaming with an error in their duration length. These three clips were removed from all subsequent analysis.

The CDF of the duration of all the 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, with a median duration of about 3 minutes (the median is about 2 minutes for video and 4 minutes for audio). 10% of 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 duration of videos stored on the Web today are significantly longer than in 1997 when 90% of video clips lasted 45 seconds or less [19].

Self-similar traffic is difficult to manage and a long-tailed distribution of transfer times may contribute to the self-similarity of Internet traffic. If the distribution of the durations for stored streaming clips is long-tailed, this lends credence to the possibility that the distribution of transfer times on the Internet for pre-recorded streaming transfers may also be long-tailed. Note, this discussion excludes live streaming events that have an undetermined duration. Figure 5 gives the CCDF of the audio and video duration distributions to allow for clearer examination of the tails of the 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 du-

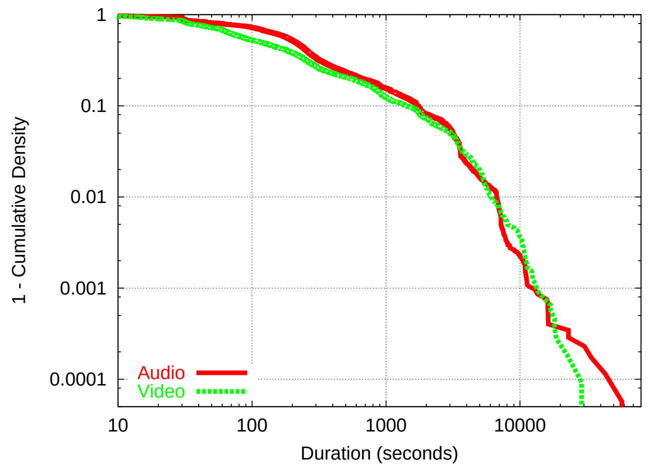


Fig. 5. CCDF of Streaming Media Duration

rations of the stored audio and video clips may be long-tailed. However, as discussed by Downey [31], certain distributions, such as lognormal, appear visually 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 [31] was applied. The details from the five steps in this process include:

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 for 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 called `aest`[13] [37]. 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 calculates 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 so the p-value is 0.498, and for the video durations, 495 differ from μ by d or more so 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 means streaming media playout durations may be long-tailed. If one assumes that the set of stored media clips are uniformly accessed, then the long-tailed distributions of the duration of stored clips would lend support to the conjecture that actual streamed media transfer times over the Internet are also long-tailed. This phenomenon would contribute to the self-similarity of Internet traffic.

Note that the streaming playout duration distributions do not include any of the live content that the crawler encountered. With live content the stream duration may not be known even by the encoder. Thus none of the commercial player APIs provide a mechanism to determine the possible duration of a live stream. It is unclear where along the distribution we should place these samples, so we ran additional curvature tests with manual placement of the clip durations along the duration CDF. We placed the live clips before the tail, at the beginning of the tail, and evenly along the tail and repeated the curvature test each time. For all cases the outcome of the curvature test, namely that streaming media durations may be long-tailed, was unchanged. However, while some live content available on the Internet could be short, since users must "tune in" at the time live content is broadcast, live streaming is likely long,[14] thus making the streaming media duration distributions even more long-tailed.

B.1 Video

Video can operate over a wide range of bitrates. Video conferences and low-bitrate videos stream at about 0.1 Mbps[15]; VCR quality videos stream at about 1.2 Mbps[16]; broadcast quality videos stream at about 2-4 Mbps[17]; studio quality videos stream at about 3-6 Mbps[17]; and HDTV

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

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

[14] Veloso et al [14] analyzed live streams that were at least 28 days long.

[15] H.261 and MPEG-4

[16] MPEG-1

[17] MPEG-2

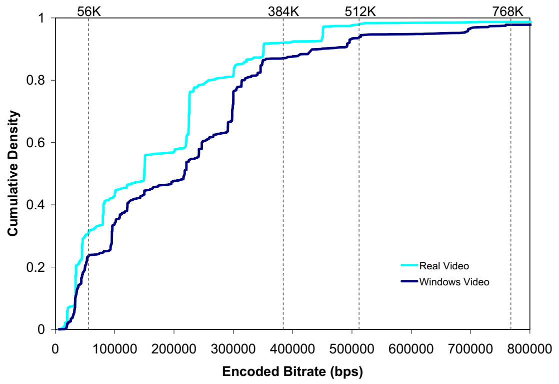


Fig. 6. CDF of Streaming Video Encoded Bitrate

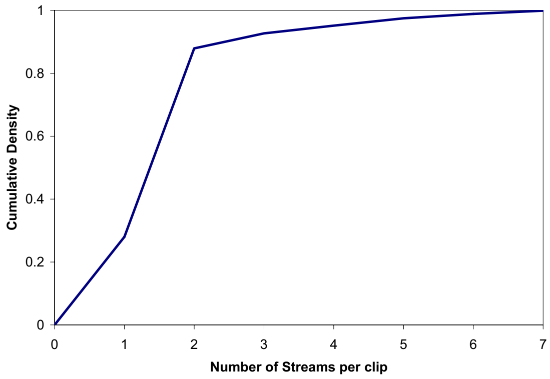


Fig. 7. CDF of Number of Windows Media Streams Encoded in One Clip

quality videos stream at about 25-34 Mbps[17]. 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 II, Quick Time Media encoding rates could not be captured). The median encoded bitrate is around 200 Kbps, with the median encoded bitrate for Windows Media being slightly higher than the median encoded bitrate for Real Media. Approximately 29% of the videos are encoded to stream over a 56 Kbps modem, a substantial increase from 1997 [19] when fewer than 1% of videos were encoded for modem bitrates. Nearly 70% of the videos are targeted for broadband (56k - 768k), up from 50% in 1997. Approximately 1% of the videos have bitrate targets above typical broadband connections (768k - 1500k), and less than 1% have bitrate targets above a T1 (1540k+), down from about 20% in 1997.

The general shift in target encoded bitrates, with 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 density of the number of encoded streams per Windows Media clip for the clips the crawler found. From the measurements, we found approximately 12.1% of Windows Media clips have multiple bitrate encoding levels. The lack of encoded bitrate choices for a media server has significant 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 to 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 [15], 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 had a standard aspect ratio of 4/3. The remaining 30% of the video clips had aspect ratios slightly above and slightly below 1.3 (see [35] for more details). The vertical lines in Figure 8 depict commonly used video resolutions: 160x120 (quarter-screen), 240x180 (three eighths-screen), and 320x240 (half-screen). The steps in the distributions correspond roughly to different resolution choices available in com-

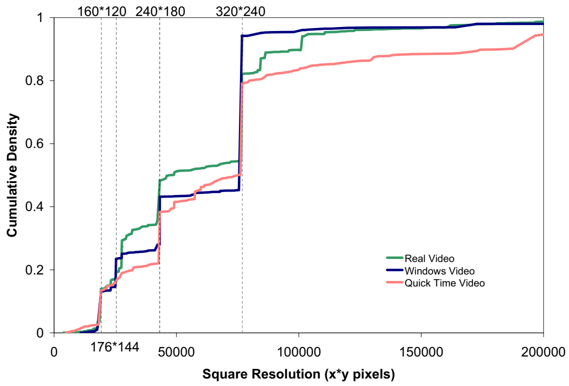


Fig. 8. CDF of Video Resolution ($length \times width$)

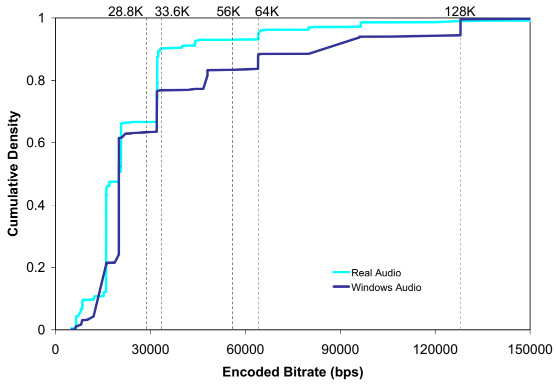


Fig. 9. CDF of Streaming Audio Encoded Bitrate

mercial 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 is likely due to the relationship between resolution and required bitrate for streaming. A video with a resolution of 320x240 will typically result in bitrates on the order of hundreds of Kbps (the target bitrates shown in Figure 6). Given current typical desktop resolutions of at least 640x480 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.

B.2 Audio

Figure 9 depicts CDFs for the encoded bitrates of the streaming audio clips for both Windows Media and Real Media. The encoded bitrates for streaming audio are low compared with the encoded bitrates for streaming video shown in Figure 6. About 90% of streaming audio is targeted for modems, with the median encoded audio bitrate suitable for streaming over older 28.8 Kbps modems. In 1999, an empirical study of streaming audio at a popular Internet audio server [13] 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.

C. 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 clearly beneficial. For example, WMv8 fills the playout buffer at the target playout rate [12], while WMv9, in a manner similar to RealPlayer [15], buffers at a significantly higher data rate.

Figure 10 and Figure 11 captures 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 clear snapshot in time, e.g., WMv9 having only 2.31% of the recorded codecs in May of 2003. However, future studies may find this data valuable in tracking the acceptability and change in market penetration over time of such innovations as WMv9.

Figure 10 shows the prevalence of different versions

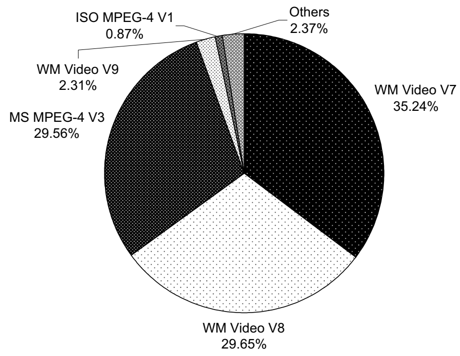


Fig. 10. Breakdown of Windows Video Codecs

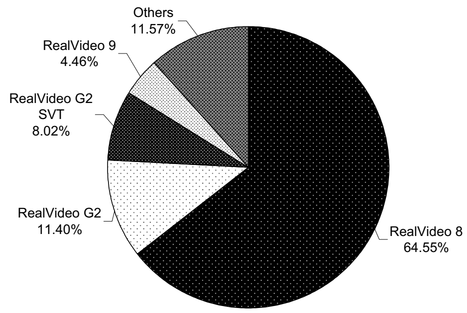


Fig. 11. Breakdown of Real Video Codecs

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 uses discrete cosine transform and motion prediction to encode and compress video content, being renamed WM Video 7 and released in May of 2001. WM Video 8 was released soon after in September of 2001, and was the only Microsoft codec until the most recent version, 9, released in January of 2003.

Figure 11 breaks down 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 amounts relative to RealVideo 8.

IV. 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 [19], but today that is impractical. Crawling the 17 million URLs used in this study took over one month. At this pace, it suggests it would take over 16 years to crawl over 3 billion pages currently on the Web. Moreover, 200 days would be needed to to actually download via streaming just the media clips analyzed in this study. To download via streaming all the freely available multimedia clips on the Web would require four years of continuous streaming. Storing these clips for subsequent use is equally problematic.

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, our strategy was to evaluate the effects of smaller sample sizes on the quality of the resultant analysis. We considered four specific questions:

- 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 less than one million unique URLs per crawl instance?
- How does the sampling, in terms of number of URLs and 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 crawling starting points, virtual experiments with fewer than one million URLs were considered. Beginning with 200,000 URLs and proceeding in increments of 200,000 URLs, up to the full one million URLS, we reviewed five separate data-gathering plateaus. Thus, the smallest data set had 3.4 million URLs ($17 \times 200,000$), and each subsequent data set increases by 3.4 million until the full 17 million URL set was reached.

For each set of media URLs found by the crawler had it stopped at a given plateau, we determined the percentage of each type of media (similar to analysis in Figure 2). Figure 12 demonstrates that at the 10.2 million URL plateau and beyond, all the percentages for the various media product types remain constant. This data suggests that, at least for this statistic, crawling more than 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

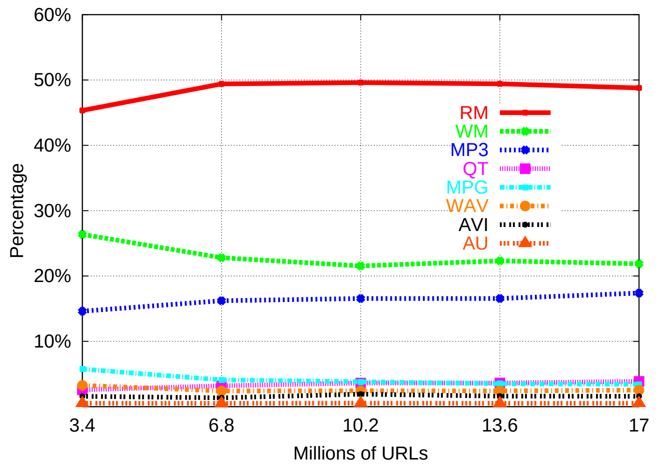


Fig. 12. Percentage of Media Types versus Number of URLs Crawled

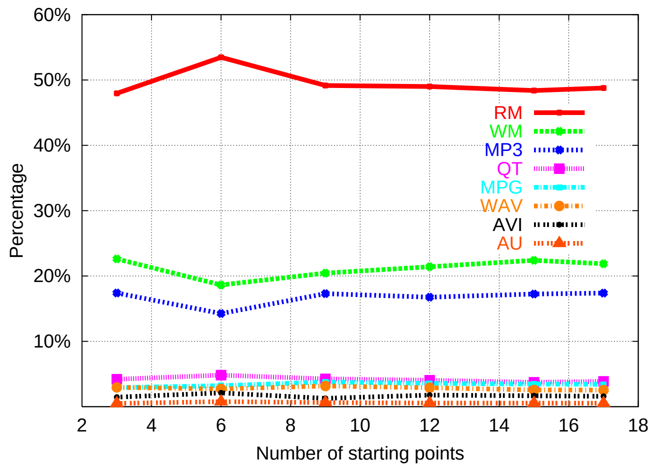


Fig. 14. Percentage of Media Types versus Number of Starting Points

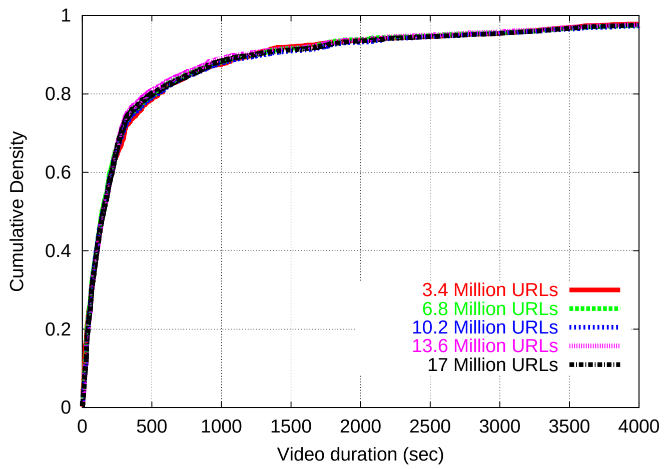


Fig. 13. CDF of Video Duration for Different Sample Set Sizes

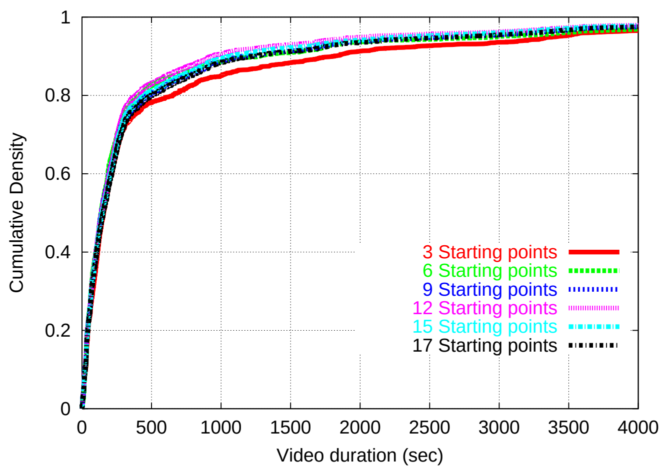


Fig. 15. CDF of Video Duration for Different Numbers of Starting Points

plateaus yields very similar results.

To drill down further, we also analyzed the impact of the data set size on the distribution of several important media clip characteristics. Only the results for video playout duration are shown here (see [35] for analysis of other streaming media characteristics for the different data set levels). 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 remarkable 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.

The next issue considered was whether the number of starting points would have a significant affect on the results obtained. From the 17 original starting points, data from five 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 stays relatively constant. Crawling from a larger number of similar 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 remarkably similar for all numbers of starting points except for a slight separation for the distribution 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 any one of the seven USA starting points and the set of URLs obtained from

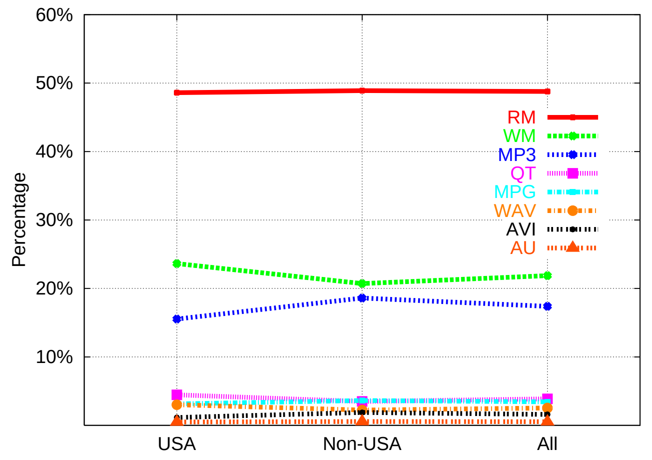


Fig. 16. Media types of USA and non-USA starting points

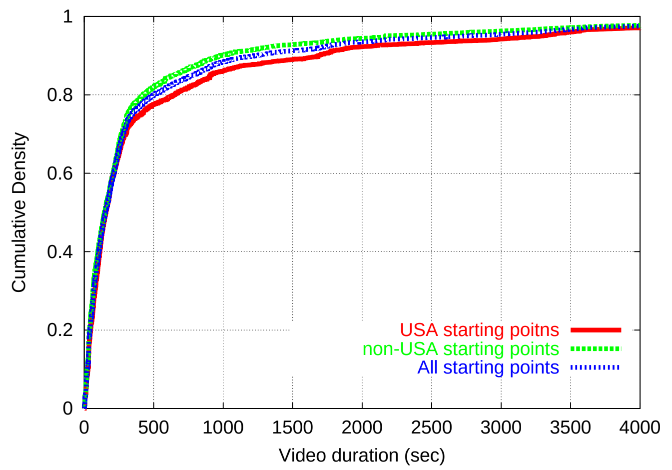


Fig. 17. CDF of video duration of USA and non-USA starting points

the 10 starting points outside the USA. 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 USA starting points have slightly more Windows Media clips but fewer MP3 clips, but they have an equivalent percentage of Real Media clips.

Combining the analysis of number of URLs, number of starting points, and cultural locations, one could argue that from the using 9 or more starting points with 600,000 URLs per starting point provides a large enough sample space to analyze the characteristics of stored streaming media on the Web. The graphs in this section lend credibility to our belief that having crawled 1 million URLs from each of 17 starting points, the resultant sample set of unique URLs is on the safe side of large "enough". Moreover, the cumulative effect of all the figures in this section is to provide confidence to the belief that the analysis presented in this paper on the characteristics of stored multimedia URLs is representative of the Web at large.

V. Application of Results

The characteristics of streaming traffic uncovered in this paper are valuable as a snapshot of the nature of fully accessible stored media on the Web today. Furthermore, this information is quite useful for researchers wishing to design and conduct experiments to evaluate the impact of streaming audio and video content on overall Internet performance.

Conducting empirical experiments involving streaming video traffic is difficult due to variable network conditions, the setup costs in deploying large numbers of streaming clients, and the effort required to build, deploy and coordinate the instrumentation tools. Consequently, using simulators, such as NS-2 [38], has become increasingly common. The results presented in this paper are of value to researchers designing simulation studies that want to model the nature of streaming media cross traffic in 2003. Figure 2 and Figure 3 provide information on the current ratio of streaming audio and video traffic for commercial products and detailed data on encoding types. The shape and steps in the bitrate distributions, shown in Figure 6 and Figure 9 provide guidance on choosing an appropriate mix of bitrates to reasonably capture the behavior of freely available streaming downloads from audio and video servers. Moreover, the duration of streaming audio and video flows can be chosen from the duration distributions in Figure 4.

The use of commercial streaming products, such as the Microsoft Windows Media Player and RealNetworks RealPlayer, has increased dramatically [20]. Understanding the performance of commercial streaming media products plays an important role in understanding the impact of streaming media on the Internet. Figure 2 provides guidance on the most prevalent commercial products. Figures 10 and 11 offer insight concerning the speed at which new media player products penetrate the marketplace and could influence the choice of which version of a product to study. The results presented in this paper may be useful for designing experiments for studies similar to [32], [16], [33], [11], [12], [15] that actively measure performance of commercial streaming media technologies. Such studies may even sample from the list of streaming media clips that were analyzed in this paper,[18] to avoid additional time-

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

consuming crawling or more biased streaming media clip selection.

VI. 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 available Web pages, it becomes difficult for network 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 multimedia 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, we crawled 17 million Web URLs and checked for validity, to yield nearly 30,000 unique audio and video clips. We then carried out in depth analysis on the stored clips by partially downloading the initial segments of each of these clips to extract header and other characteristic information about the media clips. These downloads originated from 4678 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 over 600% in the past five years. Moreover, the fraction of streaming media objects stored on the Web relative to other objects has increased over 500%.

The aggregate data analysis shows streaming audio and video content is dominated by proprietary streaming products, specifically RealNetworks Media first and 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 pre-recorded, with only a very small fraction being 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. Thus, just from this increase in the duration of media flows, it is clear that the impact of streaming video on the Internet has grown substantially.

Despite growth of broadband connections, the fact that the majority of audio encoded bitrates today are still targeted to be acceptable for modem connections is a significant. Moreover, the distribution of video bitrates implies that modems can also be used for streaming some video clips. Having streaming content suitable for modems is a useful niche given that it is estimated that half of all USA Internet subscribers will still use modems by the year 2005 [39]. 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 320x240) 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.

VII. Future Work

While the advertised target streaming bitrates presented in this report provide insight as to possible network impact, the actual streaming rates over the Internet may be quite different. The level of responsiveness for streaming media flows to Internet congestion and perceived available bitrate is expected to have a large impact on future network performance. Technologies such Windows Media “intelligent streaming” and RealNetworks “SureStream” can be used to provide multiple target bitrates in one stored media object. Previous work [15] 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 presented here depict details on the storage of audio and video on the Internet, they do not provide details on the actual streaming of the stored audio and video over a network. Future work could complement these results with measurements of actual streaming use. Such efforts would be especially useful if a media server with many audio and video encoding rates and choices were specifically studied. Specific techniques that actively query DNS caches such as in [40] could be used to provide complementary information about the popularity of Web sites sites with stored audio and video.

Our crawling methodology is specifically targeted at locating and analyzing streaming media, that is, media that will be 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 multimedia content found may provide valuable insights into the use and support of such file sharing systems.

REFERENCES

[1] Real Networks Incorporated, "RealNetworks Facts," 2001, URL: http://www.reanetworks.com/gcompany/index.html.

[2] Cooperative Association for Internet Data Analysis (CAIDA), "Characterization of Internet Traffic Loads, Segregated by Application," Oct. 2002, [Online] http://www.caida.org/analysis/-workload/byapplication/.

[3] RealNetworks, "RealNetworks and Major Media Companies Launch Streaming News, Sports and Entertainment Content to Mobile Devices," May 2003, Press Release. http://www.realnetworks.com/company/press/releases/2003/-mediaguides.html.

[4] Sally Floyd, Mark Handley, Jitendra Padhye, and Jorg Widmer, "Equation-Based Congestion Control for Unicast Applications," in *Proceedings of ACM SIGCOMM Conference*, 2000, pp. 45 – 58.

[5] Rezza Rejaie, Mark Handley, and D. Estrin, "RAP: An End-to-end Rate-based Congestion Control Mechanism for Realtime Streams in the Internet," in *Proceedings of IEEE Infocom*, 1999.

[6] R. Mahajan, S. Floyd, and D. Wetherall, "Controlling High-Bandwidth Flows at the Congested Routers," in *Proceedings of the 9th International Conference on Network Protocols (ICNP)*, Nov. 2001.

[7] W. Feng, D. Kandlur, D. Saha, and K. Shin, "Stochastic Fair Blue: A Queue Management Algorithm for Enforcing Fairness," in *Proceedings of IEEE Infocom*, Apr. 2001.

[8] Z. Cao, Z. Wang, and E. Zegura, "Rainbow Fair Queuing: Fair Bandwidth Sharing Without Per-Flow State," in *Proceedings of IEEE Infocom*, Mar. 2000.

[9] Ion Stoica, Scott Shenker, and Hui Zhang, "*Core*-Stateless Fair Queueing: Achieving Approximately Fair Bandwidth Allocations in High Speed Networks," in *Proceedings of ACM SIGCOMM Conference*, Sept. 1998.

[10] M. Chesire, A. Wolman, G. Voelker, and H. Levy, "Measurement and Analysis of a Streaming Media Workload," in *Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS)*, Mar. 2001.

[11] Yubing Wang, Mark Claypool, and Zheng Zuo, "An Empirical Study of RealVideo Performance Across the Internet," in *Proceedings of the ACM SIGCOMM Internet Measurement Workshop (IMW)*, Nov. 2001.

[12] Mingzhe Li, Mark Claypool, and Robert Kinicki, "MediaPlayer versus RealPlayer – A Comparison of Network Turbulence," in *Proceedings of the ACM SIGCOMM Internet Measurement Workshop (IMW)*, Nov. 2002.

[13] Art Mena and John Heidemann, "An Empirical Study of Real Audio Traffic," in *Proceedings of IEEE Infocom*, Mar. 2000, pp. 101 – 110.

[14] Eveline Veloso, Virgilio Almeida, Wagner Meira, Azer Bestavros, and Shudong Jin, "A Hierarchical Characterization of a Live Streaming Media Workload," in *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*, Nov. 2002.

[15] Jae Chung, Mark Claypool, and Yali Zhu, "Measurement of the Congestion Responsiveness of RealPlayer Streaming Video Over UDP," in *Proceedings of the Packet Video Workshop (PV)*, Apr. 2003.

[16] Tianbo Kuang and Carey Williamson, "A Measurement Study of RealMedia Audio/Video Streaming Traffic," in *Proceedings of ITCOM*, jul 2002, pp. 68–79.

[17] Tim Bray, "Measuring the Web," in *Proceedings of the 4th International World Wide Web Conference*, May 1996.

[18] Allison Woodruff, Paul Aoki, Eric Brewer, Paul Gautheir, and Lawrence Rowe, "An Investigation of Documents from the World Wide Web," in *Proceedings of the 4th International World Wide Web Conference*, May 1996.

[19] Soam Acharya and Brian Smith, "An Experiment to Characterize Videos Stored on the Web," in *Proceedings of the ACM/SPIE Multimedia Computing and Networking (MMCN)*, Jan. 1998.

[20] Jupiter Media Metrix, "Users of Media Player Applications Increased 33 Percent Since Last Year," Apr. 2001, Press Release. http://www.jup.com/company/pressrelease-.jsp?doc=pr01040.

[21] J. Ousterhout, H.L. DaCosta, D. Harrison, J. Kunze, M. Kupfer, and J. Thompson, "A Trace-Driven Analysis of the Unix 4.2 BSD File System," in *Proceedings of the 10th Symposium on Operating System Principles (SOSP)*, Dec. 1985.

[22] Mark Baker, John Hartman, Michael Kupfer, Ken Shirriff, and John Ousterhout, "Measurements of a Distributed File System," in *Proceedings of the 13th Symposium on Operating System Principles (SOSP)*, Oct. 1991.

[23] Kihong Park and Walter Willinger, *Self-Similar Network Traffic and Performance*, chapter 1 (Self-Similar Network Traffic: An Overview), Wiley Interscience, 2000.

[24] Vern Paxson and Sally Floyd, "Wide-Area Traffic: the Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226 – 244, 1995.

[25] Walter Willinger, Murad Taqqu, Robert Sherman, and Danlie Wilson, "Self-Similarity through High-Variability: Statistical Analysis of Eithernet LAN Traffic at the Source Level," in *Proceedings of ACM SIGCOMM*, Aug. 1995, pp. 100 – 113.

[26] Anja Feldmann, Anna Gilbert, Polly Huang, and Walter Willinger, "Dynamics of IP Traffic: a Study of the Role of Variability and the Impact of Control," in *Proceedings of ACM SIGCOMM*, Aug. 1995, pp. 301 – 313.

[27] Jacobus Van der Merwe, Subhabrata Sen, and Charles Kalmanek, "Streaming Video Traffic: Characterization and Network Impact," in *Proceedings of the 7th International Workshop on Web Content Caching and Distribution*, Boulder, CO, USA, Aug. 2002.

[28] Jacobus van der Merwe, Ramon Caceres, Yang hua Chu, and Cormac Sreenan, "mmdump - A Tool for Monitoring Internet Multimedia Traffic," *ACM Computer Communication Review*, vol. 30, no. 4, Oct. 2000.

[29] Stefan Saroiu, Krishna P. Gummadi, Richard J. Dunn, Steven D. Gribble, and Henry M. Levy, "An Analysis of Internet Content Delivery Systems," in *Usenix Operating Systems Design and Implementation (OSDI)*, Oct. 2002.

[30] S. Saroiu, P. Gummadi, and S. Gribble, "Measuring and Analyzing the Characteristics of Napster and Gnutella Hosts," *Multimedia Systems Journal*, 2002.

[31] Allen B. Downey, "Evidence for Long-Tailed Distributions in the Internet," in *Proceedings of the ACM SIGCOMM Internet Measurement Workshop*, Nov. 2001.

[32] S. Banerjee Z. Wang and S. Jamin, "Studying Streaming Video

Quality: From an Application Point of View," in *Proceedings of ACM Multimedia*, nov 2003.
[33] Tianbo Kuang and Carey Williamson, "RealMedia Streaming Performance on an IEEE 802.11b Wireless LAN," in *Proceedings of IASTED Wireless and Optical Communications (WOC)*, jul 2002, pp. 306–311.
[34] Point Topic, "DSL Passes 30m Lines Worldwide," Dec. 2002, [Online] http://www.point-topic.com/analysis.htm.
[35] Mingzhe Li, Mark Claypool, Robert Kinicki, and James Nichols, "Characteristics of Streaming Media Stored on the Internet," Tech. Rep. WPI-CS-TR-03-18, CS Department, Worcester Polytechnic Institute, May 2003.
[36] Danny Sullivan, "Search Engine Sizes," [Online] http://searchenginewatch.com/reports/sizes.html.
[37] Mark E. Crovella and Murad S. Taqqu, "Estimating the Heavy Tail Index from Scaling Properties," *Methodology and Computing in Applied Probability*, vol. 1, no. 1, pp. 55 – 79, 1999.
[38] Universiy of California Berkeley, "The Network Simulator - ns-2," [Online] http://www.isi.edu/nsnam/ns/.
[39] Eric S. Brown, "Broadband Walks the Last Mile," *Technology Review*, June 2001, [Online] http://www.technologyreview.com/-articles/print_version/brown060501.asp.
[40] Craig E. Wills, Mikhail Mikhailov, and Hao Shang, "Inferring Relative Popularity of Internet Applications by Actively Querying DNS Caches," in *In Proceedings of the Internet Measurement Conference (IMC)*, Oct. 2003.

APPENDIX

A. Selection of Crawling Domains

Table IV shows the number of broadband connections of Group 7 of the report[19] and South Korea by middle of the year 2002. Table V shows the top 10 DSL connection domains by the third quarter of 2002.[20] By combining the countries from Group 7 and the top 10 countries with DSL connections, we can create a list of the most broadband connected domains in the world, which is the 11 countries listed in Table VII.

B. Validation of our web crawling methodology

We computed the overlap between each given URL set, obtained from starting crawling in in each different domains. The overlap ratio from domain (A) to domain (B) are computed from the following equation:

$$ratio(A- > B) = overlap(A, B)/sizeof(A) \quad (1)$$

Table VI depicts the results. Therefore, overlap ratio from A to B might be different from the overlap ratio from B to A. The range of the overlap goes from 0.13% to 42.81%. The average of overlap ratio is only 3.98%.

[19] Point Topic Report. http://www.point-topic.com/cgi-bin/-download.asp?file=DSLAnalysisBroadband+penetration.htm
[20] Point Topic Report. http://www.point-topic.com/cgi-bin/-download.asp?file=DSLAnalysisQ3+2002+DSL+text+only.htm

	DSL lines	Cable modems	Total
Canada	1,320	1,772	3,091
USA	5,252	8,534	13,786
Japan	3,301	1,620	4,921
Germany	2,570	41	2,611
France	731	217	948
UK	292	453	745
Italy	550	0	550
G7 totals	14,015	12,637	26,652
South Korea	5,734	3,271	9,005

TABLE IV
DSL/CABLE MARKET REPORT (UNITS IN 1000'S)

	DSL lines
South Korea	6076.2
USA	5837.6
Japan	4223.2
Germany	2800
China	2220
Taiwan	1630
Canada	1462.1
France	882
Spain	747.8
Italy	700.4

TABLE V
DSL ONLY MARKET REPORT (UNITS IN 1000'S)

C. Sampling Issues

In sectionIV, we analyzed the major sampling issues briefly. We will discus those issues in more detail in this section.

C.1 Crawler Algorithm

Larbin[1] uses a combined bread-first and depth-first searching algorithm. It use a configurable number of parallel connection (5 in our setting) to following multiple links. Due to the recursive methodology of crawling, it stores a waiting list shared by all the threads and at the beginning of that list, most of the URLs are from the same group of sites. In another word, the crawler need a period of time to spread the tree width to the normal operation size. Therefore, we consider this period as the "warm up" time. Figure 18 depicts the number of unique sites of each 3.4 million URLs. We can find out the warming up time is approximate 2 data set, that's about 6.8 million URLs.

We also compared the absolute percentage of media

TABLE VI

OVERLAP RATIO FOR MULTIPLE STARTING PAGES

	Repub.it	Times	Ntt.jp	BT.uk	Empas.kr	Grupo.es	Espn	Free.fr	Real	T-online.de	AOL	Altavista	Gov.ca	Hollywood	Sina.cn	WMHome	News.tw
Repub.it		6.98%	8.64%	4.28%	4.45%	9.86%	7.94%	12.91%	7.42%	4.97%	4.19%	6.46%	10.91%	3.49%	1.05%	5.32%	0.70%
Times	3.27%		2.49%	1.43%	1.55%	8.46%	10.22%	5.11%	22.40%	7.44%	13.86%	6.05%	3.97%	16.15%	0.78%	15.62%	0.25%
Ntt.jp	2.19%	1.35%		15.96%	4.13%	2.32%	5.43%	2.08%	9.32%	1.46%	3.93%	1.94%	1.85%	2.58%	0.20%	5.70%	0.88%
BT.uk	2.90%	2.07%	42.81%		3.91%	6.10%	3.26%	3.85%	5.45%	4.80%	11.49%	6.93%	3.20%	1.84%	0.59%	13.62%	3.61%
Empas.kr	2.62%	1.95%	9.61%	3.39%		2.62%	19.64%	3.29%	6.94%	2.06%	5.19%	2.16%	1.85%	1.95%	5.50%	4.83%	6.17%
Grupo.es	4.37%	8.01%	4.06%	3.99%	1.97%		6.62%	8.01%	4.91%	11.92%	17.49%	15.40%	4.95%	5.34%	0.39%	2.28%	0.31%
Espn	2.81%	7.72%	7.60%	1.70%	11.80%	5.28%		2.90%	12.17%	6.08%	9.73%	6.45%	3.71%	14.95%	0.59%	10.32%	0.34%
Free.fr	12.79%	10.80%	8.12%	5.62%	5.53%	17.89%	8.12%		10.37%	12.79%	9.77%	12.27%	9.68%	5.70%	0.78%	6.48%	3.11%
Real	2.84%	18.30%	14.09%	3.07%	4.51%	4.24%	13.16%	4.01%		3.54%	10.82%	4.07%	3.67%	13.79%	0.47%	19.00%	1.14%
T-online.de	2.75%	8.78%	3.18%	3.91%	1.93%	14.86%	9.50%	7.14%	5.11%		36.85%	19.73%	10.13%	2.89%	0.58%	3.57%	0.34%
AOL	1.60%	11.31%	5.94%	6.47%	3.37%	15.08%	10.51%	3.77%	10.81%	25.49%		21.05%	4.60%	6.47%	0.93%	10.28%	0.37%
Altavista	3.83%	7.66%	4.55%	6.05%	2.17%	20.59%	10.81%	7.35%	6.31%	21.16%	32.64%		6.57%	4.97%	0.93%	5.48%	0.52%
Gov.ca	8.48%	6.58%	5.70%	3.66%	2.44%	8.68%	8.14%	7.60%	7.46%	14.25%	9.36%	8.62%		8.55%	0.61%	4.61%	0.20%
Hollywood	1.72%	17.01%	5.04%	1.34%	1.64%	5.94%	20.84%	2.84%	17.79%	2.58%	8.35%	4.13%	5.43%		0.69%	21.23%	0.73%
Sina.cn	0.69%	1.09%	0.52%	0.58%	6.15%	0.58%	1.09%	0.52%	0.81%	0.69%	1.61%	1.04%	0.52%	0.92%		2.70%	17.08%
WMHome	2.66%	16.66%	11.25%	10.03%	4.10%	2.57%	14.57%	3.27%	24.81%	3.23%	13.43%	4.62%	2.97%	21.50%	2.05%		15.09%
News.tw	0.36%	0.27%	1.78%	2.71%	5.33%	0.36%	0.49%	1.60%	1.51%	0.31%	0.49%	0.44%	0.13%	0.76%	13.19%	15.37%	

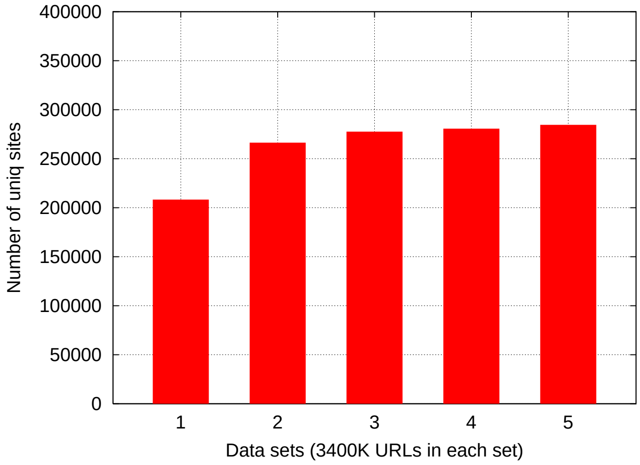


Fig. 18. Crawler “warming” up period

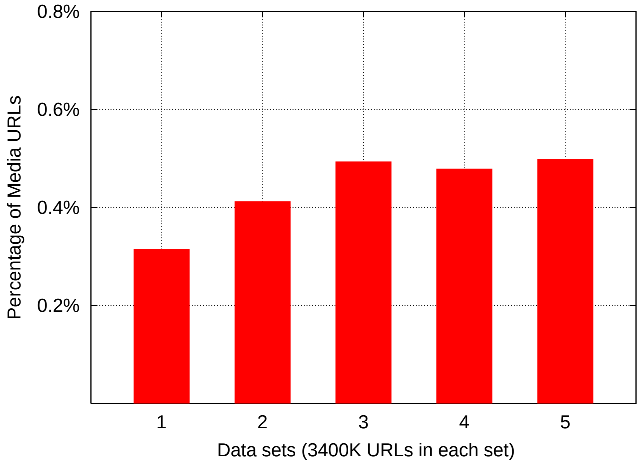


Fig. 19. Percentage of media URLs from each data set

URLs found in each set, depicted in Figure 19. For the data sets after the first 6.8 million URLs, the overall percentage of media URLs found is nearly constant, around 0.5%, suggesting that a larger number of URLs will not change the results.

C.2 More results on percentage of media URLs

As we compare the number of starting point and the geometrical location of the stating points, we found out that the the starting points is not affected much by the number of starting points. Figure 20 depicts that over 6 starting points of data will result in a relative stead percentage. And as shown in figure 21, USA starting points produced a higher overall percentage of media URLs than those starting point outside USA. However, the difference is not big enough to be considered.

C.3 Effects on tail analysis

To evaluate the effects of sampling issues on our heavy tail analysis, we also compare the complementary cumulative distribution of the video durations. Those results are shown in figure 22, figure 23, and figure 24. By comparing the shape of the tail in each figure, we conclude that after crawling 10.2 million URLs, we can get a steady CCDF tail shape, while 12 starting points if we consider the number of stating points.

However, we get different shapes in the comparison of USA and non-USA starting points. Non-USA data site come up with a longer tail, which means the data from non-USA starting points have a small number of long duration video clips.

D. Additional Results

D.1 Crawling statistic

The Web crawling took place between Feb 13, 2003 to March 18, 2003 and totally crawled 17 million URLs starting from 17 different domains.

The number of media URLs for each type of media and the number of unique media URLs are listed in Table VIII. The column labeled “Percentage” is the percentage of that media type over the total number of media URLs. The column labeled “Unique” is the ratio of unique URLs over number of URLs of that particular media type. We can

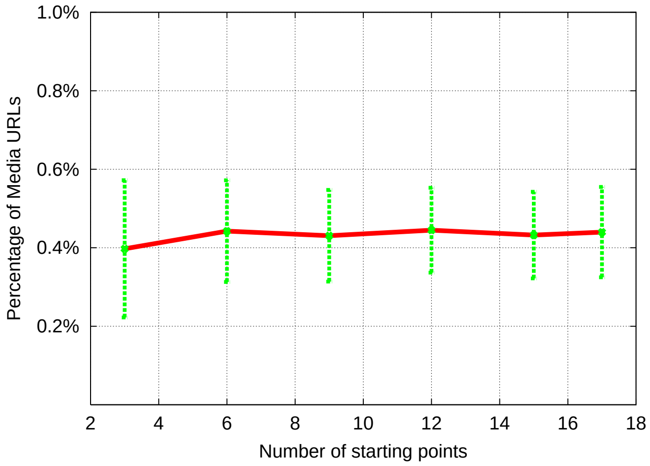


Fig. 20. Media and number of starting points

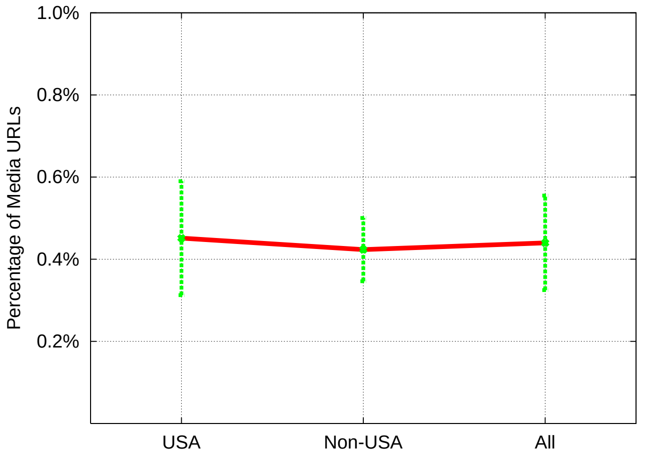


Fig. 21. Media from USA and non-USA starting points

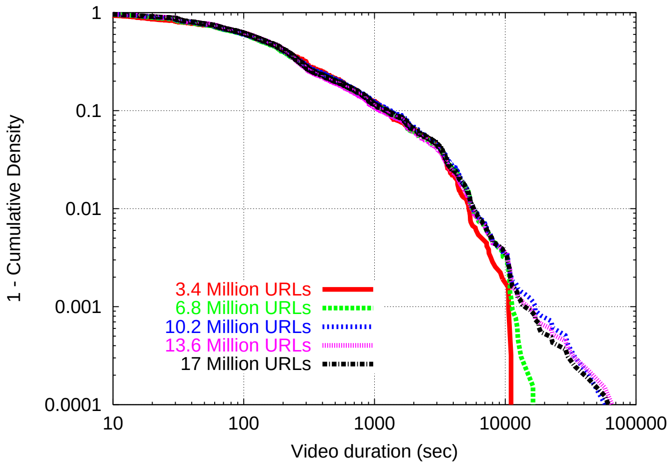


Fig. 22. Video Duration and number of URLs

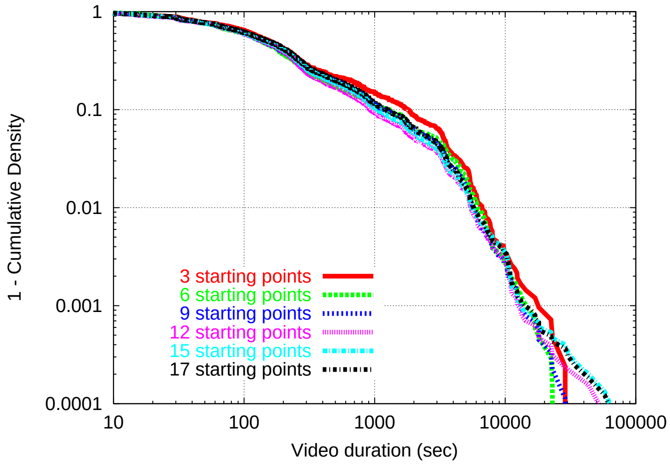


Fig. 23. Video Duration and number of starting points

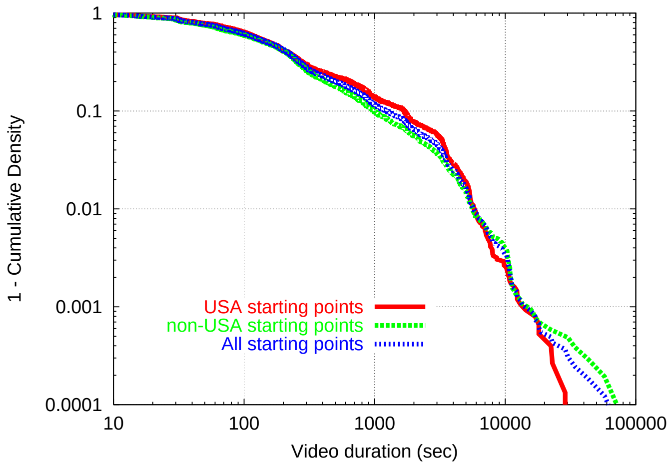


Fig. 24. Video duration from USA and non-USA starting points

see the most popular media type over the Internet is Real Networks media and Microsoft Windows media. For audio only URLs, the most popular media types are MP3s.

The media URLs distribution over the 17 domains are shown in Figure 25. Over all domains, Real Networks media has the largest population, except in the data set that started in South Korea; that data set has a slightly larger portion of Microsoft Windows media URLs.

Crawling times are highly related with the round-trip time from the server to the crawling client. From Figure 26 we can see that the Asian domains have significantly longer crawling times for the same number of URLs (1 millions URLs, in our experiments).

D.2 Media URLs Validation

During analysis, the Real Media, Windows Media and Quick Time Media URLs that we could stream we labeled as "valid". The valid ratios over all domains and for each type of media clips are shown in Figure 27 and Figure 28,

	Domains
American	Canada, USA
Asian	China, Japan, South Korea, Taiwan
European	France, Germany, Italy, Spain, UK

TABLE VII
MOST BROADBAND CONNECTED COUNTRIES

Media Type	URLs	Unique	Percentage	Unique
RM	33443	23405	42.74%	69.98%
WM	16360	13948	25.47%	85.26%
MP3	13566	10277	18.77%	75.76%
QT	2898	2137	3.90%	73.74%
MPEG	2580	2155	3.94%	83.53%
WAV	2201	1558	2,85%	70.79%
AVI	1255	1073	1.96%	85.50%
AU	406	209	0.38%	51.48%
Total	72709	54762	100.00%	75.32%

TABLE VIII
MEDIA URLS RESULTS

respectively. To find out the relationship between Internet traffic status and the validation rate, we consulted the Internet traffic report[21] during our media analysis. We find out domains such as China and Taiwan have a lower Traffic Index[22]. The Internet traffic status might be one issue that affect invalid ratio for the media URLs.

From Figure 28, the different media types have similar valid ratios according to our analysis, with a valid ratio around 72% to 76%.

D.3 Analysis on unavailable URLs

As we discusses in last section, there are approximately 24% to 28% URLs are unavailable but still being linked to web pages. 24% to 28% is a considerable large portion of the whole crawling result. We perform another simple tests on the error events caught from our Media analyzer.

We crawled 1 million URLs from Altavista on Nov. 25, 2003, and applied the Media Analyzer to the URLs on Nov. 26, 2003 (Within 24 hours, identical to what we did in this research). The Windows Media URL available rate is 73%, while the count of available Windows Media URL 437 out a total number 601. Table VII-D.3 lists all the

[21] http://www.internettrafficreport.com/
[22] The *Traffic Index* is a unit to measure to Internet traffic status. The higher, the better Internet condition, meaning low congestion, delay, and loss

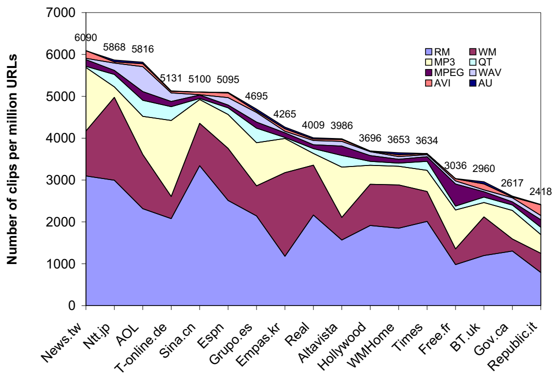


Fig. 25. Number of media URLs out 1 million URLs

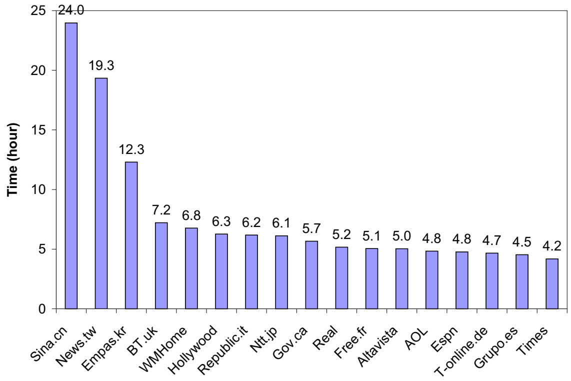


Fig. 26. Crawling time for 1 million URLs

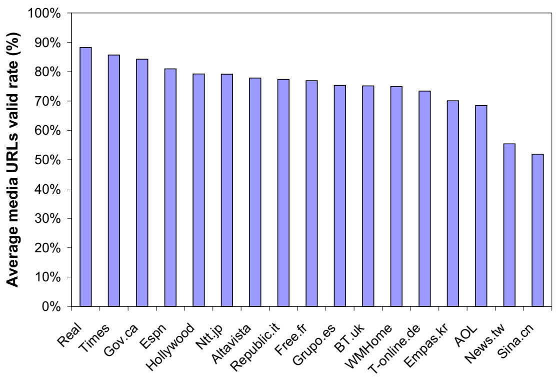


Fig. 27. Valid ratio for media URLs for each domain

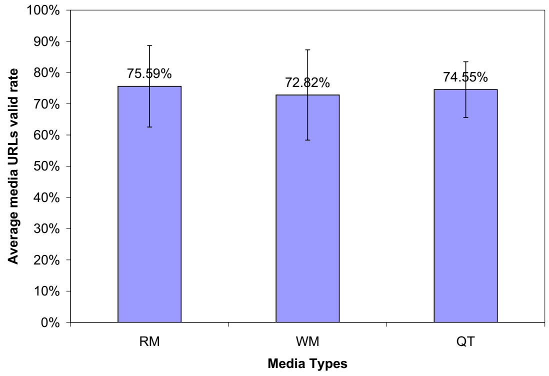


Fig. 28. Valid ratio for each media URLs

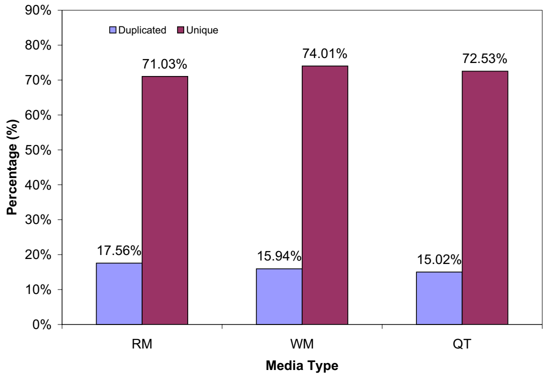


Fig. 29. Duplicated and unique for each media type

Errors	Count	Percentage
Source filter can't be loaded	53	8.8%
File not found	46	7.8%
Can't connect server	37	6.2%
Authorization Fail	23	3.8%
others	5	0.8%

TABLE IX
ERROR RETURNED FROM UNAVAILABLE URLS

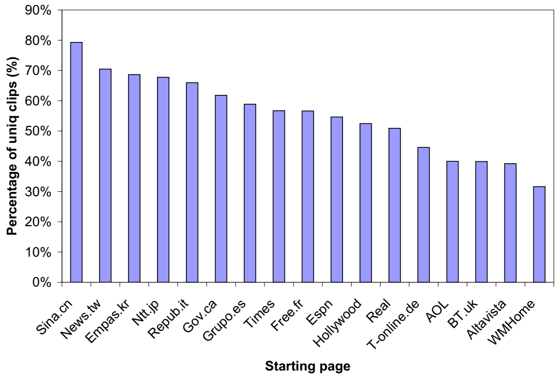
errors returned from Media Analyzer.

The most unavailable URLs is due to the "Source filter can't be loaded". However, by manually testing some of the URLs with "Filter" error, we found out that the error is not caused by filter in most of the cases. A few contents providers use embedded Windows Media Player in HTML page to play the contents or use CGI to redirect the connect request to another HTML page. Since WMP is not able to open HTML page using media filter, it will generate the Filter error. Therefore, the "filter not able to load" doesn't really means the media URL is unavailable. Therefore, we didn't consider it when we discuss the most popular causes of URLs unavailable.

D.4 Duplicate URLs

We did some analysis on the percentage of duplicated URLs and Unique URLs. As shown in Figure 29, the duplicated URLs may show up in multiple data sets, even though they started from different starting pages. We also created a list of the 10 most duplicated media clips for each media type, shown in Table XI.

The unique URLs contribution from each starting page is another interesting issue to examine. However, since the



crawling is not limited to a specific domain, the country starting page is not necessarily a data set entirely in from the country. From Figure VII-D.4 we can see the most unique page are from non-English speaking Asian countries, while the U.S. and other English speaking countries have fewer unique URLs for each data set. This is evidence of culture and language barriers for Web sites links.

D.5 Multiple Encoded Level Analysis

Media encoded bit rates had been discussed in Section III-B.1 and Section III-B.2, in detail. However, we put together a complete range of data from our measurement of Windows Media and Real Media. The encoded bit rates were divided in four ranges: $<=$56 Kbps for modem connections; 56 Kbps -768 Kbps for general broadband connections; 768 Kbps - 1.5 Mbps for higher broadband and T1 connection; and $>$1.5M for other broadband and LAN

Media	<=56K	56-768K	768K-1.5M	>1.5M
RM Audio	93.0%	7.0%	0	0
RM Video	31.5%	67.2%	0.8%	0.5%
WM Audio	83.4%	16.6%	0	0
WM Video	23.6%	74.2%	1.8%	0.4%
All Audio	90.8%	9.2%	0	0
All Video	27.9%	70.4%	1.2%	0.5%

TABLE X
ENCODED BIT RATE RANGES

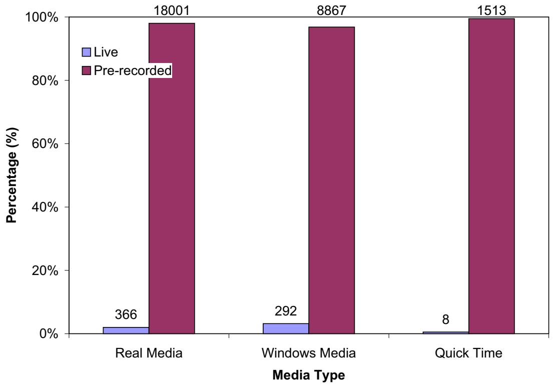


Fig. 30. Live and Pre-recorded media for each type of media

connections. The complete range distribution is shown in Table X. The units are in Kbps.

D.6 Live versus Pre-Recorded

Figure 30 depicts the ratio of live vs pre-recorded content across three streaming media applications. From our media analysis, there are 2.3% live streaming URLs out of 23,381 total valid URLs. Although all of the 3 major commercial media streaming applications support live streaming, most live content is provided in Microsoft Windows media and RealNetworks media formats.

D.7 Video Aspect Ratios

Figure 31 depicts a cumulative density graph of of aspect ratios. Most of the video clips (70.1% out of all videos) have an aspect ratio that follows the Academy Standard of Television (4:3 or 1.33:1). However, 7.1% of videos have an aspect ratio of 11:9 or 1.22:1, which are the aspect ratios of CIF (Common Intermediate Format: 352 x 288) and QCIF (Quarter CIF: 176 x 144). The Quick Time videos have the largest range of aspect ratios, most of them are from HDTV (16:9 or 1.78:1) and variant film standards (1.85:1, 2.35:1, etc.).

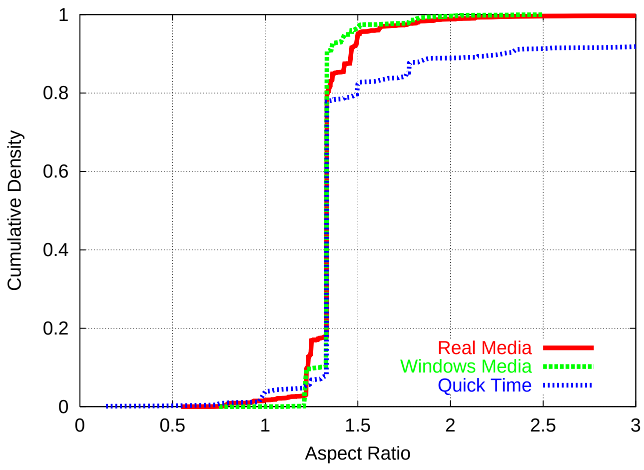


Fig. 31. Cumulative density function of video aspect ratio

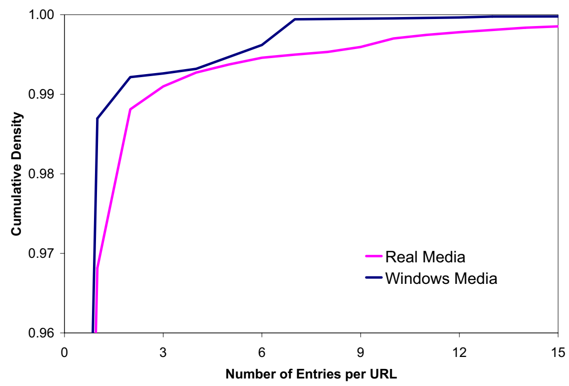


Fig. 32. Number of entries in each media URL

D.8 Playlists

Both Windows Media Server and Real Media Server provide support for server-side playlists on the media servers. A server-side playlist is used to simplify the clip management by the content provider and also to provide additional wrapper functions. These function can be used to specify additional content to be played out before or after the content requested by the user, or to provide a single URL composed with multiple items requested by the user.

Figure 32 depicts a cumulative density graph of the number of items in one media URL. The number of items in one media URL is typically 1, indicating the URL is linked to the clip directly or the server side playlist with only one item in it.

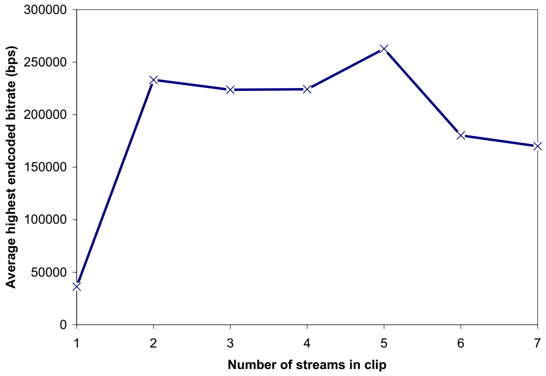


Fig. 33. Level of encoded streams vs. Max. encoded bitrate

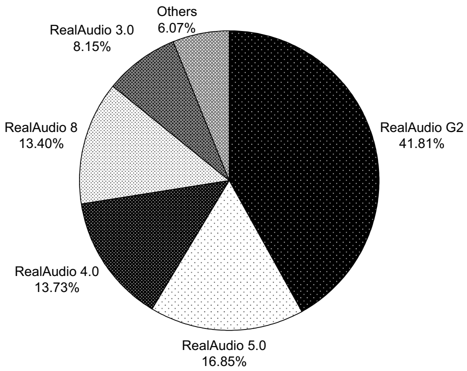


Fig. 34. Breakdown of RealMedia Audio Codecs

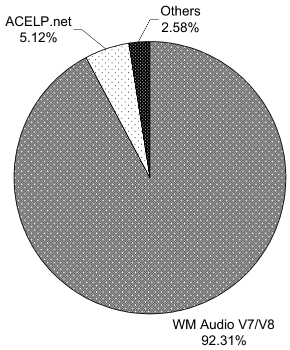


Fig. 35. Breakdown of WindowsMedia Audio Codecs

D.9 Multiple encoded level analysis

We also tried to find some relationship between the maximum encoding bitrate and number of streams in clip. As seen in Figure 33, there are no clear visual corelation between those two parameters. That is, even a low encoded bitrate clip could have a large number of encoded bitrate levels.

D.10 Codec Results

Section III-C discusses the video codecs for Microsoft Windows video and RealNetworks video. Figure 34, Figure 35 and Figure 36 depict the codecs used for Real Networks audio, Microsoft Windows audio and Apple QuickTime video, respectively.

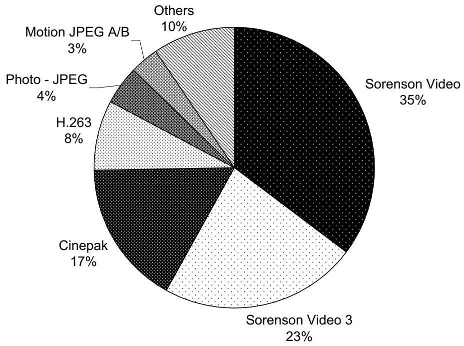


Fig. 36. Breakdown of QuickTime Video Codecs

Times	URLs
RM:	
15	http://www.bbc.co.uk:80/worldservice/news/summary.ram
15	http://europe.real.com:80/smil/vidzone.smil
14	http://www.npr.org:80/atc3.smil
14	http://www.bbc.co.uk:80/go/homepage/int/sport/vi/-/newsa/n5ctrl/sport/bulletins/video_daily.ram
14	http://www.bbc.co.uk:80/go/homepage/int/sport/au/-/news/olmedia/cta/sport/programmes/bulletins/daily.ram
14	http://www.bbc.co.uk:80/go/homepage/int/news/vi/-/newsa/n5ctrl/tvseq/n24.ram
14	http://www.bbc.co.uk:80/go/homepage/int/news/au/-/news/olmedia/cta/progs/rn/bulletin.ram
13	http://www.undp.org:8080/ramgen/oa/ronzid.rm
13	http://www.npr.org:80/realmedia/news2a.ram
13	http://www.npr.org:80/realmedia/24hour.ram
WM:	
13	http://www.npr.org:80/windowsmedia/programstream.asx
13	http://www.npr.org:80/windowsmedia/newscast.asx
13	http://www.nasdaq.com:80/reference/JetBlue_WPP_Sun.wmv
13	http://www.nasdaq.com:80/reference/Cisco-Intel-Staples.wmv
12	http://www.npr.org:80/webevents/npr.asx
12	http://www.nasdaq.com:80/reference/DELL_MSFT_SBUX.wmv
12	http://www.nasdaq.com:80/reference/Costco-Staples-Starbucks.wvm
12	http://www.nasdaq.com:80/reference/AppliedMaterials-Costco-Dell.wmv
11	http://www.npr.org:80/webevents/news.auto.asx
10	http://www.nab.org:80/conventions/nab2003/exhibitors/video/avid.wmv
QT:	
10	http://www.perl.org:80/yapc/2002/movies/2002-06-24-perl6-handwaving.mov
8	http://www.iscb.org:80/webmovs/bourne03.mov.mov
8	http://www.iscb.org:80/webmovs/bourne02.mov.mov
8	http://www.iscb.org:80/webmovs/bourne01.mov.mov
7	http://reason.com:80/ReasonMagazine.mov
7	http://alberta.indymedia.org:80/uploads/kyotororbust1.mov
6	http://planetmirror.com:80/pub/movie_trailers/L2Towers.mov
6	http://downloads.warprecords.com:80/bushwhacked2.mov
5	http://www.gfdl.noaa.gov:80/ jps/images/gallery/fran_anim_title_A_D.qt
5	http://www.gfdl.noaa.gov:80/ jps/images/gallery/emily_A2_C_B_2x_q3.qt

TABLE XI
TOP 10 DUPLICATED MEDIA URLS