Understanding User Behavior in Large-Scale Video-on-Demand Systems*

Hongliang Yu†, Dongdong Zheng†, Ben Y. Zhao§ and Weimin Zheng†
† Computer Science Department, Tsinghua University, Beijing China
§ Computer Science Department, U. C. Santa Barbara, CA, USA
†{hlyu@,zdd03@mails.,zwm-dcs@}tsinghua.edu.cn, §ravenben@cs.ucsb.edu

ABSTRACT

Video-on-demand over IP (VOD) is one of the best-known examples of "next-generation" Internet applications cited as a goal by networking and multimedia researchers. Without empirical data, researchers have generally relied on simulated models to drive their design and developmental efforts. In this paper, we present one of the first measurement studies of a large VOD system, using data covering 219 days and more than 150,000 users in a VOD system deployed by China Telecom. Our study focuses on user behavior, content access patterns, and their implications on the design of multimedia streaming systems. Our results also show that when used to model the user-arrival rate, the traditional Poisson model is conservative and overestimates the probability of large arrival groups. We introduce a modified Poisson distribution that more accurately models our observations. We also observe a surprising result, that video session lengths has a weak inverse correlation with the video's popularity. Finally, we gain better understanding of the sources of video popularity through analysis of a number of internal and external factors.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed Applications

General Terms

Measurement, Performance, Human Factors

Keywords

Video-on-demand, User behavior, modeling, Poisson distribution

1. INTRODUCTION

Streaming Video-on-Demand (VOD) over the Internet is the next major step in the evolution of media content delivery. For several years, cable and satellite providers (Comcast, Dish Networks), video rental companies (Netflix) and other media companies (TiVo) have been developing online streaming video systems for an increasingly demanding and growing consumer population [24]. By leveraging the increasing availability of broadband access, VOD systems offer users the ability to browse, select, view, and scan through media content from a large content repository on an on-demand basis, all from the comfort of their homes. With recent studies [16, 4, 19] that show a significant shift in Internet traffic from the web to multimedia content, it is clear that both the necessary capacity and demand for streaming multimedia have arrived.

While current VOD systems remain in the prototype or design stages, a key challenge companies face is how to design an architecture that scales smoothly to a large number of customers, all while maintaining low access latency, high video quality and reasonable operational costs. Designers use simulations based on common assumptions to evaluate and drive their architectures. Unfortunately, the lack of deployed VOD systems meant few of these assumptions have been validated on real measurement data.

In this paper, we present one of the first measurement studies of a large deployed VOD system. Our data set comes from detailed logs of a video-on-demand system deployed by China Telecom, covering a total of 1.5 million unique users for a period of seven months in 2004. Our analysis seeks to validate and adapt existing assumptions about streaming media, focusing on user behavior and content access patterns. More specifically, we examine the accuracy of existing models for user arrival rates and request patterns. By comparing and contrasting our analysis with existing models, we rectified some of the common misunderstandings about VOD users and their access patterns. In addition, we derived more accurate models of user behavior and access patterns based on our analysis, which will not only increase the accuracy of existing simulations, but also serve as building blocks in more sophisticated experiments.

Initial analysis revealed several key facts. First, user distributions follow clear patterns with respect to time and arrival rates match a modified form of the Poisson distribution. Second, the average session is quite short, due to users sampling movies by "scanning" through them, much like the "intro-sampling" mode supported by portable CD-players. In addition, session lengths are influenced by file popularity. But surprisingly, the correlation is an inverse one, where less popular videos actually see longer session times. Third, our system does not follow the "fetch-at-most-once" model, and our file popularity matches the Zipf's distribution much better than prior work suggested. Finally, we find the change of file popularity over time is greatly influenced by external factors such as highly visible "recommended videos" and "most popular videos" lists, suggesting that system designers can use them as predictable guidance metrics for near-future user access requests.

*This work is supported by the National Natural Science Foundation of China under Grant No. 60433040. This work is also partially supported by DARPA BAA04-11 and by the NSF under CAREER Award CNS-0546216.

EuroSys'06, April 18–21, 2006, Leuven, Belgium.

Table 1: Components in the Powerinfo VOD system

Role	# of Servers	Functionality
Application server	1	Task scheduling, load balancing
Management server	1	System monitoring, management, accounting
Media server	$\geqslant 1$	Media streaming

The rest of this paper is organized as follows. We begin by first describing the China Telecom video-on-demand system and our source data in Section 2. Then in Section 3, we present our detailed analysis on user behavior characteristics and content access patterns. Next, we compare our approach study to related work in Section 6. Finally, we summarize our results and conclude in Section 7.

2. THE POWERINFO VOD SYSTEM

Measurement data used in our study come from logs collected during daily operation of the PowerInfo Video-on-demand system. Currently deployed VOD systems are mainly operated as a free value-added services by telecommunication companies in China. Paying customers can access the large video libraries free of charge on an unlimited basis. Many of these are Internet Service Providers with large bandwidth resources, whose primary goal is to attract new users. PowerInfo is one of the leading video streaming software providers in China. Its system provides service to over 20 cities in China, most of which are generally managed by regional branches of China Telecom. To date, the PowerInfo system user base exceeds 1.5 million, the large majority of whom are connected using broadband (512 kb/s) to the home.

The PowerInfo Video-on-Demanding system uses a distributed architecture for media streaming. Customers are divided into regional networks, each served by one or more server clusters known as Units. Nodes in each VOD Unit cluster serve one of three different roles: *application servers*, *management servers* and *media servers*. In each unit, one application server caches video metadata, performs authentication, and interacts directly with users to schedule streaming requests. Tasks route to one or more media servers, which stream video directly to the user. Application servers load balance tasks appropriately across media servers. Finally, a management server monitors all servers in the VOD Unit, and perform accounting functions based on user requests. Statistics gathered on user requests are used to determine the optimal number and placement of replicates for each individual video file. We list these components and their responsibilities in Table 1.

For the analysis presented in this paper, we used a complete segment of the PowerInfo system log ranging from May 16th to December 20th of 2004. The system log includes both an extensive record of user accesses and a full metadata listing of available video files. For each streaming session, the user log includes the user's IP address, ID number of video requested, timestamps for the start and end of the session, and the media and application servers used. A sample of the user log is shown in Table 2.

We focus our analysis to logs from a single representative city with a total user base of 150,000 users served by 3 VOD Units. We list the hardware configuration of this regional system in Table 3. Note that like many other regions, all servers in these Units are connected to the main bone of China Telecom through a gigabyte network.

We summarize the logged statistics for our representative city in Table 4. During the logged period of 219 days, users in our representative city issued more than 21 million video requests covering a total of over 6700 unique video files. The total length of videos streamed is more than 317,000 minutes, or roughly 5300 hours. Figure 2 shows the daily user distribution during our tracking period. In addition, the first week of May and October are both Chinese national holidays, and we were able to isolate user behavior changes during these two vacation weeks.

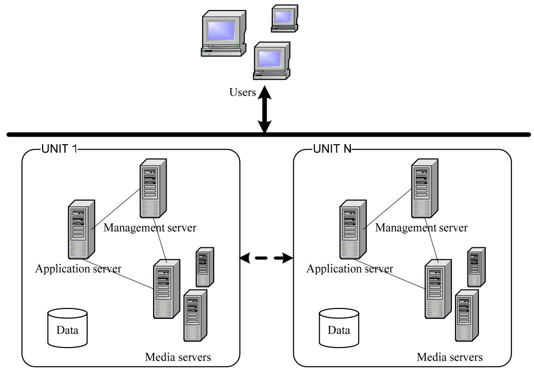

Figure 1: Architecture of the PowerInfo VOD system

Table 2: Log Samples

Prog #	Starttime (s)	Endtime (s)	Traffic	MS	AS
2884	1097397599	1097400153	764192	22	1
16742	1097397600	1097397619	3980	21	1
2021	1097397600	1097400053	357888	22	1
...	...	...	...	...	...

Finally, we briefly describe the types of videos available in the PowerInfo system. The large majority of videos in the library are recordings of older television shows and Chinese movies, encoded via MPEG1, MPEG2 and MPEG4 codecs at relatively low resolution. A typical file is an older movie, roughly 100 minutes in length, and stored as a 300 MByte MPEG1 file. We note that while PowerInfo is a commercially deployed VOD system, videos can be accessed by members free of charge on an unlimited basis. Note that this value-added service model is also being adopted by Comcast Cable as they begin their video-on-demand service deployment in the US [11].

3. USER ACCESS PATTERNS

In this section, we discuss the basic characteristics of user behavior in our large scale VOD system. Understanding how users access the system can help system designers optimize resources in order to produce the best user experience at minimal cost. We begin our analysis with user access patterns over time, then discuss the modeling of the user arrival distribution, and conclude with a study of user session length distributions.

3.1 User Accesses over time

We begin by trying to gain a basic understanding of user access patterns in the system. The distribution of user accesses can be characterized as a function of time. We examine its distribution first across hours of the day, and then across days of the week.

3.1.1 Daily Access Patterns

Looking at how user accesses change during the course of a day, we found that as expected, the user accesses follow a clear daily

Table 3: Hardware configurations

UNIT	Components	Num	Hardware configurations
1	Application server	1	HP DL360 G2 (2*PIII XEON 1.4G, 1G RAM)
	Management server	1	HP DL580 (2*PIII XEON 700, 1G RAM)
	Media server	4	HP DL580 (8*PIII XEON 1.4G, 2G RAM)14*73G SCSI HD
2	Application server	1	Co-located with one of the media servers
	Management server	1	Co-located with one of the media servers
	Media server	3	HP DL5807(8*PIII XEON 1.4G, 2G RAM) 10*73G SCSI HD
3	Application server	1	Co-located with one of the media servers
	Management server	1	Co-located with one of the media servers
	Media server	2	Compaq DL380 G3(4*XEON 3G, 1G RAM)10*36G SCSI HD

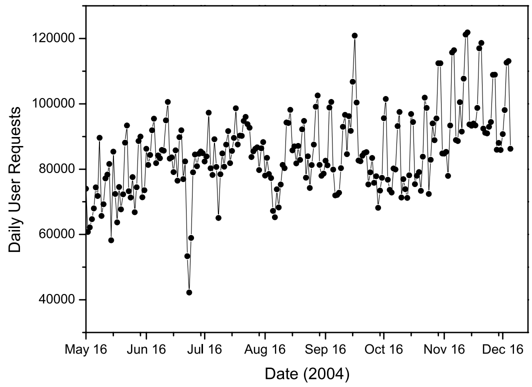

Figure 2: Number of daily user accesses across the entire log period of 219 days.

Table 4: Statistics summary of logs

Sessions	Length	Files Avail.	Files accessed
21,498,338	219 days	7036	6716

pattern. While our data shows a consistent pattern across most days, we focus specifically on the days with the highest traffic volume. Therefore, we focused on seven days during the first week of October 2004, when the PowerInfo system experienced its highest number of user accesses due to the arrival of the week-long national holiday.

Figure 3 shows a time-series plot of the total number of users. As expected, within one single day the number of users drops gradually during the early morning (12AM-7AM) and the afternoon (2PM-5PM), while it climbs up to a peak when users are in noon break (Noon-2PM) or after work (6PM-9PM). This pattern reflects the expected behavior as users entertain themselves during breaks and after work hours. Naturally, the second daily peak in the number of users usually arrives after the "prime time" (7PM-10PM) of the commercial television industry.

3.1.2 Weekly Access Patterns

To get a broader view of the user distribution over time, we chose data for a period of seven consecutive weeks from our records. We plot the daily access count in Figure 4, with the division into weeks shown as vertical lines marking each Sunday night at midnight. As we can see, while the daily number of user requests can fluctuate during the first three workdays of the week, the average number of daily requests increases steadily in the second half of the week, reaching its peak on Sunday. As expected, this shows a direct correlation with users' work habits during the week, and more relaxed time at home during the weekends.

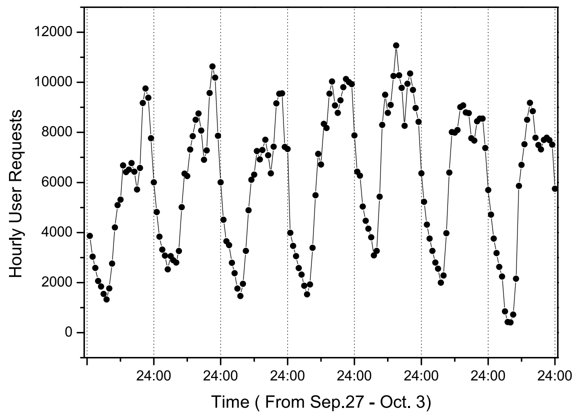

Figure 3: Hourly distribution of user accesses during the course of a single week. (China's week-long national holiday begins annually on October 1st.)

We observe an interesting phenomenon just days before the start of the national holiday on October 1, 2004. There is a substantial increase in user requests starting on September 30th, the day before vacation starts. Clearly, worker productivity is impacted by employees watching more videos at work in anticipation of the arriving holiday week. While requests hit an all-time daily high on the first day of vacation (10/1/2004), they quickly drop to below-normal levels for the next two weeks. This is consistent with the traditional annual boom in domestic travel that occurs every year during and following the national holiday. While the average subscriber is vacationing on the road, active requests on PowerInfo drop significantly.

3.2 User Arrival Rates

User arrival distributions in multimedia systems are often modeled using a Poisson distribution. Here, we study the validity of that model applied to our usage data. First, we look at the general user arrival rate. We measure user arrival by counting the number of arrivals in 5 second buckets. Figure 5 shows the results of this measurement across the entire log period. From this simple result, we can draw two conclusions. First, the number of arrivals usually ranges from 0 to 27 users per 5 seconds, or 0 to 5 users per second. Second, this arrival rate does not match a Poisson distribution.

One of the main challenges in the design of a VOD system is how to handle a large number of simultaneous users. Therefore, we take a closer look at user arrival patterns under periods of heavy load. We isolate the user arrival data from 6PM-9PM for each day

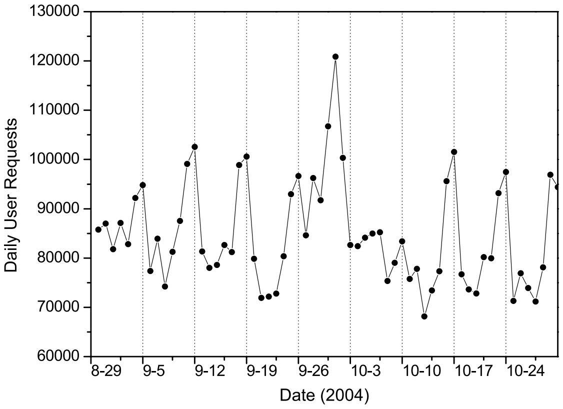

Figure 4: A view of daily user requests covering a period of seven weeks, including the week-long national holiday from 10/1 to 10/7.

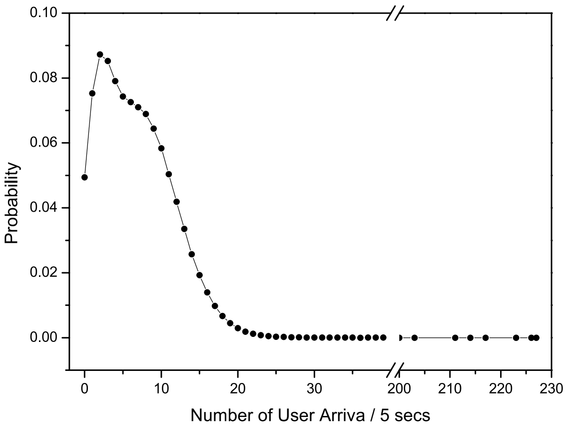

Figure 5: User arrival distribution showing the number of arrivals per 5 second intervals across the entire log period.

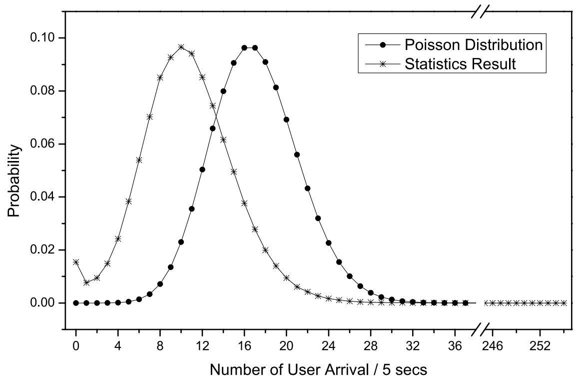

Figure 6: Statistics of user arrival versus Poisson distribution (with lambda=15)

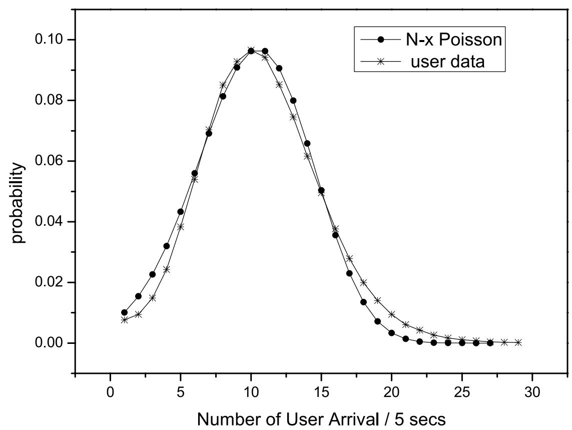

Figure 7: Statistics of user arrival versus modified Poisson distribution

of our log period. As shown in Figure 3, this is the daily period with the heaviest user traffic. When we try to match this arrival distribution with one derived from the Poisson distribution, we get Figure 6. It seems that while the heavy-load user arrival distribution looks similar to the Poisson distribution, the two do not match well. From the analysis, we see that the Poisson distribution underestimates the possibility of small arrival cases and it over-estimates the probability of large arrivals. So any VOD system or simulation model based on a Poisson distribution will likely result in an over-provisioned system.

To provide a more accurate model, we introduce a modified version of the Poisson distribution by replacing the independence variant x with $(N-x)$, where N is the maximum number of user arrivals in our records. As Figure 7 shows, our tracking results match the modified Poisson distribution very well. This modified version of Poisson distribution can be defined as:

$$P(X) = \frac{\lambda^{(N-X)}e^{-\lambda}}{(N-X)!}, X = 0, 1, 2... \quad (1)$$

Here, N is the maximum number of user arrivals in a target system. In Figure 7, we used the values $\lambda = 17$ and $N = 27$. Clearly, our modified Poisson distribution can be used as a reference model to study the user behavior of large scale streaming services in a heavy workload.

3.3 Session Lengths

We now turn our attention to an analysis of the lengths of streaming sessions measured on the PowerInfo VOD system. In particular, we wish to better understand why and how users terminate a streaming session. We first present general results on session lengths across all videos. Next, we examine session length statistics for two representative video streams to explain user behavior artifacts. Finally, we explore the relationship between session length and video popularity.

We first note that the videos in the VOD library span a wide range of lengths, including everything from television shows that range between 30 to 60 minutes to a large number of movies that range from 90 to more than 120 minutes. To adjust for this large variance, we use as our metric the *normalized session length* (NSL), the proportion of the video viewed before the session terminated ($SessionLength/VideoLength$). In addition, because we are using session length data to help us better understand user behavior, we focus on sessions proactively terminated by the user, and remove from our dataset sessions that ran to completion. Specifically, we examine for Table 5 and Figure 8 sessions that lasted less than 85% of their video lengths. This accounts for roughly 86.33% of the entire dataset.

Table 5: Statistics of session length

Session length	5 min	10 min	25 min	50 min
percentage	37.44%	52.55%	75.25%	94.23%

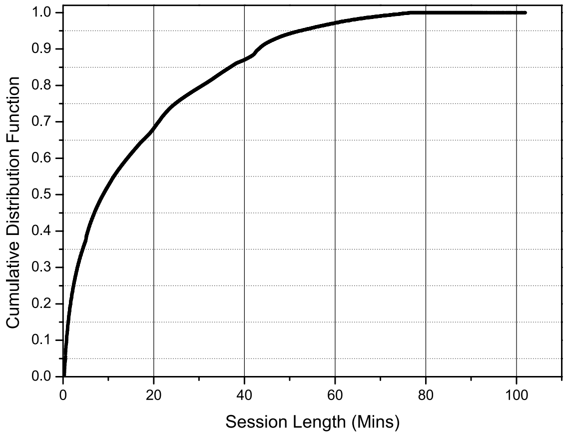

Figure 8: Distribution of session length

We summarize some session length statistics in Table 5, and plot the cumulative distribution function (CDF) of session lengths in Figure 8. As shown in both, the majority of partial sessions (52.55%) are terminated by the user within the first 10 minutes. In fact, 37% of these sessions do not last past the first five minutes. Within 25 minutes, more than three-quarters of all sessions have been terminated. Note that since this dataset covers nearly 90% of the entire dataset, we expect similar results to hold across the full dataset.

These results show an extremely "impatient" audience, who despite the availability of program guides and movie information, often scan through the beginning of videos to quickly determine their interest. This evidence suggests that prefix cache systems such as [20] can significantly improve user response time by caching the beginning of a large portion of videos in easily accessible memory or disk. Our results predict that caching the first 10 minutes of videos will be sufficient to serve 50% of all user sessions. While we can attribute some of this user impatience to the fact that users do not pay per video, this evidence also suggests that even pay-per-video VOD systems can significantly improve user satisfaction if they can offer users the ability to scan through the beginning of movies for free. This approach is currently offered by some Pay-Per-View movie services such as DirecTV.

To shed additional light on this phenomenon, we next examine in detail session lengths for two popular videos. We collect the lengths of all sessions that streamed these two videos, and show the session length distribution in Figure 9. Note that for clarity between the lines, we used a "non-normalized" CDF. We note the presence of two large spikes in sessions terminated within the first 10 minutes. A large number of sessions end within the first minute. We can only speculate that these users quickly determined they had ordered the wrong video after seeing the initial title screen. This suggests that VOD systems need to do a good job of providing information about movies in order to reduce the frequency of these types of "fruitless" sessions. Another major spike occurs around the five minute mark, when a large number of sessions end. We can presume that these users are "scanning" through videos, and having seen enough of

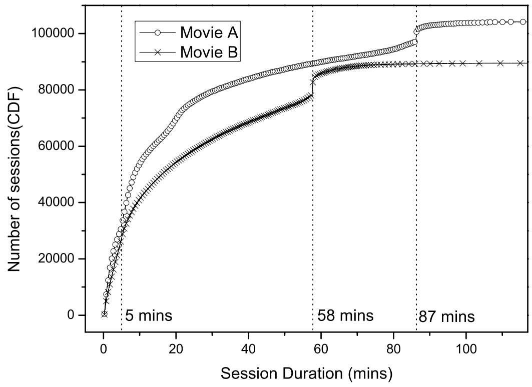

Figure 9: The session length distribution for two popular movies. Movie A has a total length of 87 minutes, while movie B has a total length of 58 minutes.

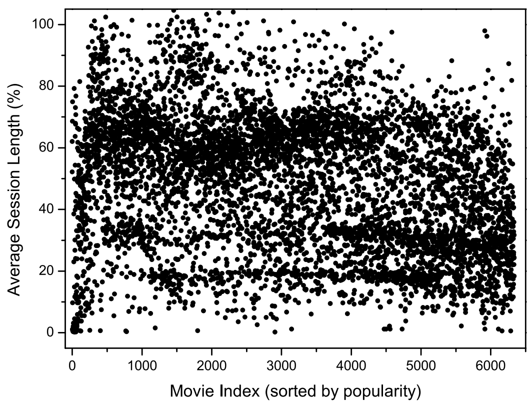

Figure 10: A comparison of session lengths to video popularity. Average session lengths for videos are sorted by the video's popularity from least popular to most popular.

the movie, decided to move on to the next video. Finally, the rest of the user sessions are spread out in length, except for two clusters of user sessions that match the length of each video (58 minutes for B and 87 minutes for A). Those account for users who watched the entire length of the video. The small portion of sessions that go over the video length are users who have extended the session by rewinding and replaying certain segments of the video.

Finally, we explore the relationship between session length and the popularity of a video. One might assume that the video with the highest demand will have the best chance of holding on to its audience for the duration of the video. Surprisingly, our data shows the opposite result.

We first plot the average NSL of each video with videos sorted in order from least to most popular. Here popularity is measured by the total number of user accesses throughout the log period. As seen in Figure 10, results are fairly scattered, showing a weak correlation. However, we do see a general trend showing that more popular videos often have lower NSL values than less popular videos.

To more clearly detect any possible correlation between the two, we partitioned all movies from Figure 10 into four quartiles according to their popularity. We then plotted the CDF of all NSL values of each quartile in Figure 11, where the line 0-25% indicates the most popular quartile of movies and 75-100% denotes the

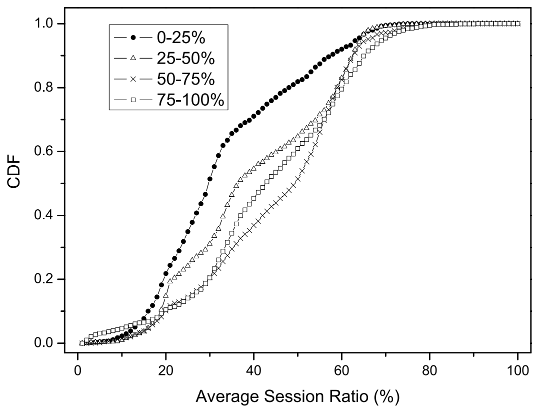

Figure 11: A CDF comparison of session lengths to video popularity, with four lines representing the four quartiles of popularity. 0-25% represent the most popular quartile, and 75-100% represent the least popular quartile.

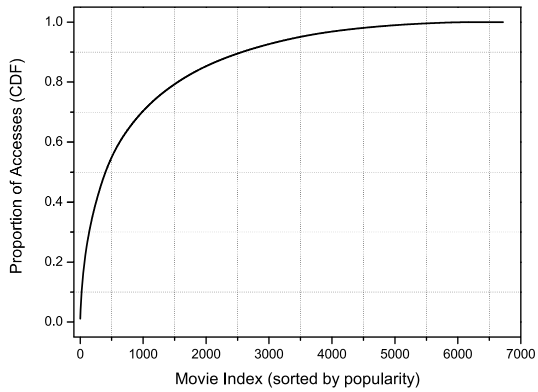

Figure 12: CDF of videos accessed sorted by popularity. A total of 6716 videos were requested at least once.

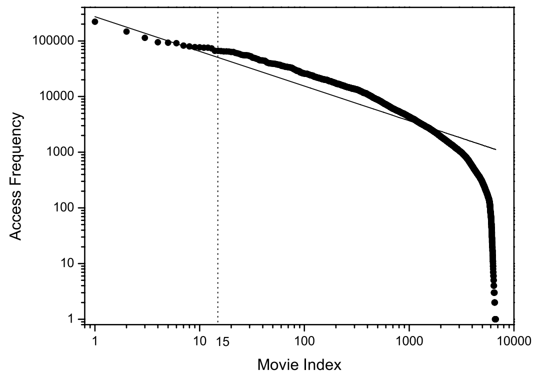

Figure 13: Fitting the video popularity distribution of videos across the 219 day log period to a Zipf distribution using a log-log graph.

least popular movies. We see that while weak, there is a distinct inverse relationship between popularity and session lengths. Note that the session lengths for videos of different popularity are similar at the beginning and end of videos. The main difference occurs in the middle of videos, where less popular videos do a better job of keeping the user's attention.

This inverse correlation is an interesting result that may provide insight into the way users watch videos. We hypothesize that the highly popular videos suffer from loss of interest after repeat viewings. In other words, people watching the most popular videos are likely to have seen them before, either in another medium (theater or DVD) or in a prior VOD session. Therefore, they lose interest more easily during the movie, resulting in shorter session times.

3.4 Implications

These results on file popularity and session times suggest a number of guidelines for optimizing performance in VOD systems. First, as expected, the clear diurnal patterns in user access patterns mean that maintenance and upgrade operations should be scheduled for early morning hours (5-8AM) in order to minimize impact on users. Note that we do not observe any effects of different timezones, since all of China operates on a single timezone. For systems deployed in the USA or Europe, we would expect a more even distribution of accesses in the morning hours. Next, our observation of short session times suggest that a high proportion (70%) of sessions are terminated in the first 20 minutes. Therefore, system caches can maximize their effectiveness by allocating the majority of their capacity to storing beginning segments of movies. Finally, our initial results show shorter sessions for popular movies, suggesting that beginning segments of popular movies should be prioritized over latter segments in any caching scheme. Clearly, VOD systems need to exploit time-varying user interest patterns by intelligently partitioning videos into segments and taking their time index in replica and cache management.

4. POPULARITY AND USER INTEREST

We now examine the issue of user interest and video popularity. Having an accurate model of how user requests spread across videos can help system designers choose system parameters such as cache size and replication factors for popular items. In this section, we look at how a static snapshot of video popularity compares to the Pareto principle, explore the applicability of the Zipf distribution to VOD requests, and examine how video popularity changes over time.

4.1 Pareto Principle

The Pareto Principle, or 80-20 rule, is the most popular rule used to describe the skew of user interest distributions. To test the accuracy of the Pareto principle on our data, we analyze user logs for the entire 219 days, and sort all objects who were accessed at least once according to how often they were requested. The results are plotted in Figure 12.

As we see from the figure, given the opportunity to choose from a wide selection of videos, user requests are spread more widely than predicted by the Pareto principle, with 10% of the most popular objects accounting for approximately 60% of all accesses while 23% of the objects account for 80% of the accesses. This is a more moderate result than the frequently referred to 80/20 or 90/10 rule. We believe that this moderate Pareto principle is relevant to VOD systems with relatively large libraries. Consequently, VOD systems are likely to require larger than expected caches in order to achieve the same hit rates predicted by the traditional Pareto Principle.

4.2 Request (Popularity) Distribution

In examining video popularity, one of our main goals is to explore the applicability of Zipf-like distribution to VOD requests.

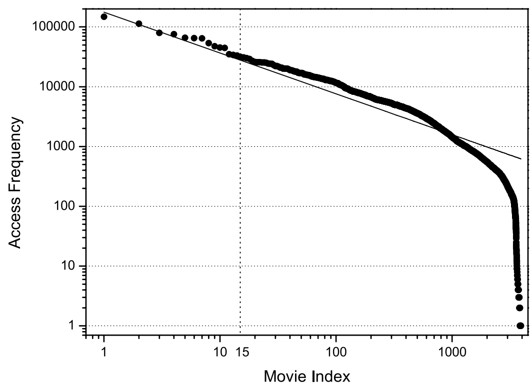

Figure 14: Fitting the popularity distribution of 3969 videos that were originally introduced at the launch date of PowerInfo.

Zipf's law is commonly used as a sound model to capture the distribution of media accesses. Early in 1994, Dan and Sitaram [10] considered the distribution of hits on the available videos and chose Zipf distribution to model video popularity. Wolf and Yu, in 1997, noticed that in their study that Zipf-like distribution roughly matched their access pattern, with a varying degrees of skew week by week. Breslau et al. [5] confirmed this result when analyzing characteristics of webpage request distribution. Later in 2000, Acharya and Smith [1] showed that Zipf distribution with a fixed parameter α does not accurately model the video file popularity distribution. In 2002, Cherkasova and Gupta [7] argued that although the distribution of client accesses to media files can be approximated by a Zipf-like distribution, the time scale plays an important role in this approximation. Finally, measurements by Gummadi et. al [13] showed that file downloads on the Kazaa [15] network did not follow the Zipf distribution, and instead proposed a *fetch-at-most-once* model.

In this subsection, we set out to determine how accurately Zipf-like distributions apply to our data and the characteristics of the varying skew factor. The Zipf-like distribution is defined as

$$\sum_{i=1}^{N} P_i = 1, P_k = \frac{\lambda}{K^{1-\alpha}}, \lambda = \frac{1}{\sum_{i=1}^{N} \frac{1}{i^{1-\alpha}}} \qquad (2)$$

In this formula, N is the number of available movie titles, i is the index of a movie title in the list of N movies sorted in order of decreasing popularity, and α is the skew factor. Setting $\alpha = 0$ corresponds to a so-called pure Zipf distribution, which is highly skewed. Setting $\alpha = 1$ corresponds to a uniform distribution. So Zipf-like distributions model a wide range of skew alternatives. It is also noteworthy that this distribution, typically used as the basis for investigations on video server operations, is completely independent of the number of users in the system.

In prior work by Gummadi et. al [13], the authors used a log-log plot to argue that popularity data in VOD systems did not in fact fit the Zipf distribution. The log-log graph showed that accesses for both the most and least popular items were lower than those predicted by Zipf. Instead, they proposed a "fetch-at-most-once" model to fit existing VOD data from a 1992 video-rental data set. To verify the applicability of Zipf distribution to our data, we plot the access frequency of videos in a log-log graph against a Zipf distribution. The results in Figure 13 show that unlike the 1992 video rental data set, most of our data fit the Zipf distribution well, with the exception of a heavy tail of unpopular items. This contradicts the "fetch-at-most-once" model, but fits within a video-on-demand model, where users cannot store streamed movies locally and must re-fetch the video for repeat viewing.

Table 6: Statistic summary of skew factors

N	Min.	Max.	Mean	Std. Dev.
209	0.000	0.348	0.199	0.070

Table 7: One-Sample Kolmogorov-Smirnov Test

N		209
Normal Parameters (a,b)	Mean	0.199
	Std. Dev.	0.070
Most Extreme Differences	Absolute	0.037
	Positive	0.028
	Negative	-0.037
Kolmogorov-Smirnov Z		0.531
Asymptotic Significance (2-tailed)		0.940

We speculated that the long tail of low popularity items might be due to the aging of old videos. To test this hypothesis, we identify the 3969 videos that were introduced into PowerInfo at the launch of the system on January 9, 2004. When we plot the log-log graph of their accesses through our log period in Figure 14, we see that the result is very similar to the overall data set, and does not confirm our hypothesis. We performed further analysis by plotting the log-log graph of all videos introduced *after* the launch date, and again the results are similar.

While popularity statistics across the entire log fit Zipf well, daily segments show Zipf fits with highly variable skew factors, as shown in Table 6. To determine the characteristics of the constantly-changing skew factor, we use the Kolmogorov-Smirnov Goodness-of-Fit Test [6]. The Kolmogorov-Smirnov test is useful in deciding if a sample comes from a population with a specific distribution, and it is defined as

$$Z = \sqrt{N} \cdot D, D = \max_{1 \leq i \leq N} |F(Y_i) - \frac{i}{N}| \qquad (3)$$

Here F is the theoretical cumulative distribution of the distribution being tested. The Z test statistic is the product of the square root of the sample size (N) and the largest absolute difference between the empirical and theoretical CDFs (D). Also, we calculated the two-tailed significance level (Asymptotic Significance), testing the probability that the observed distribution would not deviate significantly from the expected distribution in either direction. If the significance level result is above 0.05, then it was safe for us to assume that the data tested was not significantly different from the hypothesized (*e.g.* normal) distribution.

Using Kolmogorov-Smirnov Goodness-of-Fit, we tested our skew factor values against different distributions, and found that the normal distribution matched our data best. As shown in Table 7, the Asymptotic Significance value (0.940) was well above 0.05, indicating that the normal distribution is a good fit for the skew factors we observed in fitting Zipf's law.

Further, the Normal Probability-probability Plot, or Normal P-P Plot, plots the cumulative proportion of a single numeric variable against the cumulative proportion expected if the sample were from a normal distribution. If the sample is from a normal distribution, the points will cluster around a straight line. This is one method of assessing whether a variable is normally distributed. As shown in

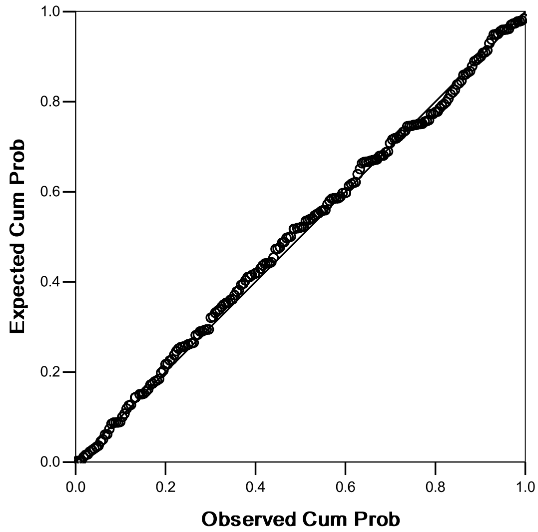

Figure 15: Normal P-P Plot of Skew Factor

Table 8: Descriptive statistics of skew factor

Skewness	Std. Error	Kurtosis	Std. Error
-0.253	0.168	0.048	0.335

Figure 15, our data points are well represented by a straight line. This together with our Kolmogorov-Smirnov result above, shows that the Zipf's skew factor is normally distributed in our VOD system.

Table 8 also presents the descriptive statistics of all of the 209 skew factors we collected during our analysis. Here, a negative skewness value means that our data has a long left tail, but it still can be taken as a symmetrical distribution since the skewness value is not more than twice its standard error. Meanwhile, the positive kurtosis indicates that our observations cluster more and have longer tails than those in the normal distribution.

4.3 Rate of Change in User interest

Our analysis of video popularity shows user interest is spread widely across a number of videos. We now examine the rate at which user interest changes, an important design consideration for VOD system architects. To optimize the performance of VOD systems, a commonly used approach is to move popular objects around the network based on the transfer of user interest. So it is essential to consider how frequently the content in the buffer or peer server should be refreshed. This subsection analyzes the rate of change in user interest by examining the videos that make up the top-10, top-100 and top-200 in accesses for different time periods.

Figures 16, 17 and 18 show the percentage of change in the top 10/100/200 videos over different time scales. From Figure 16, we see that user interest varies significantly on an hourly basis. The highest rates of change seem to occur near the morning, when the rate of change almost reaches 30% in the top 10 videos, and around 50% in the top 100. It is noteworthy that the top 10 is significantly more stable compared to the top 100, and only changes an average of 10-20% for most hours of the day. Note that during the most busy hours from 11AM to midnight, the top-10 list is remarkably stable. We hypothesize that the increased variance in the top 10 between midnight and 8AM is due to the more varied and unpredictable access patterns of a smaller user set between those hours. In general, these results suggest that an adaptive cache that focuses on storing the top 10 accessed videos will be able to serve the top 80-90% of the most frequently requested videos.

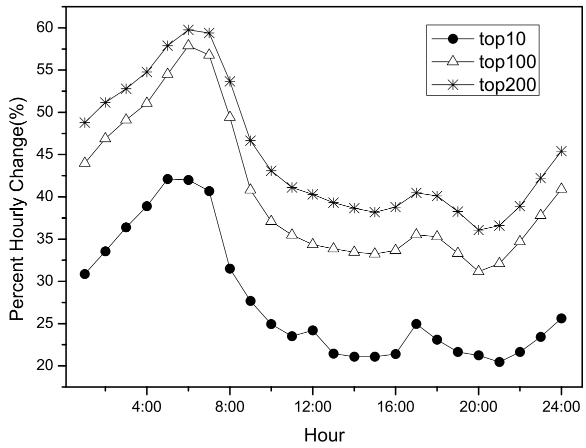

Figure 16: Rate of change in user interest, as seen over hours in a single day.

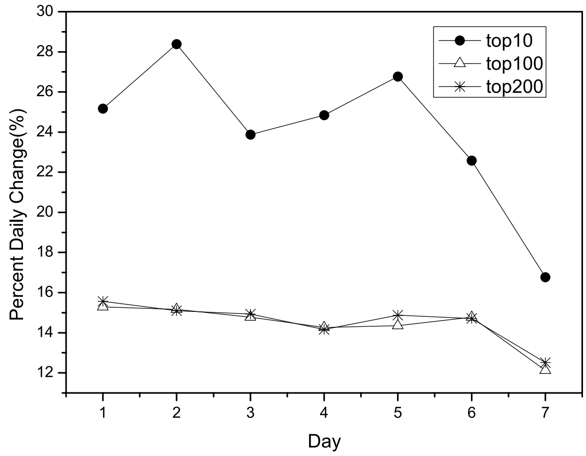

Figure 17: Rate of change in user interest, seen over days in a single week. Day 1 represents Monday while day 7 represents Sunday.

In order to get some idea of how fast users' interest changes day by day, we calculate the daily change across the course of a week, and average those results for all weeks during our log period. The result shown in Figure 17 shows that the top-10 list fluctuated significantly through the week, while the top-100 and top-200 lists were much more stable with a steady change rate between 12 and 15 percent. The underlying reason for this result is that our system inserts some new objects into the content server every day, and that the latest video files usually supplant the current favorites within a day of being released.

Finally, Figure 18 shows the rate at which the top movies changed week by week. Here we can draw the same conclusion as from the daily change pattern. The only difference is that the top-10 list fluctuates more radically at this time scale. Note that in the top 200 videos on Figure 17 and Figure 18, the rate of change closely matches the top 100 curve. This suggests that there is likely very

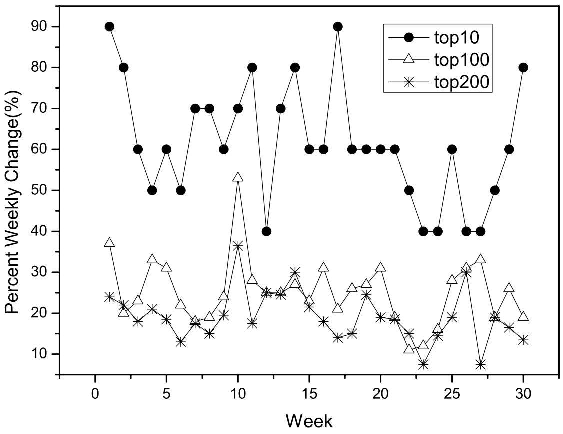

Figure 18: Rate of change in user interest, seen over weeks across the log period.

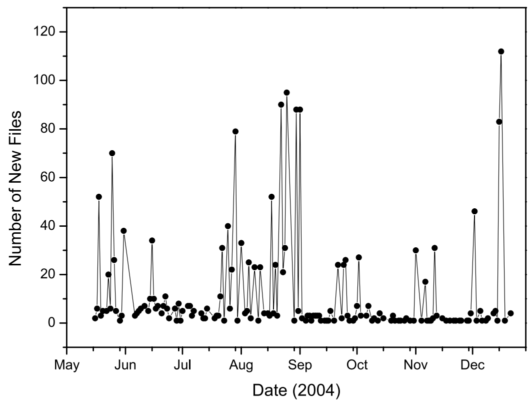

Figure 19: New video files introduced on a daily basis.

little performance gain between caching the top 200 videos and caching the top 100.

The key implication of these results is that the top 10 and top 100 most popular videos exhibit churn on different time scales. The top-10 lists exhibits relative stability on an hourly basis, but shows high churn on a daily or weekly basis. The top 100 videos however, shows stability over the long term. This suggests a two level caching model, where a fast but relatively small adaptive cache focuses on capturing a small number ($\sim$ 10) of the most popular videos, while a potentially slower but larger secondary cache serves content based on the second tier of popular videos.

5. UNDERSTANDING POPULARITY

Our analysis has shown that a small number of videos account for a large proportion of total user accesses. It is unclear whether this is purely a social phenomenon, or whether it has been influenced by external factors. In this section, we seek to understand this issue by analyzing data on the introduction of new videos into the system, and examining the impact of external factors such as official recommended movie lists or lists of most popular videos.

5.1 Introduction of New Content

Tang[22] introduced a new file introduction process in HP corporate media servers. He argued that the time gap between two

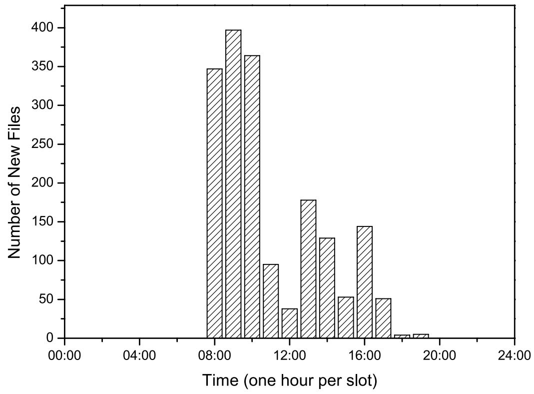

Figure 20: Hourly histogram of when new videos are introduced into the PowerInfo system.

introduction days followed a Pareto distribution. However, the data in our system shows no such distribution. In fact, new movies are generally added to system on a daily basis.

While PowerInfo launched in January 2004 with an initial video library of 3969 videos, our log of new file introductions began on May 16, 2004. Figure 19 presents the overview of the new file introduction process in our VOD system on a daily level (from May 16th, 2004 to December 22nd, 2004). Note that on some days more than 30 new files are introduced into the system. In such cases, those clusters of files are usually TV mini-series movies or cartoons. From the user interest perspective, multiple episodes of the same TV mini-series are similar to a single movie, and usually do not result in the same disruption of user interest as would if the same number of independent videos were added to the system.

The next logical question we answer is what time of the day is new content most often introduced. Figure 20 shows a cumulative histogram showing when new videos were introduced into the system. We see that most new content is introduced during the morning hours (8AM-10AM), when the system is lightly loaded. If we revisit our data on the rate of change in user interest shown in Figure 16, this corresponds to the relatively high rate of change between 9AM and 11AM. Clearly, the availability of new content captures users' attention and requests, thereby changing the distribution of user requests.

These results demonstrate the direct impact new content introduction has on user requests. Correlation to fluctuations in user interest suggests that dynamic caches should be re-calibrated shortly after a large amount of new content is introduced to the system. This way, users can quickly adapt to new video popularity patterns emerging after the integration of new material.

5.2 Impact of Recommendations

To enhance the user experience, the PowerInfo web interface includes some user-friendly features, including two features. One is a list of the 15 most popular movies of the month, the top-15 hot list, with a link to each movie. The other is a list of 20 movies recommended by the system. Most of these are recent additions to the system. As our analysis will show, the appearance of movies on these two lists has a significant impact on their popularity within the VOD system.

We start our analysis by taking a closer look at the rise and fall of a typical video in our system, movie 14102. We choose movie 14012 because its history of popularity is quite representative of most videos in our system. In Figure 21, we plot both its rank

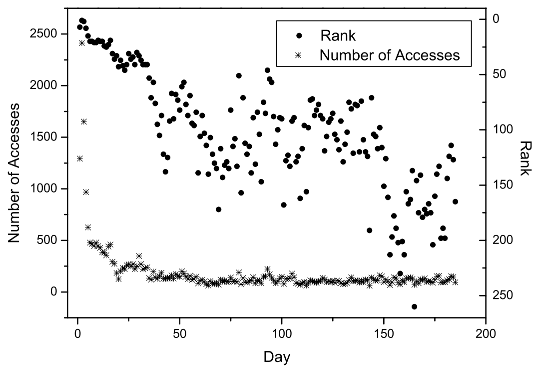

Figure 21: The "life span" in rank and accesses of an average movie: 14102.

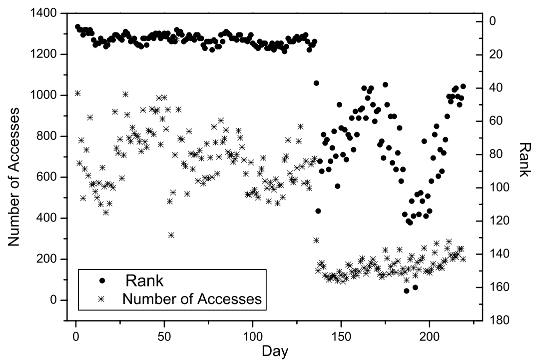

Figure 22: The "life span" of one of the most popular movies, 116, showing the impact membership on top-15 hot lists.

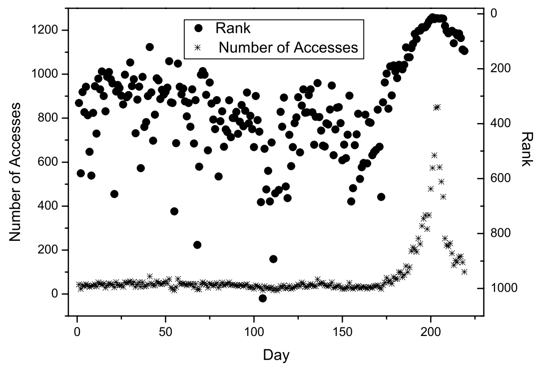

Figure 23: The "life span" of an older movie, 9757, showing the impact of a movie remake rekindling interest.

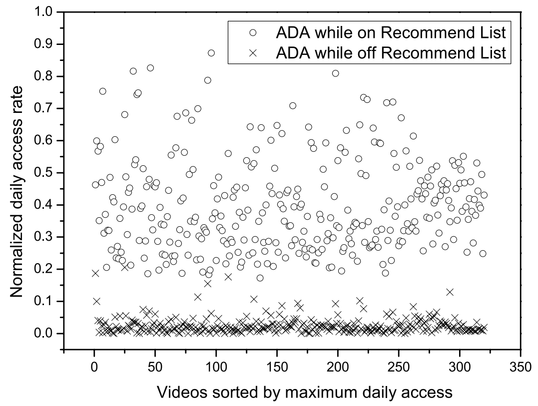

Figure 24: The impact of membership on the recommendation list on normalized average daily accesses.

in terms of video popularity and also its number of user requests against time. When it is first introduced into the system on June 19, 2004, 14102 became the 7th most popular video in the system. On the next day, it became the most frequently requested video with 2413 daily accesses. But in the next few days, its popularity rank decreased rapidly.

We observed that while most videos share a similar burst of popularity, their ability to maintain popularity is what sets them apart from each other. While some movies stay consistently popular, others exit the limelight quickly, dropping from 1000 daily accesses to 50 in one week.

The effect of the hot movie list can be seen from the life span of movie 116, shown in Figure 22. This movie is one of the top 5 in total accesses across our entire log period, and maintains its popularity ranking in the top 15 for a significant amount of time. But once 116 dropped out of the ranks of the top 15, it was never able to recover its popularity, and both accesses and rank dropped significantly. What makes the story of movie 116 even more interesting is the unusual way that it dropped out of the top-15 hot list. We found out from the system administrator that on day 136 of our log, he saw movie 116 had been on the top-15 list for several months, and manually removed it from the top-15 hot list in favor of more recent popular movies. While the top-15 list is usually generated monthly by the system, this rare external intervention explains the drastic change in 116's popularity.

In another case study, Figure 23 shows the lifetime of movie #9757. Based on an old Chinese novel, 9757 is an older movie in the video library on PowerInfo's launch date. However, the later release of a new film based on the same book dramatically increased interest in this older version, accounting for a dramatic rise in ranking and requests near the end of our log period.

While these individual case studies prove interesting, we need to further quantify the impact of the recommendation list and the top-15 hot list on the access popularity of videos. To show this for each movie, we plot the movie's average daily access rate while it is on and off the manager's recommend list, and plot both as a normalized ratio of its maximum daily access rate. For example, a video might reach a maximum daily access rate of 3000 requests on a particular day. We would calculate the average daily access rate for all days when it was on the recommend list, and normalize that against 3000.

In Figure 24, we plot this pair (on/off) of normalized average daily access (ADA) rates for all videos that have ever appeared on the recommend list. sorted by its maximum daily access rate. We observe that the inclusion on the recommend list drastically increases the average daily access rate, generally by an order of magnitude or more. Most videos only average less than 5% of their maximum daily access rate when they are not included in the recommend list, but can average anywhere from 20% to 90% of their maximum rate while listed on the recommend list. Clearly, an external opinion, or perhaps just membership on an "official" list, can dramatically influence users on their choice of video selections. This is likely a social phenomenon that applies to movie rental out-

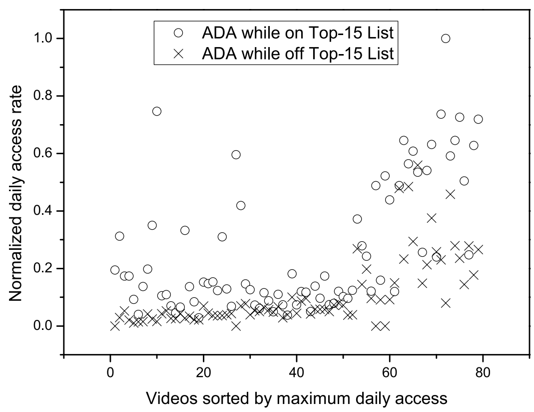

Figure 25: The impact of membership on the Top-15 list on normalized average daily accesses.

lets in general, and not specific to the PowerInfo VOD system. We also note that this impact appears to be independent of the actual video popularity. Thus we would expect the same relative "burst of popularity" from recommending a title, regardless of how popular it was originally.

Next, we use the same technique to analyze the impact of membership on the top-15 hot list, and show the results in Figure 25. Note that because the top-15 list is calculated on a monthly basis, only a small number (< 80) of videos have ever been listed. Because videos have to sustain their popularity for a month in order to enter the top-15 list, our sample set only includes highly popular videos, and the relative impact of membership is less significant. In contrast, popularity-independent lists like the recommendation list covers a greater selection of videos, and is likely to achieve greater impact. Finally, we note that the largest difference between membership and non-membership occurs for videos with lower popularity, suggesting a scenario where removal from the list resulted in a significant drop in daily access rates similar to the history of movie 116 as shown in Figure 22.

6. RELATED WORK

Many studies have been carried to analyze the user behaviors in different media services, including web services, file sharing services, media broadcasting services and on-demand video streaming services.

Due to the lack of a real deployed large-scale VOD streaming system, previous VOD studies mainly relied on data from video rentals, small-scale systems or web-based Internet streaming services. While studies of web-based video streaming is similar in scale to a large VOD system, such services were deployed on lower bandwidth links compared to the broadband connections used in VOD systems. As observed in [8] and [14], lower bandwidth connections resulted in lower quality of service and impatient users. Another implication of the lower bandwidth is seen in the relatively smaller object sizes. Finally, these systems did not provide features like movie recommendation or top-ranked video lists. These features not only make the user experience more enjoyable, but as our results have shown, also have significant impact on user access patterns. In comparison, we believe our study provides results much more relevant to the design of future large-scale VOD systems.

The pioneering study on video-on-demand services modeled user behavior according to a week's worth of empirical data on video rentals in various video stores [10]. Later, Griwodz [12] introduced a model of movies' long-term life cycle using data from a video store and movie magazine. Conclusions such as the Zipf assumption from these offline data sets were highly influential.

Acharya [1] presented the analysis of a little more than six months of trace data from a multicast media on demand (mMOD) system with a mix of educational and entertainment videos. Its analysis was based on a smaller system covering only 139 videos, and had a much larger average request inter-arrival time of 400 seconds.

The sequence of studies in [2], [3] and [9] focused on user behavior and data access patterns in video streaming systems. Their studies included two kinds of video streaming services. The first is a small scale VOD system named eTeach and BIBS, and include a small number of files being accessed by a specific group of users: undergraduate students. The difference in content, size and audience probably accounts for the difference in results between our studies. Since our log data provided relatively little information about specific users, we believe our studies can be viewed as complementary. The second was a study of web-based streaming services, which as we described above, are significantly different from VOD systems.

Cherkasova [7] and Tang [22] explored workload characteristics based on logs from two internal media servers at Hewlett-Packard. The servers were limited in delivering company-related content to employees. We believe the limited choice of topics limit the applicability of their results on general VOD systems. Finally, these services only observed a light workload of less than 1 million sessions in 29 months. In contrast to these studies, the system used to gather our data served more than 150,000 users and 21 million requests under varying amount of load.

Veloso [23] and Sripanidkulchai [21] characterized live Internet streaming media workloads. Veloso showed that the object-driven nature of interactions between users and objects for live streaming is fundamentally different from stored objects. Sripanidkulchai affirmed the feasibility of application level multicast with enough resources and nonlethal dynamics.

While most of the above works used media server logs, Mena [17], Chesire [8] and Guo [14] used different traffic tracing methods to trace the user accessing data in Internet streaming systems. They gathered data across multiple sites and multiple media types, but cannot detect properties specific to a single site or media type. Gummadi et. al [13] used tracing at the network border of Univ. of Washington to gather data on multimedia file-sharing in Kazaa, and extended their results to VOD systems.

Finally, Paxson [18] investigated a number of wide-area TCP arrival processes to determine the error introduced from modeling them using Poisson processes, and showed that not all network arrivals are well-modeled by Poisson distributions. We show that even user-initiated session arrivals could not be modeled by pure Poisson distribution, but instead, by a modified version of it.

7. CONCLUSION

Multimedia streaming has become a dominant factor in today's Internet, and Video-on-Demand is one of the most promising killer applications for the Internet of the future. The lack of data from deployed VOD systems has limited researchers to relying on assumptions about user behavior and content access patterns. In this study we tracked, analyzed and modeled user behavior and relevant access patterns in a large scale VOD environment. Our results show that the timing of user accesses are predictable, but user arrival rates do not match the Poisson distribution. Instead, we propose a modified version of the Poisson distribution which matches our empirical data. We also found low but significant inverse correlation between a video's average session length and its popularity.

Our analysis showed that video popularity matched the Zipf distribution better than predicted using the "fetch-at-most-once" model. In addition, we found change in video popularity was strongly influenced both by the introduction of new content as well as external factors such as recommendation lists and popularity rankings. Results of our analysis should help VOD system designers make intelligent decisions about resource allocation and techniques for performance optimization. Finally, we are working to make our data set publicly available to researchers worldwide.

8. ACKNOWLEDGMENTS

We wish to thank Yi Li, CTO of Powerinfo Software Co. of China for helping us collect log data from the Powerinfo streaming servers. We also thank our shepherd Thomas Gross and the anonymous reviewers for their valuable feedback.

9. REFERENCES

[1] ACHARYA, S., SMITH, B., AND PARNE, P. Characterizing user access to videos on the world wide web. In *Proc. of ACM/SPIE Multimedia Computing and Networking* (January 2000).

[2] ALMEIDA, J., KRUEGER, J., AND VERNO, M. Characterization of user access to streaming media files. In *Proc. of ACM SIGMETRICS / Performance* (2001).

[3] ALMEIDA, J. M., KRUEGER, J., EAGER, D. L., AND VERNON, M. K. Analysis of educational media server workloads. In *Proc. of NOSSDAV* (Port Jefferson, NY, June 2001).

[4] Approaches to controlling peer-to-peer traffic: A technical analysis. White paper from `http://www.p-cube.com`, 2003.

[5] BRESLAU, L., CAO, P., FAN, L., PHILLIPS, G., AND SHENKER, S. Web caching and zipf-like distributions: Evidence and implications. In *Proc. of INFOCOM* (New York, NY, March 1999).

[6] CHAKRAVARTI, I. M., LAHA, R. G., AND ROY, J. *Handbook of Methods of Applied Statistics*, vol. I. John Wiley and Sons, 1967.

[7] CHERKASOVA, L., AND GUPTA, M. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *Proc. of NOSSDAV* (May 2002).

[8] CHESIRE, M., WOLMAN, A., VOELKER, G. M., AND LEVY, H. M. Measurement and analysis of a streaming-media workload. In *Proc. of USITS* (San Francisco, CA, March 2001).

[9] CRISTIANO, C., CUNHA, I., BORGES, A., RAMOS, C., ROCHA, M., ALMEIDA, J., AND RIBERIO-NETO, B. Analyzing client interactivity in streaming media. In *Proc. of WWW* (New York, NY, 2004).

[10] DAN, A., SITARAM, D., AND SHAHABUDDIN, P. Scheduling policies for an on-demand video server with batching. In *Proc. of ACM Multimedia* (October 1994).

[11] GOUGH, P. J. Comcast video-on-demand hits 1 billion mark. Yahoo News, October 2005.

[12] GRIWODZ, C., BAR, M., AND WOLF, L. C. Long-term movie popularity models in video-on-demand systems. In *Proc. of ACM Multimedia* (1997).

[13] GUMMADI, K. P., DUNN, R. J., SAROIU, S., GRIBBLE, S. D., LEVY, H. M., AND ZAHORJAN, J. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proc. of SOSP* (October 2003).

[14] GUO, L., CHEN, S., XIAO, Z., AND ZHANG, X. Analysis of multimedia workloads with implications for internet streaming. In *Proc. of WWW* (Chiba, Japan, May 2005).

[15] KaZaa media desktop. `http://www.kazaa.com`. Using Fasttrack: `http://www.fasttrack.nu`.

[16] Managing peer-to-peer traffic with cisco service control technology. `http://www.cisco.com`, 2005.

[17] MENA, A., AND HEIDEMANN, J. An empirical study of real audio traffic. In *Proc. of INFOCOM* (March 2000).

[18] PAXSON, V., AND FLOYD, S. Wide-area traffic: The failure of poisson modeling. *ACM/IEEE Transactions on Networking 3*, 3 (June 1995).

[19] SAROIU, S., GUMMADI, K. P., DUNN, R. J., GRIBBLE, S. D., AND LEVY, H. M. An analysis of internet content delivery systems. In *Proc. of OSDI* (December 2002), ACM, pp. 315–328.

[20] SEN, S., REXFORD, J., AND TOWSLEY, D. Proxy prefix caching for multimedia streams. In *Proc. of INFOCOM* (New York, NY, March 1999).

[21] SRIPANIDKULCHAI, K., GANJAM, A., MAGGS, B. M., AND ZHANG, H. The feasibility of supporting large-scale live streaming applications with dynamic application end-points. In *Proc. of SIGCOMM* (Portland, OR, August 2004).

[22] TANG, W., FU, Y., CHERKASOVA, L., AND VAHDAT, A. Medisyn: Synthetic streaming media service workload generator. In *Proc. of NOSSDAV* (2003).

[23] VELOSO, E., ALMEIDA, V., MEIRA, W., BESTAVROS, A., AND JIN, S. Hierarchical characterization of a live streaming media workload. In *Proc. of Internet Measurement Workshop* (Marseille, France, November 2002).

[24] ZATZ, D. Tivo inches towards video on demand. RealTechNews Article, September 2005. `http://www.realtechnews.com/posts/1834`.