Streaming Video Traffic : Characterization and Network Impact

Jacobus van der Merwe, Subhabrata Sen, Charles Kalmanek
AT&T Labs Research
{kobus,sen,crk}@research.att.com

Abstract

The emergence of the Internet as a pervasive communication medium, and the widespread availability of digital video technology have led to the rise of several networked streaming media applications such as live video broadcasts, distance education and corporate telecasts. This paper studies the traffic associated with two major categories of streaming content - on-demand streaming of pre-recorded content and live broadcasting. Using streaming logs from a commercial service, we analyze the traffic along a number of dimensions such as session characterization, object popularity, protocol choice and network load. Among our findings, (i) high bandwidth encodings account for about twice as many requests as low bandwidth ones, and make up about 94% of the traffic, (ii) Windows Media streams account for more than 75% of all requests, when the content is available in both Windows and Real formats, (iii) TCP based transport protocols dominate over UDP being used in about 70% of all bytes transfered (iv) Object popularities exhibit substantial skew with a few objects accounting for most of the load, (v) A small percentage of IP addresses (or routing prefixes or origin autonomous systems (ASes)) account for most of the traffic demand across a range of performance metrics. This last behavior suggests that substantial bandwidth efficiency can be realized with a distribution infrastructure comprised of a relatively small number of replicas, placed close to the heavy-hitter ASes. We also found very high variability in terms of the traffic volume with an order of magnitude or more increase in the offered load over tens of minutes, suggesting the potential benefit of a shared infrastructure that can exploit statistical multiplexing.

1 Introduction

The emergence of the Internet as a pervasive communication medium, and the widespread availability of digital video technology have led to the rise of several networked streaming media applications such as live video broadcasts, distance education, corporate telecasts, etc. It is therefore important to understand and characterize the traffic associated with these applications in terms of end-system behavior and network impact, in order to develop workload models as well as insights into network traffic engineering and capacity planning for such applications.

Demand for streaming media is surging. According to a recent industry study [5], there were 60 million people listening to or watching streaming media each month, 58 US TV stations performing live webcasting, 34 offering on-demand streaming media programs, and 69 international TV webcasters. The study also finds that 6000 hours of new streaming programming are created each week. The ongoing deployment of an array of broadband last mile access technologies such as DSL, cable and high speed wireless links will ensure that a growing segment of the population will have sufficient bandwidth to receive streaming video and audio in the near future. According to Forrester Research [8, 19], by 2005, 46 million homes in the US alone will have broadband Internet connectivity. This is likely to dramatically increase the use and popularity of streaming media.

However, due to the high bandwidth requirements and the long-lived nature (tens of minutes to a couple of hours) of digital video, server and network bandwidths are proving to be major limiting factors in the widespread usage of video streaming over the Internet. Audio and video files tend to be large in size, e.g., 4.8 MB for a 5 minutes long 128 Kbps MP3 audio clip, 450 MB for a 2 hour long MPEG-4 video clip encoded at 500 Kbps.

There is a rich body of literature on end-system and network mechanisms for delivering streaming media across the Internet. There has been a significant amount of work modeling the multi-timescale bursty bandwidth profile of compressed variable-bitrate (VBR) videos [7, 16, 10, 11], and on tech-

niques [18, 17, 21] for efficiently delivering such streams across the network. A number of experimental studies address the quality of service (delay, loss, jitter etc.) experienced by multimedia streams transmitted across networks [23, 2, 13, 12, 22]. However, there has been very little work on characterizing requests for streaming content and the associated server and network workload distributions for such requests. Historically, a main reason for this has been the paucity of streaming video content and the absence of large user base for whatever content was available. Only recently have a number of factors, such as the growth in broadband users, and the development and spread of new compression techniques such as MPEG-4 that can deliver good quality at low bandwidths, converged to a point where many large content providers now offer a range of streaming content such as news, shopping, short video clips and trailers, and entertainment. In this paper, we analyze session logs from a commercial streaming service, and analyze the workload for two different types of content - stored on-demand media, and a live, real-time streaming presentation.

Existing empirical work on streaming media can be categorized as either measuring the performance of individual streams across the network, or as characterizing streaming workloads. [15] examined interactions of around 200 University users in 1997 with a courseware application composed of lecture notes (in HTML) with accompanying synchronized audio lectures. [14] analyzed five audio traces (RealAudio packet traces corresponding to long-lived Internet radio channels at Broadcast.com), ranging from 83 seconds to 18.2 hours long, and containing up to 1460 distinct audio data flows and 1397 distinct user IP addresses.

Prior work on streaming video workload characterization, includes [4], which analyzes 58808 RTSP sessions from 4786 University users to 23738 distinct streaming media objects from 866 servers across the Internet, and compares the characteristics to Web workloads. [1] analyze streaming video workload associated with two University course projects.

This work is based on log files containing several orders of magnitude more sessions and users than any previous work. We extracted and analyzed 4.5 million session-level log entries for two streaming services over a period of 4 months. We also integrated information from the streaming logs with BGP (Border Gateway Protocol) routing information gleaned from multiple border routers on a tier-1 ISP. We used this combination of streaming and routing information to study the network implications of streaming traffic. Specifically we used network routing-aware clustering techniques [9] to determine the traffic distribution for different IP address prefixes and ASes. To our knowledge, this is the first network-routing-aware study of streaming traffic distributions.

The remainder of the paper is organized as follows. Section 2 presents our methodology for analyzing the streaming traffic as well as the data set we used. We report our analysis and results in Sections 3-7. Section 3 discusses the session composition by protocol family, stream bandwidth and transport protocol. In Section 4 we consider the traffic distribution at different levels of aggregation and its implications for content distribution. Traffic dynamics over various time-scales as well as object popularity is investigated in Section 5. The session characteristics of a few highly popular objects is presented in Section 6. Section 7 contains a summary of our results and we conclude the paper in Section 8 with a conclusion and indication of future work.

2 Methodology

We first outline our data collection and analysis methodology.

2.1 Measurement approach

This study is based on an analysis of a large dataset of application level session logs from a commercial streaming service. A session corresponds to all the interactions associated with a single client requesting and viewing a clip containing both audio and video. From the log data, we analyze the breakdown of traffic by protocol family, stream bandwidth, and transport protocol to get a quantitative understanding of the breakdown of streaming traffic for these key parameters of interest.

A streaming session is initiated when a new request for a streaming object is received at a streaming node. During the session, while the video is being streamed to the requesting client, user requests for interactive operations (such as fast forward, rewind, pause, restart) can arrive. The session terminates

either when the client sends a termination request, or due to some error situation. At termination, a single entry is created in the log summarizing a range of information for that session. The fields in each log entry include: requesting IP address, particulars of requested resource, whether the file is a Real or Windows Media object, transport protocol used for streaming (TCP or UDP), total data transmitted, session end-time, total session time, status/error codes, etc. Content providers utilizing streaming services typically develop their own naming convention for streaming objects from which further information about the stream (e.g. its encoding rate) can be determined.

From the streaming session logs we extracted all the log entries associated with two particular streaming sites that serve different types of content - stored on-demand media, and a long-lived, real-time streaming presentation. For the on-demand data set session logs were collected over a four month period of time, whereas for the live data set, logs were collected for a two month period.

We characterize the workload by looking at a number of different measures: number of requests, traffic volume, number of active connections, etc. We then look at these workload measures at different levels of address aggregation, from client IP address, to network prefix and Autonomous System (AS). This aspect of the study focuses on understanding the spatial (topological) workload distribution.

In order to better understand the traffic dynamics, we also present a time series analysis and present several measures such as traffic volume over several time scales of interest, ranging from several minutes to 4 months. This type of analysis is potentially useful in understanding long-term trends in traffic, as well as shorter time-scale variations such as flash crowds. Longer time-scale trends are important for capacity planning, while shorter time scale variations are important both in planning for load peaks as well as in developing load balancing strategies if streaming services are supported via a content distribution network.

Analyzing the traffic at the level of individual IP addresses is useful for several reasons. First, a single session entry in the application log always corresponds to a single client, allowing us to explore intra-session client behaviors. Second, IP level information provides a fine-grained view of the demand and load distribution across the network. For example, if a single user generated a substantial amount of request traffic, this would show up in an IP level analysis. Due to the use of techniques such as dynamic address assignment, NAT (Network-address-translation) and forward proxy servers at the edge of the network, an IP address may not correspond to a unique client in general. However, since each IP address maps to a unique interface (subnet) at the edge of the network, it is still useful for understanding the overall traffic distribution.

We use network prefixes as an intermediate level of aggregation. An IP router uses longest prefix matching to map from the destination IP address of an incoming packet to a list of prefixes in its forwarding table that determine the next-hop router to which the packet should be forwarded towards its destination. All packets mapping to the same prefix are forwarded to the same next hop router. Hence, the prefix-level aggregation allows us to group IP addresses (and clients) that are topologically close together from a network routing viewpoint. All IP routing decisions are made at the granularity of the routing prefix, and so understanding traffic at this level is important for the purpose of network or CDN traffic engineering. For similar reasons, we also study the traffic at larger routing granularities, including AS level (all prefixes belonging to a single AS are part of a single administrative domain). For instance, if we observe that a few prefixes (or ASes) account for a substantial fraction of the total traffic, this might be used by network designers responsible for setting up ISP peering or transit service relationships or placing network servers, in order to optimize the network, reduce network load, and potentially improve streaming performance.

2.1.1 Integrating Routing with Streaming Data

As mentioned above, we correlate the streaming logs with routing data collected from multiple routers across a Tier-1 ISP. BGP (Border Gateway Protocol) table dumps obtained from the routers each day are collated to obtain a table of (routing prefix, originating AS number(s)) pairs for that day. In our data, we note that we do not necessarily have a unique mapping from a client IP address to a unique network prefix or originating AS. A routing prefix might be mapped to multiple originating ASes if for example, multiple ASes advertise the same prefix. In addition, because IP routing is dynamic, the

routing table entries can change: a prefix can appear or disappear, or its mapping to an AS can change. When looking at data for time-scales up to a day, we integrate the routing information with the session logs as follows: for each session log entry, we use longest prefix matching on the requesting IP address to determine (from the table for that day), the corresponding network prefix(es) and originating ASes for that session. If this does not result in a unique mapping, we assign the traffic for a address mapped to both ASes AS1 and AS2 to a separate logical AS represented by AS1+AS2. Since we look at logs over a period of four months, we need to consider carefully how to combine routing information with the streaming log data for time scales longer than a day. To understand the extent of routing information change, we collected the routing data for a 22 day period in our 4 month logging period. We then developed a list of prefix-AS pairs by merging the prefix-AS mappings into a single combined table, and discarded any prefix-AS mapping for which there was no corresponding streaming log entry. This combined table contained some prefixes that map to multiple ASs.

For the entries in the combined table, we determined the number of days that each prefix and prefix-AS pair appeared. This list contains 30843 unique prefixes of which 26781 (87%) were present all 22 days. In addition, out of a total of 31247 unique prefix-AS pairs, 26485 (85%) were present all 22 days. This suggests that the large majority of the prefixes and prefix-AS pairs are stable across the 22 days.

The results of the analysis presented in the rest of the paper use routing table data from selective single days in the log file analysis. We believe, based on the above observations, that the results are not significantly affected by this simplification.

2.2 Description of Data Set

For this study we used session level logs from two data sets:

- On demand streaming of pre-recorded clips from a current affairs and information site - the *On Demand* data set.
- A commerce oriented continuous live stream-the *Live* data set.

Table 1 shows the collection period, number of sessions, number of distinct requesting IP addresses and number of distinct requesting ASes for the two data sets. For each data set, the total traffic over the measurement period was of the order of several Terabytes. For *On Demand*, all content is offered in both Windows Media (MMS) and Real Media (Real) formats, and for each format, the video is encoded at two different bandwidths: a higher bandwidth version at 250 Kbps and a low bandwidth version at 56 Kbps. There were 4296 unique clips accessed during the measurement period for this set. *Live* consisted of a single 100 Kbps stream in Windows Media format.

3 Session Composition

We profiled the sessions in terms of protocol family (Real and Windows Media), stream bandwidth, and transport protocol used. Note that for *On Demand*, all content is offered in both media formats, and as both high and low bandwidth encodings. Hence the choice of a particular protocol family or stream bandwidth will be driven by a combination of client-side and network factors such as user preference, network connectivity and software deployment.

Table 2 reports the breakdown for On-Demand, and Table 3 depicts the transport protocol breakdown for *Live*. These breakdowns show relatively little change across the different months and are considered in more detail below.

3.1 Composition by Protocol Family

Table 2 shows that Windows Media requests dominate by far over the four months - there are 3.35 times as many sessions and 3.2 times as much traffic generated by Windows media requests as compared to Real sessions. Note that the relative ratios are quite stable across the different months. Fig. 1(a)-(b) depicts the breakdown among the top ranked ASes that either generate 80% of all the requests or account for 80% of all the traffic, across the four months. We see that the overwhelming majority of these heavy-hitter ASes receive much more Windows traffic than Real. All this suggests a widespread prevalence and use across the Internet of the Windows Media software. This dominance

Data	Dates	Number of sessions (million)	Number of unique IPs (million)	Number of distinct ASes
On Demand	12/01/2001 - 03/31/2002	3.5	0.5	6600
Live	02/01/2001 - 03/31/2002	1	0.28	4000

Table 1: *Data set*: Statistics.

Dates	Metric (% of total)	Protocol Family		Bandwidth		Transport		
		MMS	Real	Low	High	Proprietary Streaming		HTTP
						UDP	TCP	TCP
Dec, 2001 - Mar, 2002	Requests	77	23	35	65	34	29	37
	Traffic Volume	76	24	5	95	28	45	27
Dec, 2001	Requests	77	23	35	65	32	28	40
	Traffic Volume	76	24	6	94	26	45	29
Jan, 2002	Requests	78	22	36	64	34	30	36
	Traffic Volume	78	22	7	93	30	45	25
Feb, 2002	Requests	75	25	35	65	39	29	32
	Traffic Volume	74	26	7	93	33	45	22
Mar, 2002	Requests	76	24	32	68	33	33	34
	Traffic Volume	76	24	6	94	29	45	25

Table 2: *On-Demand* : Popularity breakdown by protocol family, encoding bandwidth, and transport protocol, for a number of time intervals. Metrics are number of requests and traffic volume, expressed as a percentage of the total amount over the corresponding time period.

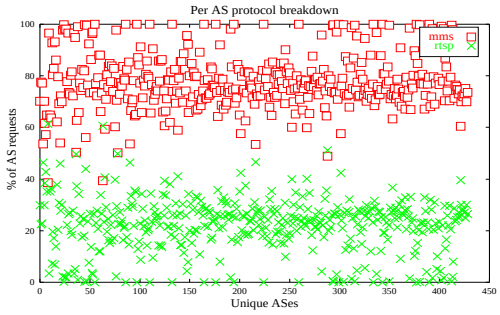
Dates	Metric (% of total)	Transport		
		Proprietary Streaming		HTTP
		UDP	TCP	TCP
Feb, 2001 - Mar, 2002	Requests	28	17	55
	Traffic Volume	17	38	47
Feb, 2001	Requests	30	18	52
	Traffic Volume	17	36	47
Mar, 2001	Requests	26	17	56
	Traffic Volume	16	36	48

Table 3: *Live*: Popularity breakdown by transport protocol. Metrics are number of requests and traffic volume, expressed as a percentage of the total amount over the corresponding time period.

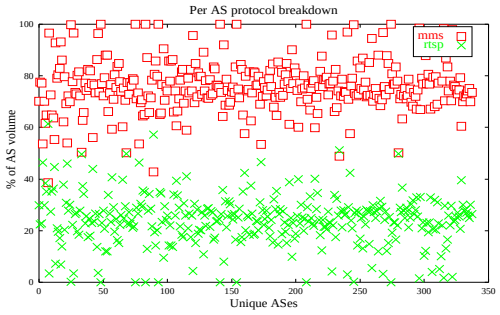
could be at least partially attributed to the Windows strategy of bundling the encoder, server and player software with their operating system. Still the fact that Real continued to command about 23% percent of the requests across the 4 months, suggests that at least for the time-being, content providers should continue providing the content in both formats.

3.2 Composition by Stream Bandwidth

We observe from Table 2 that overall there are almost twice as many sessions downloading (or requesting) high bandwidth streams compared to low bandwidth streams. The high bandwidth content accounts for 95% of the total traffic, and the relative ratios are nearly identical across the different months. Our logs reveal that the breakdown is similar within individual protocol families. High bandwidth content accounts for 67% and 94% of all MMS sessions and traffic respectively (60% and 92% of all Real sessions and traffic respectively). Given that these are streaming downloads, the above statistics seem to indicate that a large majority of requests for the streaming content are sourced by clients with good end-to-end broadband connectivity. Fig. 2(a)-(b) depicts the breakdown, by content bandwidth, among the top ranked ASes that either generate 80% of all the requests or account for 80% of all the traffic, across the four months. We find that for a large majority of these heavy-hitter ASes, sessions

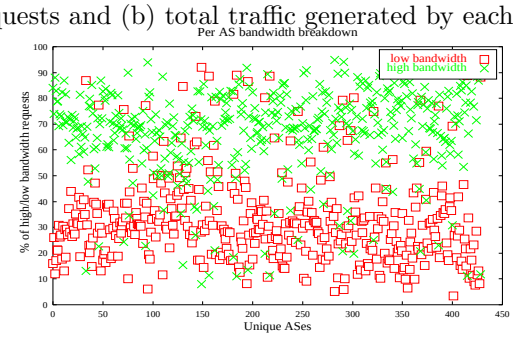


(a) requests

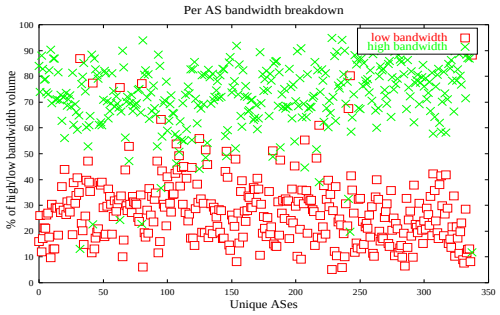


(b) traffic volume

Figure 1: On demand: per-AS protocol (Windows Media (mms) or Real (rtsp)) breakdown for ASes generating 80% of requests, and data volume. X-axis numbers the ASes. Y-axis is in percentage of (a) total requests and (b) total traffic generated by each AS.



(a) requests



(b) traffic volume

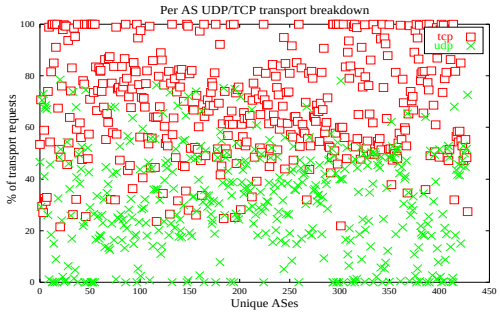
Figure 2: On demand: per-AS breakdown by stream bandwidth (high versus low encoding rate) for ASes generating 80% of requests, and data volume. X-axis numbers the ASes. Y-axis is in percentage of (a) total requests and (b) total traffic generated by each AS.

requesting high bandwidth dominate, both in number of sessions and generated traffic. 90% of all the ASes originated some sessions for broadband content for *On Demand*. For *Live* which is streamed at 100 Kbps, there were 4000 requesting ASes. All this suggests a fairly wide presence of clients with broadband connectivity (either at home or through corporate or campus LANs) across the Internet.

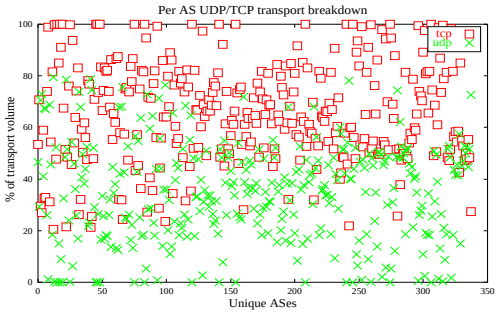
3.3 Composition by Transport Protocol

We next consider the transport protocol used to stream the video to the clients. Both Windows Media and RealNetworks recommend that the video be streamed using their respective proprietary streaming protocols running preferably over UDP. To overcome firewall restrictions, the protocol can also run over TCP. There is also the option to stream the clip using standard HTTP or some variant of it. This is the fall-back option for going through firewalls (almost all firewalls allow outgoing HTTP requests), and also for older versions of the player software.

For *On Demand*, Table 2 shows that for the Dec, 2001-March, 2002 period, the majority (63% of the sessions accounting for 73% of the traffic) use proprietary streaming protocols over either UDP or TCP. Still, a significant 37% of the sessions use HTTP, the recommended last option. In addition, overall 66% of all the sessions use TCP (HTTP or proprietary protocol), and only 34% use UDP. For the 100 Kbps *Live* stream, over Feb-March, HTTP is used by 55% of requests accounting for 47% of the traffic (HTTP appears to be more prevalent for *Live* than for *On Demand*), and overall, 72% of the sessions accounting for 83% of the traffic use TCP. As shown by the above tables, for both data sets, the overall breakdown between UDP, TCP and HTTP sessions remains similar across the months, though there are some variations in the actual percentages for each category. Fig. 3(a)-(b) show that TCP accounts for the majority of the traffic for most heavy-hitter ASes. This observed widespread use of TCP occurs in spite of the conventional wisdom that the congestion-control and reliability mechanisms in TCP make it less suitable than UDP for meeting the the real-time constraints associated with stream-



(a) requests



(b) traffic volume

Figure 3: On demand: per-AS breakdown by transport protocol (TCP (represented by blocks) or UDP (represented by "+")) ASes generating 80% of requests, and data volume. X-axis numbers the ASes. Y-axis is in percentage of (a) total requests and (b) total traffic generated by each AS.

ing. Firewall restrictions may be a key determining factor behind such widespread usage of TCP for streaming, even for high-bandwidth streams.

A consequence of the above composition is that the bulk of the streaming traffic, by virtue of using TCP, is still congestion-control friendly. We also find that the TCP sessions can be long - for example for *Live*, 9% of sessions using TCP are on for at least 20 minutes. This seems to indicate that even for high bandwidth streams, the quality of media streamed using TCP is considered good enough by a large proportion of the end-users to continue viewing the content for extended periods. This in turn again suggests that these clients experience good enough end-to-end connectivity that the TCP throughput is sufficient to deliver the video at its encoded bitrate.

4 Traffic Distribution

We next study how the traffic is distributed across the network at the IP, network prefix and AS aggregation grains. For *On Demand*, Figs. 4(a)-(b) plot the ranked CDF of (i) the number of requests generated by each entity, and (ii) the total traffic generated by each entity, where an entity is either an IP address, a network prefix or an AS. Fig. 4(c) plots the ranked CDF of the total number of unique IP addresses per prefix and AS. The ranked CDF is obtained by first ordering the IPs (or prefixes or ASes) in order of decreasing volume, and then plotting the cumulative volume for the ranked list. Figs. 5(a)-(c) present the corresponding plots for *Live*. The graphs reveal substantial variability in the number of requests as well as in the traffic volume among different IP addresses, prefixes and ASes. For *On Demand*, 75% of the total sessions and 80% of the traffic is attributable to just 30% and 20%, respectively of the total IP addresses. For *Live*, 94% of the total sessions and 96% of the traffic for *Live* is attributable to just 30% and 20%, respectively of the routing prefixes. We note that for each aggregation grain, the distribution of traffic volume and number of requests is more skewed towards a few heavy contributors (IP/prefix/AS) for *Live* compared to the distribution for *On Demand*. There is a similar difference between the distribution of requesting IP addresses at the prefix and AS levels for two datasets. The skew in the distribution of the number of sessions increases with larger aggregation grains - from IP to prefix to AS, for both data sets (Figs. 4(a) and 5(a)). The same behavior holds for the total traffic distribution at the prefix and AS levels for both data sets. However, the IP-level distribution of traffic volume exhibits the least and the most skew, respectively, among the different aggregation levels, for *On Demand* and *Live* (Figs. 4(b) and 5(b)). For both data sets, a few top-ranked ASes together account for almost all the requests as well as all the traffic. Fig. 4(c) shows that a tiny percentage of all the prefixes (or ASes) account for most of the requesting IP addresses.

We find that the ASes ranked highest in each of the three metrics have significant overlap. For instance, for *On Demand*, the top-ranked 300 ASes (5% of all the ASes) for all three rankings have 71% common members, while 310 ASes appear on the top 300 list for at least 2 of the three rankings. This suggests a high degree of positive correlation between the number of requests, traffic volumes and IP addresses for an AS.

We also found that a large proportion of ASes consistently remain among the top traffic contributors

across the months. Considering *On Demand* for instance, 207 ASes are among the the top-ranking 300 ASes (this set contributes around $79 - 80\%$ of the monthly traffic) for each month between Dec. and March. The significant skew in traffic contributed by different ASes as well as the persistent high ranking of many heavy-hitter ASes suggests that there can be benefits from distribution schemes that target the heavy hitter ASes. We shall explore distribution architectures in more detail next.

4.1 Impact on Content Distribution

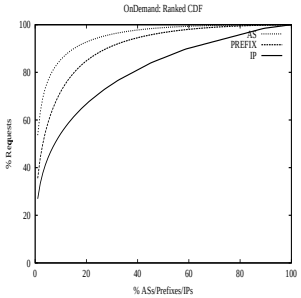
In this section we use the data from our streaming logs together with BGP routing information from a tier-1 ISP to investigate different tradeoffs for the distribution of streaming content. In all cases we assume that the streaming content is being served from a hosting center in the tier-1 ISP or through a content distribution network (CDN) originating from the ISP. The user perceived quality of a streaming presentation is determined by many factors including encoding rate, frame rate and image size. However, from a network perspective sustaining the required bandwidth and reducing the packet loss appears to be the most important factors in determining streaming quality. Maintaining a (near) congestion free end-to-end connection between a streaming client and server is therefore important to maintain streaming quality. AS hop count is in general not a good indicator of the congestion that might be experienced in traversing an end-to-end path other than the fact that the probability of experiencing congestion increases with every network element on the path. However, direct connectivity to a tier-1 ISP normally avoids congested public peering links. Also, tier-1 ISPs are normally well connected with other tier-1 ISPs allowing users to benefit from their collective rich connectivity. We therefore make the assumption for this discussion that a low number of AS hops (e.g. 2) between a tier-1 ISP and a streaming client will in general ensure adequate streaming quality.

First we determine how much of the content would be served to clients no more than one AS hop away. This is shown in the first lines of Table 4 and Table 5 for the month of March for the *On Demand* and *Live* data sets respectively. (We performed the analysis across all of the months in the data set and observed similar results.) We consider the traffic volume, number of IP addresses and the number of ASes that would fall in this subset of the data expressed as a percentage of the totals for the time period. For both data sets the percentages of volume and number of IP addresses exceed 50% even though less than 20% of the ASes are covered. This is as expected given that the BGP data is from a tier-1 ISP which is well connected to other major networks and given the highly skewed per-AS distribution of the data that was presented in Section 4.

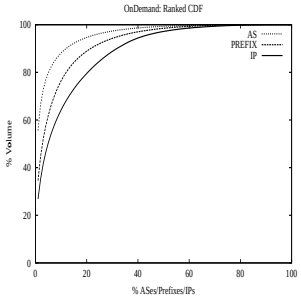
Next we consider content that would be served to clients no more than two AS hops from the tier-1 ISP. The results for this analysis is shown in the second lines of Tables 4 and 5. The *On Demand* data set show substantial increase in all three metrics considered. The *Live* data set on the other hand show a similar increase in the percentage of ASes covered, but show only a modest increase in the volume and number of IP addresses. This seems to suggest that in the case of the *Live* content a number of significant contributor ASes fall outside the 2 AS hop boundary.

Given the very skewed nature of the per-AS distributions presented in Section 4 we next considered the effect of selective additional arrangements with consistently heavy contributing ASes. In practice, such arrangements could take the form of content internetworking or peering relationships with such ASes, or replica placement at or near to such ASes. We determined the set of consistently significant contributor ASes as follows. For each month covered by our data we determined the list of top ASes that contributed 90% of the traffic volume. We then generated a list of ASes for each month corresponding to the ASes in the 90% list but not in the one-AS-hop-or-less list. Finally we picked the set of ASes that was present in all of these monthly lists (across both data sets) to make the *consistent contributor AS list* which consisted of 116 ASes.

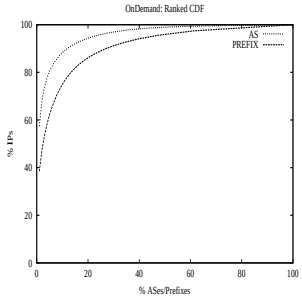
Combining the consistent contributor AS list with the one-AS-hop-or-less list corresponds to a content distribution approach where a service provider attempts to reach clients either directly through one AS hop or by selective peering or replica placement with ASes. The result of such an approach with our data is presented in the third lines of Tables 4 and 5. As expected the AS coverage in both cases increase very little as a relatively small set of ASes were selected. There is roughly a 40% and a 30% improvement in both the traffic volume and the number of IP addresses respectively for the *On Demand* and *Live* data sets. In the case of the *On Demand* data



(a) On Demand: % of requests against % of IPs/Prefixes/ASes

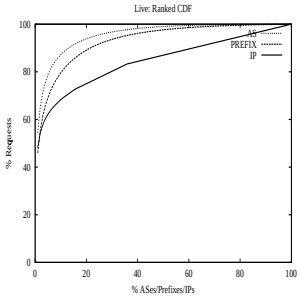


(b) On Demand: % of volume against % of IPs/Prefixes/ASes

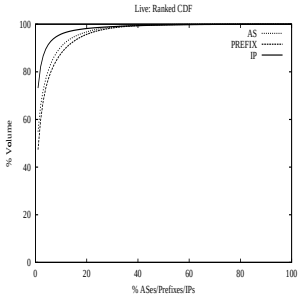


(c) On Demand: % of IP addresses against % of Prefixes/ASes

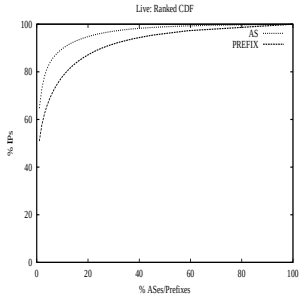
Figure 4: *OnDemand*: Ranked CDF plots



(a) Live: % of requests against % of IPs/Prefixes/ASes



(b) Live: % of volume against % of IPs/Prefixes/ASes



(c) Live: % of IP addresses against % of Prefixes/ASes

Figure 5: *Live*: Ranked CDF plots

set the improvement is less significant than for the two-AS-hops-or less approach, whereas for the *Live* data set the improvement is more significant.

Finally we repeated the same exercise but included on the consistent contributor list only those ASes not in the two-AS-hop-or-less set. The number of ASes in this set is only 15. Combining this AS set with the two-AS-hop-or-less set corresponds to an approach where the service provider combines the coverage provided by existing peering arrangements with selective peering or replica placement in a small number of heavy contributing ASes. The result for this approach is shown in the last lines of Tables 4 and 5.

While our data set is not large enough to make general conclusions, the analysis suggests that:

- A tier-1 ISP covers a significant portion of endpoints through 2 or fewer AS hops.
- If needed this coverage can be augmented with selective relationships with a small number of ASes.

While the economic implications of CDN architectures are beyond the scope of this paper, the analysis hint at the tradeoffs that exists between deploying and managing more caches versus maintaining appropriate peering relationships.

5 Traffic Dynamics

In this section we consider the traffic dynamics across various time-scales. Figures 6(a)-(b) plots the bandwidth usage across a one-month period for both On-Demand and Live. The data indicates substantial variability in the bandwidth demand. For On-Demand, the mean, median and peak bandwidths are 4.6 Mbps and 1.1 Mbps, and 141 Mbps, respectively. The peak is 31 times the mean. For Live, the mean, median and peak bandwidths are 13.4 Mbps, 10.3 Mbps and 122 Mbps respectively. The graphs also show there are daily local peaks, and that there can be substantial differences in the peak bandwidth requirement across days.

Figures 7(a)-(b) focus on the bandwidth requirements for each of 2 days (Dec12 and Dec 13, 2001), for On-Demand. The bandwidths here are averaged

Result Set	Traffic Volume (% of total)	# IP addresses (% of total)	# ASes (% of total)
One AS hop (or less)	52.5	53.5	17.5
Two AS hops (or less)	88.7	89.7	72.7
One AS hop & selected ASes	73.4	72.5	20.1
Two AS hop & selected ASes	91.7	92.1	73

Table 4: *On Demand*: Content Distribution Statistics.

Result Set	Traffic Volume (% of total)	# IP addresses (% of total)	# ASes (% of total)
One AS hop (or less)	60.1	64.4	18.9
Two AS hops (or less)	63.9	68.7	71.2
One AS hop & selected ASes	79.7	80.1	22.2
Two AS hop & selected ASes	94.6	95.4	77.5

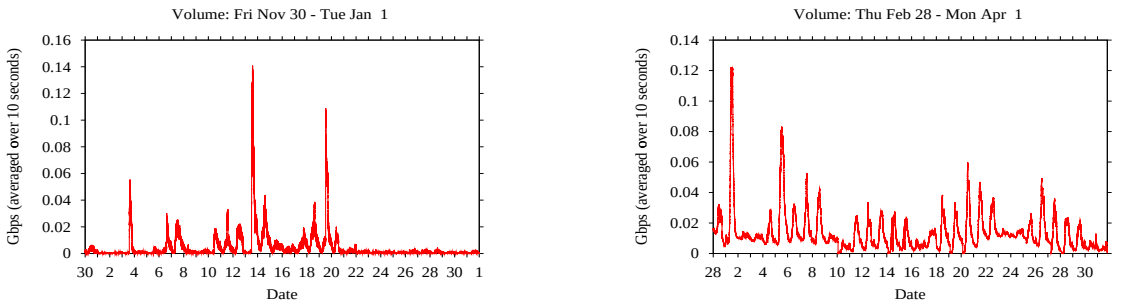
Table 5: *Live*: Content Distribution Statistics.

over 1 sec. intervals. The graphs reveal the following time-of-day effect. In both cases, the bandwidth curve shows a low demand early in the day. This is followed by an increase in demand (steeply for Dec 13, more gradually for Dec 12), followed by a region of high bandwidth requirement. Finally, the demand drops off. The mean, median and peak bandwidths for Dec 12 and 13 respectively are (9 Mbps, 8.76 Mbps, 26 Mbps) and (28 Mbps, 15 Mbps, 153 Mbps), indicating that there can be a significant variation in bandwidth load across the entire day. Note that Dec 13 has a much higher peak than Dec 12 (almost six times higher) and is among the three high peak days in Fig. 6(a). On each of these three days, the high load was traced to heavy demand for a small number of clips. Fig. 7(c) shows that the increase in bandwidth usage can be quite sudden. For Dec 13, the load increases from 1.35 Mbps (by a factor of 57) to 77 Mbps within a span of just 10 minutes. The above data suggests that we have a "flash-crowd" effect for Dec 13. We also find that the bandwidth variations across time (Fig. 7(b)) are due to variations in the number of requests across the day. This can be seen in Fig. 7(d), where the graph showing the number of concurrent connections, closely resembles the bandwidth usage across the day. Fig. 8 indicated that the high level daily trends are similar for the live streaming data. Figure 8(b) shows the initial 5 hours of the ramp up. This is clearly happening much more gradually than for the *On Demand* data for Dec 13. A more gradual buildup in demand, by providing more reaction time, should make it easier to handle the increased load, than the sudden surge witnessed for the *On Demand* data.

The above graphs show that there can be significant bandwidth variability with substantial difference between peak bandwidth requirement within a day and across days. In addition, the demand can spike by several factors within a few minutes. All this makes it a challenging problem to provision server and network resources to handle such a variable workload in a resource-efficient manner. Provisioning for the peak demand would keep the system and network resources under-utilized most of the time, and may be uneconomical for individual content providers. If the content were hosted at a single location, the sudden large traffic surges we see might create hot-spots and performance problems for the network provider and the end-users of the content. Instead, a distribution infrastructure (such as a CDN) shared among many different content providers might be useful as it offers the potential for statistical multiplexing of resources. This would allow more efficient and economical resource usage, with different providers getting access to additional burstable bandwidth when required. Appropriate distribution mechanisms can be used to distribute the request load across the CDN to prevent hot-spots.

5.1 Object popularities

Understanding how the observed traffic relates to the different clips will be useful for developing traffic models, and for determining appropriate techniques for handling the workload for such streaming content. Figs. 9(a)-(c) show the per-object traffic contribution for three days in Dec, 2001 and March 2002 for *On Demand*. Figs. 10(a)-(c) shows the distribution of the number of sessions per clip



(a) On-demand (b) Live

Figure 6: Bandwidth demand across time for On-demand (Dec 2001) and Live (March 2002). Each point represents the average bandwidth across a 10-sec interval

for the same three days. Both set of graphs indicate that a small number of "heavy hitter" clips account for the bulk of the traffic (volume as well as number of sessions). For instance, for Dec 13, 2001, the top 5 clips (out of a total of 320 clips requested that day) together accounted for 85% of the total traffic. This suggests that distribution and bandwidth management techniques focused on realizing resource-efficient delivery for the few "heavy hitter" clips, might be required. For instance, in a CDN infrastructure, the few heavy hitters could be replicated and served from a large number of the CDN nodes. Also promising are scalable delivery techniques such as patching and periodic broadcast [6, 20, 3] which can deliver a popular clip to a large number of clients, with significant bandwidth savings.

For certain types of content, clients may be willing to view a clip later in time. In such cases, providing a "delayed download capability" as an option may be an attractive alternative for the content provider, network and end-users in times of high load. (Such an approach clearly offers a different user experience than the instantaneous viewing enabled by streaming. However, the use of such techniques in popular peer-to-peer systems, indicate that it might be acceptable for certain types of content.) The server could schedule the downloads to occur automatically to the clients during off-peak time. This would help reduce/smooth out the traffic peaks (Fig. 6(a),Fig. 7) while still satisfying many requests using essentially time-shifted playback.

Finally, for the live streaming, Figure 8 indicates that there can be a significant number of concurrent connections for the event. Using multicast delivery seems a promising way to reduce the bandwidth usage in this context.

6 Session Characteristics

We next study the distribution of session durations and data download sizes for streaming content. Table 6 depicts the sizes, durations and bandwidths of four popular on-demand clips. These clips all correspond to the same content and differ in the format (Real or Windows) and bandwidth (low or high).

Fig. 11 shows the CDF of the amount of data downloaded by the sessions requesting each clip. The graphs suggest that the data download session can be highly variable across different sessions requesting the same clip. For all the clips, a large fraction of sessions download only a small part of the video. For instance for clip 1 (clip 2 is similar), 62% of the sessions download at most 10% of the video, and only 10% download more than 90% of the clip. This behavior may be an indication of users either viewing a prefix of the video or of using forward index jumps to browse the clip. The behavior may also be an indication that the user-percieved reception quality may be inadequate in many cases. We note that for both low bandwidth clips, sessions tend to download a smaller proportion of the object than for the high bandwidth clips. For instance for clip 3, 82% of the sessions download at most 10% of the video, and less than 2% download more than 90% of the clip. This difference could be due to a combination of (i) the poorer viewing quality of the low-bandwidth encodings, and (ii) poorer connection quality experienced by users with low bandwidth network connectivity (e.g. dial-up users) - they are the most likely audience to request a low bandwidth encoding in preference to a higher bandwidth version.

For all clips, the CDF shows a spike (more pronounced for clips 1 and 2) around the region where the data download is 100% of the video size. This is

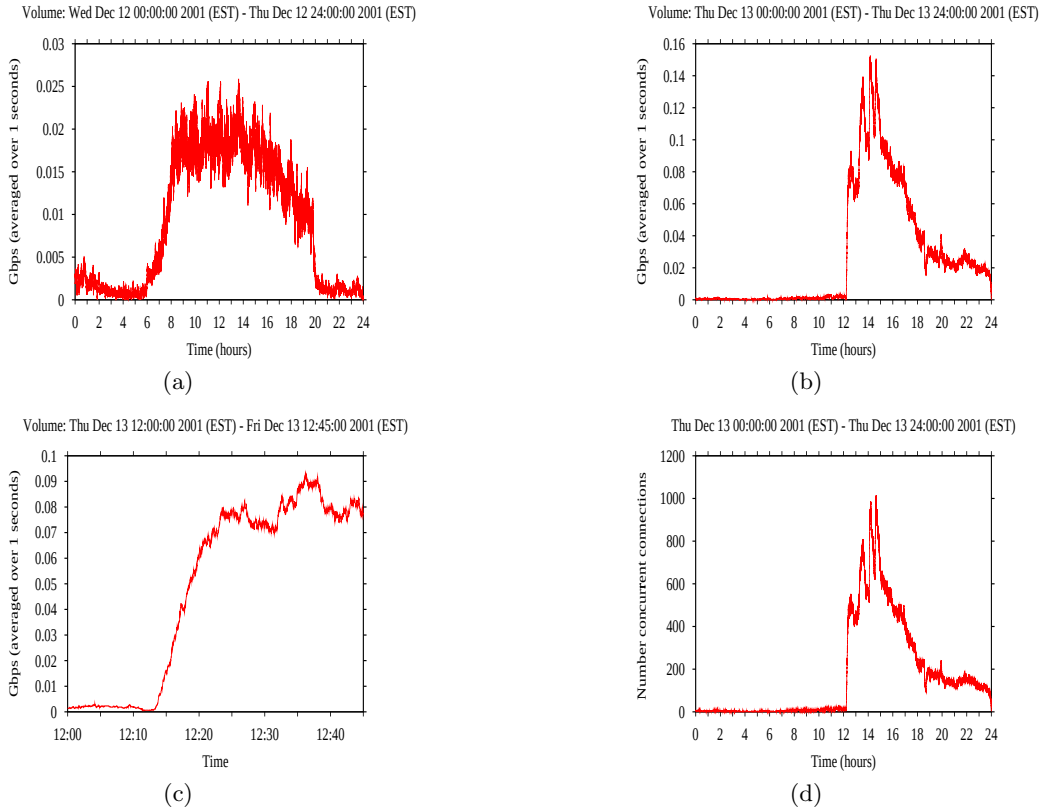


Figure 7: On Demand - (a)-(b) plot the bandwidth across time for Dec 12 and Dec 13, 2001. For the latter day, (c) plot the bandwidth for a 50 min. time interval, and (d) number of concurrent connections for the entire day.

Name	Format	Bandwidth	Duration (sec)
clip1	MMS	High	268
clip2	Real	High	272
clip3	MMS	Low	271
clip4	Real	Low	272

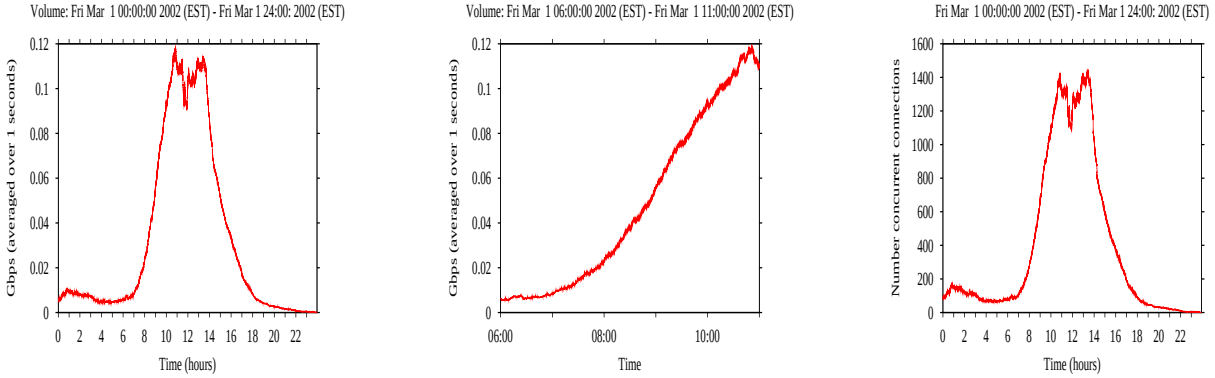
Table 6: *Popular clips*: Properties.

due to a mass of sessions downloading the complete clip.

For all the clips we note that there are some sessions that each download data in excess of the video size. The effect is more pronounced for the high bandwidth clips. For instance, 0.03% of sessions download more than twice the video size for clip 1. We are currently investigating the reason for this behavior.

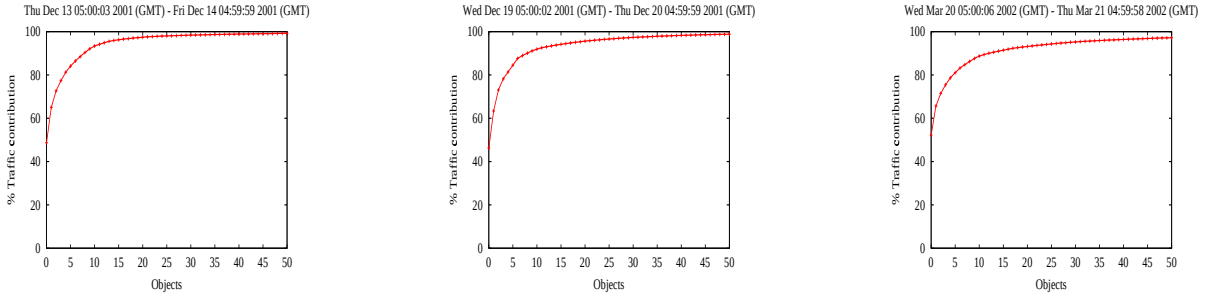
Fig. 12 shows the CDF of the connection times for the sessions requesting each clip. The graphs indicate that session length can be highly variable across different sessions requesting the same clip. A large fraction of sessions last for only a short time period, and small fraction tends to be long-lasting. We note that a smaller fraction of low-bandwidth sessions are long-lasting compared to high bandwidth ones. For instance for clip 1, 36% of the sessions last at most 10 sec, and 16% last more than 200 sec. In comparison, for instance for clip 3, 38% of the sessions last at most at most 10 sec, and 5% last more than 200 sec. The spike in the graphs occurs at around 270 sec, the video length.

Fig. 13(a)-(b) depict the distribution of session connection times for *Live*. A large proportion of the sessions (69%) are on for 2 minutes or less. However the distribution exhibits a long tail (as seen



(a) (b) (c)

Figure 8: Live - (a)-(b) plot the bandwidth across time for the entire day and for a 5 hour period, for March 1, 2002. (c) plots the number of concurrent connections for the entire day.



(a) Dec 13, 2001 (b) Dec 19, 2001 (c) Mar 20, 2002

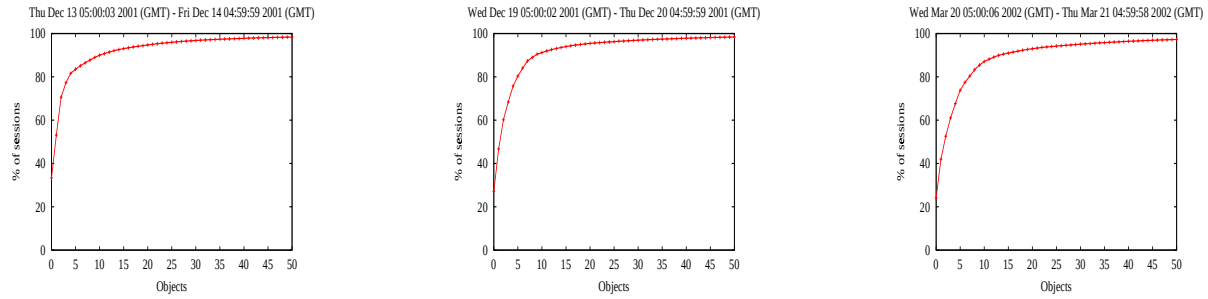
Figure 9: *On Demand* Traffic volume distribution. X-axis: clips ranked in decreasing order of traffic contribution (top 50 clips shown). Y-axis: cumulative traffic contribution (percentage of total).

from Fig. 13(b)). About 12% of the sessions are at least 10 minutes long, while 8% of the sessions are longer than 20 minutes. This suggests that there exists an audience for long-duration network-based streaming video presentations. For on-demand content, this in turn argues for expanding the content offering from the short-duration clips that are the norm today to more comprehensive presentations.

7 Summary of results

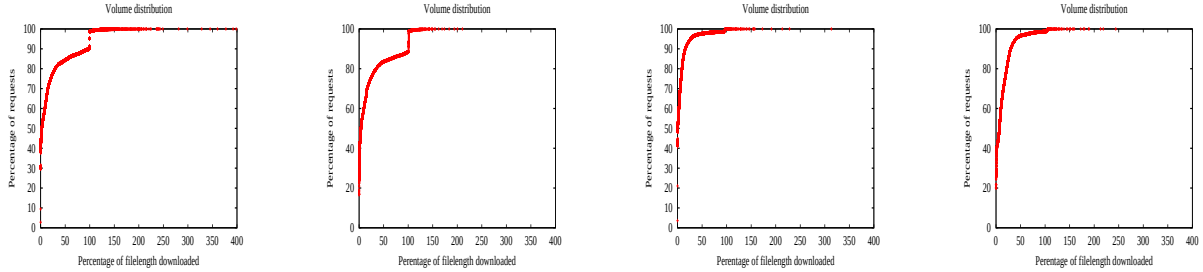
In this Section we list a summary of our findings:

- Requests for Windows Media dominate those for Real where content is available in both formats.
- Requests for content encoded at a higher bitrate dominate where high and low encoding rates are available.
- Sessions using transport protocols running over TCP dominate those using UDP.
- Request and traffic volumes are highly skewed at different levels of aggregation (IP address, routing prefix and AS).
- For a tier-1 ISP a significant percentage of streaming clients are within 2 AS hops of the ISP.
- Selective arrangements with a modest number of consistently high contributing ASes yield significant gain in improving coverage to streaming clients.
- Streaming traffic exhibits regular daily patterns with very high variability in terms of request, traffic volume and concurrent number of connections.
- Ramp up to daily peaks can be gradual over several hours or very sudden over tens of minutes.
- Streaming traffic exhibit very high variability in terms of daily peaks.



(a) Dec 13, 2001 (b) Dec 19, 2001 (c) Mar 20, 2002

Figure 10: *On Demand* Request distribution. X-axis: clips ranked in decreasing order of number of sessions (top 50 clips shown). Y-axis: cumulative number of sessions (percentage of total).



(a) (b) (c) (d)

Figure 11: Figs.(a)-(d) plot the cumulative distribution of data download associated with each request for clips 1-4 respectively

- A small number of streaming objects is responsible for significant portions of the request and traffic volume.
- Where the same content is encoded in high and low bitrates, the higher bitrate clients tend to watch more of the content.

8 Conclusions and future work

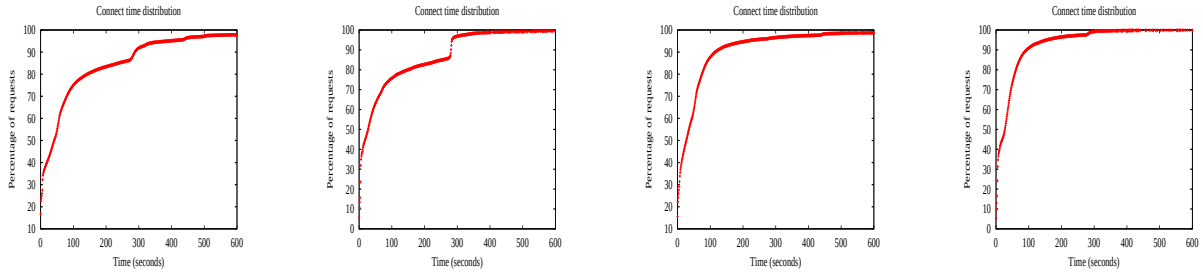
This study revealed various important aspects of streaming traffic on the Internet. For one it showed the widespread use of streaming with content being accessed by many endpoints across many different networks. However a lot of work remains to be done to fully characterize streaming traffic and applying such knowledge to deliver streaming content in the most efficient way to large numbers of clients.

The first obvious future direction for our work is to determine how the various session compositions we investigated will develop over longer periods of time and whether it holds over other larger data sets.

In this paper, we have taken a first pass over the data towards developing a workload model for streaming traffic. However, coming up with a parameterized model for streaming traffic will require a more detailed look at the relationships between the request arrival process, the popularity distribution of streaming objects, object sizes and play times etc.

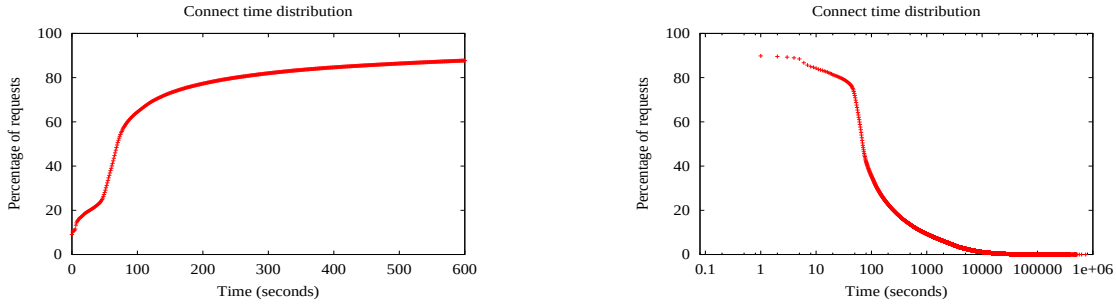
Similarly, on the network side, the relative stability of distributions across longer time scales will be important in order to engineer scalable content distributions strategies. In particular, we need to investigate the general applicability of the suggested approach of selective relationships with high contributing networks.

Finally, some of the busy days we encountered in our data set exhibited “flash crowd” behavior. Coupled with the relatively high per-client bandwidth requirements of streaming media, this can have a substantial impact on the various resources associated with a streaming service. Studying these events in detail will be instrumental in developing techniques for dealing with or reducing the impact of this phenomena.



(a) (b) (c) (d)

Figure 12: Figs.(a)-(d) plot the cumulative distribution of connect times associated with each request for clips 1-4 respectively



(a) (b)

Figure 13: *Live*: (a) CDF of connect times (b) CCDF of connect time (X axis logscale)

9 Acknowledgments

We thank Tim Griffin who's ongoing efforts provided the routing data as well as Matt Roughan for discussions on appropriate statistical methods. We also thank Jennifer Rexford for many helpful comments on an earlier version of the paper, and the anonymous reviewers, whose suggestions benefited the final version of the paper. Finally we thank the anonymous sources of our streaming logs who made this analysis possible.

References

[1] Jussara Almeida, Jeffrey Krueger, Derek Eager, and Mary Vernon. Analysis of educational media server workloads. In *Proc. Inter. Workshop on Network and Operating System Support for Digital Audio and Video*, June 2001.

[2] J. M. Boyce and R. D. Gaglianello. Loss effects on MPEG video sent over the public Internet. In *Proc. ACM Multimedia*, September 1998.

[3] Michael K. Bradshaw, Bing Wang, Subhabrata Sen, Lixin Gao, Jim Kurose, Prashant Shenoy, and Don Towsley. Periodic broadcast and patching services - implementation, measurement, and analysis in an internet streaming video testbed. In *Proc. ACM Multimedia*, October 2001.

[4] Maureen Chesire, Alec Wolman, Geoffrey M. Voelker, and Henry M. Levy. Measurement and analysis of a streaming media workload. In *USENIX Symposium on Internet Technologies and Systems*, March 2001.

[5] Barclay Dutson, Claudia Dutson, and Stephen Drayson. New opportunities in streaming media report. In *Vision Consultancy Group*, September 2000.

[6] Derek Eager, Mary Vernon, and John Zahorjan. Minimizing bandwidth requirements for on-demand data delivery. In *Proc. 5^{th} Inter. Workshop on Multimedia Information Systems*, October 1999.

[7] M.W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proc. ACM SIGCOMM*, September 1994.

[8] Bruce Kasrel, Josh Bernoff, and Meredith Gerson. Broadband content splits. In *Forrester Research*, October 2000.

[9] Balachander Krishnamurthy and Jia Wang. On Network-Aware Clustering of Web Clients. In *Proceedings of ACM Sigcomm*, August 2000.

[10] Marwan Krunz and Satish K. Tripathi. On the characteristics of VBR MPEG streams. In *Proc. ACM SIGMETRICS*, pages 192–202, June 1997.

[11] T. V. Lakshman, A. Ortega, and A. R. Reibman. Variable bit-rate (VBR) video: Tradeoffs and potentials. *Proceedings of the IEEE*, 86(5), May 1998.

[12] Dmitri Loguinov and Hayder Radha. Measurement study of low-bitrate Internet video streaming. In *Proc. ACM SIGCOMM Internet Measurement Workshop*, November 2001.

[13] N. F. Maxemchuk and S. Lo. Measurement and interpretation of voice traffic on the Internet. In *Proc. International Conference on Communications*, June 1997.
`http://www.research.att.com/~nfm/ref.1443.ps`.

[14] Art Mena and John Heidemann. An empirical study of real audio traffic. In *Proc. IEEE INFOCOM*, March 2000.

[15] J. Padhye and J. Kurose. An empirical study of client interactions with a continuous-media courseware server. In *Proc. Inter. Workshop on Network and Operating System Support for Digital Audio and Video*, 1998.

[16] Amy R. Reibman and Arthur W. Berger. Traffic descriptors for VBR video teleconferencing over ATM networks. *IEEE/ACM Trans. Networking*, 3(3):329–339, June 1995.

[17] J. Rexford and D. Towsley. Smoothing variable-bit-rate video in an internetwork. *IEEE/ACM Trans. Networking*, 7(2):202–215, April 1999.

[18] James D. Salehi, Zhi-Li Zhang, James F. Kurose, and Don Towsley. Supporting stored video: Reducing rate variability and end-to-end resource requirements through optimal smoothing. *IEEE/ACM Trans. Networking*, 6(4):397–410, August 1998.

[19] Gregory J. Scaffidi and Mark Zohar. Consumer broadband hits hypergrowth in 2001. In *Forrester Research*, October 2000.

[20] Subhabrata Sen, Lixin Gao, Jennifer Rexford, and Don Towsley. Optimal patching schemes for efficient multimedia streaming. In *Proc. Inter. Workshop on Network and Operating System Support for Digital Audio and Video*, 1999.

[21] Subhabrata Sen, Jennifer Rexford, Jayanta Dey, James Kurose, and Don Towsley. Online Smoothing of Variable-Bit-Rate Streaming Video. *IEEE Transactions on Multimedia*, pages 37–48, March 2000.

[22] Yubin Wang, Mark Claypool, and Zheng Zuo. An empirical study of Realvideo performance across the Internet. In *Proc. ACM SIGCOMM Internet Measurement Workshop*, November 2001.

[23] Maya Yajnik, Jim Kurose, and Don Towsely. Packet loss correlation in the MBone multicast network. In *IEEE Global Internet*, November 1996.